Project Milestone 1
(Final Submission)

Deep Learning and Software Engineering

TEAM 10

Agustín Tamagnone

Birk Bregendahl

Karol Swiderski

Milestone 1 – Initial

Index

Revision History

Revision Number	Revision Date	Comments
Version #1	22/10/2025	Initial version.
Version #2	26/10/2025	Added sub-section System Architecture (1.2). Added sub-section 'Models Considered' (2.1) Modified body text in Section 2.1

1. Context and Requirements

This project focuses on developing an interactive web application that enables students, journalists, and the general public to explore the credibility of online news articles. The goal is to create an educational tool that helps users understand how language models can analyze writing style and tone to estimate the *likelihood* that an article appears misleading or credible. The system does not perform fact-checking but instead provides a probability-based indication of how "real-like" a piece of text appears, based on patterns learned from data.

The application integrates a Deep Learning (DL) model based on DistilBERT, a transformer architecture capable of understanding contextual relationships between words. The model is trained using the ISOT Fake News Dataset from the University of Victoria (Canada), which contains labeled *True* and *Fake* articles with four key attributes: title, text, subject, and date. By analyzing these components, DistilBERT captures both linguistic and semantic cues commonly present in unreliable or trustworthy news sources.

The system's objective is to support the academic community within UPM (Universidad Politécnica de Madrid) by providing a publicly accessible platform for experimentation and learning. It will allow users to test how modern language models interpret written content and assess the limitations of AI-based credibility analysis.

1.1 System Overview

The Fake News Detection system will be deployed as a web-based application accessible to students, professors, and other authorized UPM users via institutional credentials. Users will be able to paste or upload a news article (title and text), and the application will send this data to a backend inference service that hosts the DistilBERT model.

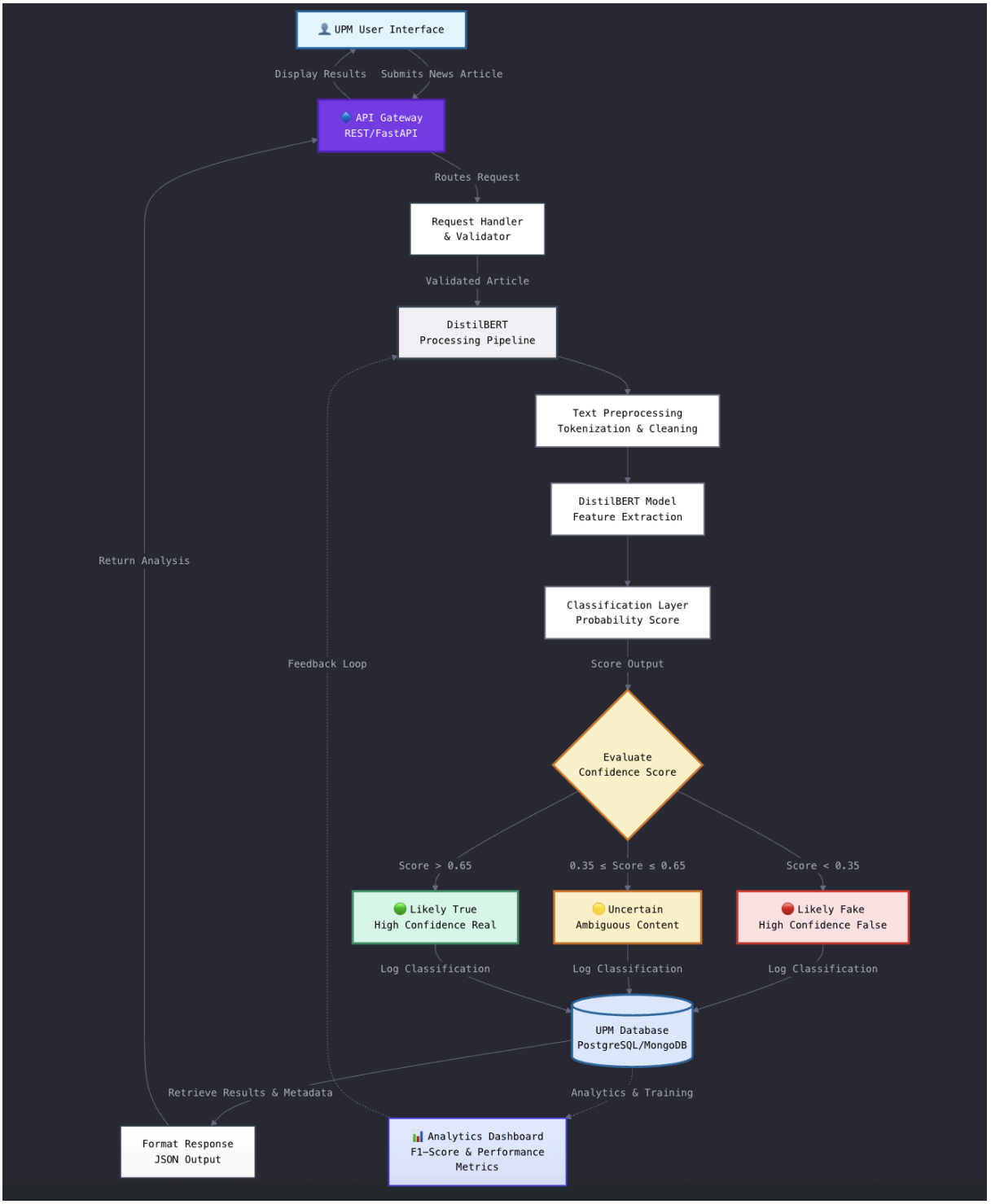
The backend processes the input, runs inference through DistilBERT, and returns a "true likelihood" score ranging from 0 to 1. This score represents how similar the writing appears to true or reliable news, based on patterns identified during model training. The web interface will display the result visually—using colors or gauges—and may highlight specific words or phrases that most influenced the prediction, increasing model transparency and user understanding.

Users will also be able to provide feedback on predictions by marking them as "correct" or "incorrect." These responses will be stored anonymously and can later be used for model retraining and continuous improvement.

From a software engineering perspective, the system is designed around clear separation of components:

- a **frontend** for interaction and visualization,
- a **backend API** for processing and model inference, and
- a **logging** and **data storage** layer for recording predictions and feedback.

1.2 System Architecture



1.3 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 true.

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.

1.4 Non-Functional Requirements

The system must operate efficiently on standard hardware. The deep learning component, based on the DistilBERT transformer model, should process an article within approximately five seconds on a CPU or standard workstation, while the web interface is expected to display results shortly after receiving the model's response. Although designed primarily as an educational and research tool accessible to UPM users, the application aims to maintain around ninety-five percent uptime during use. In the event of a technical issue—such as a model or API failure—the system will respond gracefully by displaying an informative message instead of crashing.

All communication between the web client and backend service is encrypted via HTTPS, and access to the system is limited to authenticated users through UPM credentials. To ensure privacy, article texts and predictions are stored only when

required for evaluation or retraining, and all stored data are anonymized according to privacy principles.

The architecture follows a modular design, separating the frontend, model inference, and data-handling layers, which simplifies maintenance and future development. This design supports scalability, enabling additional features such as feedback-based retraining or integration with classroom dashboards.

1.5 Success Criteria

At this stage, the focus is on defining and implementing the initial architecture of the DistilBERT-based model rather than evaluating its performance. No formal testing has been conducted yet; however, model training and evaluation will take place in subsequent milestones.

The main objective for future stages is for the DistilBERT model to achieve strong performance in distinguishing *true-like* writing styles, with a target F1-score of approximately 0.85 or higher on a held-out test set. Predictions should be computed within a few seconds per article on standard hardware, and the training process will be fully reproducible. A short error analysis will be prepared once evaluation results are available, including examples of misclassifications and user feedback integration.

1.6 Key Risks and Mitigations

The project faces several challenges typical of NLP applications, including bias, domain limitations, and interpretability issues. Because the ISOT dataset primarily covers political topics, the model may still associate emotional or sensational language with fake content rather than factual inaccuracy. This limitation will be documented, and results will be analyzed by subject category to detect bias.

Another concern involves domain drift when applying the model to new or non-political articles. To reduce this effect, the system will allow continuous retraining using user feedback collected through the web interface.

While DistilBERT mitigates some issues of sparse features seen in TF-IDF models, there remains a risk of overfitting to the training data or misinterpretation of context. These risks will be addressed through early stopping, dropout, and performance monitoring on validation sets. Continuous user evaluation will further help identify weaknesses and guide fine-tuning.

1.7 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 the article is to be real. The problem is therefore framed as a binary text-classification task, where values close to 1 represent higher likelihood of real news and values close to 0 indicate fake or unreliable content.

Our system uses DistilBERT, a compact and efficient version of the BERT transformer architecture. DistilBERT is an encoder-only model pretrained on a large corpus of English text and fine-tuned for text-classification tasks. Unlike earlier approaches that rely on sparse statistical representations such as TF-IDF, DistilBERT uses contextual embeddings that capture the meaning of each word relative to its surroundings. This allows the model to understand nuanced differences in phrasing, tone, and semantics that often distinguish credible reporting from deceptive writing.

Each article is first tokenized using the DistilBERT tokenizer, which divides the text into sub-word units and maps them to integer IDs. Positional encodings preserve word order, and the tokens are then passed through multiple self-attention layers that compute relationships between all words in the sequence simultaneously. The resulting contextual embeddings are aggregated and fed into a classification head—a fully connected layer with a sigmoid activation—that outputs a single probability between 0 and 1.

The model will be fine-tuned on the ISOT Fake News dataset using binary cross-entropy loss and optimized with AdamW, a variant of the Adam optimizer tailored for transformer models. During evaluation, we will monitor metrics such as accuracy, precision, recall, and F1-score to measure predictive performance.

2.1 Models Considered

Several model families were reviewed before selecting DistilBERT for this project:

- **Feed-Forward Dense Neural Networks (FF DNNs):**
  Simple and computationally lightweight; however, they depend on fixed vector representations (e.g., TF-IDF) that ignore word order and contextual

relationships. Their performance typically degrades on linguistically complex or longer texts.

- **Recurrent Neural Networks (RNNs / LSTMs):**
  Effective for sequential data and capable of retaining short-term dependencies, but training is slow and they struggle with long-range relationships common in multi-sentence articles.

- **Convolutional Neural Networks (CNNs):**
  Capture local n-gram patterns efficiently but are less suited for modeling sentence-level meaning and global context.

- **Transformer-based Models (BERT, DistilBERT, RoBERTa):**
  Leverage self-attention to analyze all words in a text simultaneously, enabling a far deeper understanding of semantics and writing style. Among these, DistilBERT offers nearly the same accuracy as full BERT while being approximately 40 % smaller and 60 % faster, making it ideal for CPU-based inference and limited-resource environments such as this educational project.

Based on this comparison, we selected DistilBERT as the single model for all milestones. It provides an optimal balance between accuracy, interpretability, and computational efficiency. DistilBERT's pretrained contextual understanding allows it to generalize well even with moderate data volumes, aligning perfectly with the project's educational focus and hardware constraints. The model's output can be easily integrated into the system's user interface and feedback mechanisms, ensuring consistency across all future milestones.

3. Model Card

3.1 Model Details

- Developed by students at Universidad Politécnica de Madrid, 2025, v1.
- It uses a transformers based model, DistilBERT.
- Binary classification, the model predicts the likelihood for a news article to be *False* or *True*.

3.2 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.

3.3 Factors

- Results may vary depending on:
    - Topic of the article (for example, politics, world news, sports).
    - Writing style of the article. Opinion and biases from the dataset (tone and topic imbalance) could affect predictions.
    - Language (our dataset only contains English), and the DistilBERT model was only trained in English, so it will not predict articles news in other languages.

3.4 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 = true, below = fake).
    - Likely fake (< 0.35)
    - Uncertain (0.35 < x < 0.65)
    - Likely true (> 0.65)

3.5 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) CSVs*.
- Each record includes: title, text, subject, and date.
- Fine-tune DistilBERT directly on the text, using both the title and body of each article.
- The model's tokenizer will automatically turn the text into the numerical format needed for training.

3.6 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.

3.7 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.

3.8 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

4.1 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.

4.1.1 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.

4.1.2 Languages

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

4.2 Dataset Structure

4.2.1 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

4.2.2 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

4.3 Dataset Creation

4.3.1 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.

4.3.2 Source Data

4.3.2.1 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.

4.3.2.2. 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.

4.3.3 Annotations

4.3.3.1. Annotation Process

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

4.3.3.2. Who are the Annotators?

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

4.3.4 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.

4.4 Considerations for Using the Data

4.4.1 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.

4.4.2 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)

4.5 Additional Information

4.5.1. Dataset Curators

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

4.5.2. Licensing Information

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

4.5.3 Citation Information

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

4.5.4 Contributions

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