System analysis and design

Unintended Bias in TOXICITY CLASSIFICATION

Hugo Mojica. Andrey Gonzales. Laura Paez

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

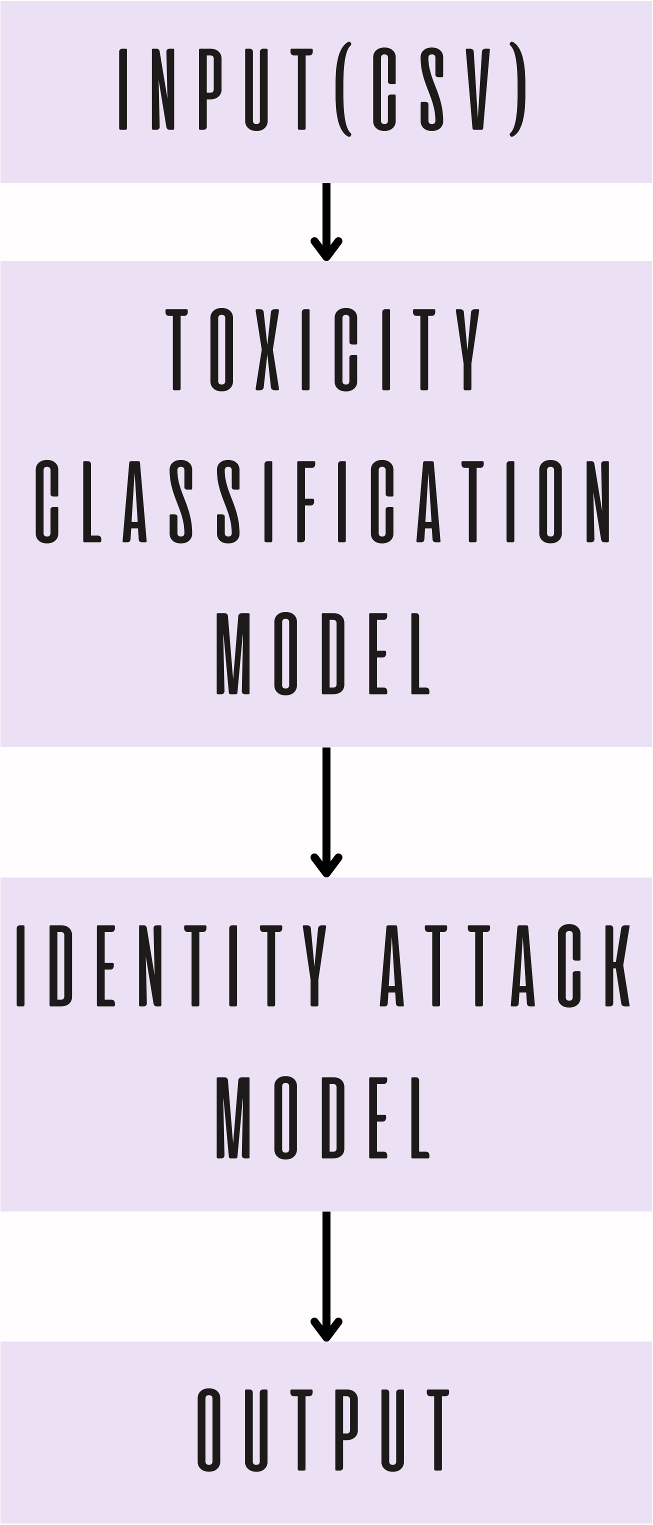
INTRODUCTION

Kaggle competition: Unintended Bias in toxicity classification:

- Conversation AI detects problems in learning models associating frequently attacked identity names with toxicity
- Models predicted a high likelihood of toxicity for comments containing those identities (e.g. "gay"), even when those comments were not actually toxic (such as "I am a gay woman").
- The problem arises in the data to train the model, as it contains these identities which are used in a negative context

PROBLEM AND GOAL

- Build a model that can identify seven types of toxicity without fail for the identity mentions.
- Reduce false positives for identity-related comments using fairness-aware techniques.
- Replace TF-IDFCallan (2003), by the incorporation of LLMsDevlin et al. (2019), to
- improve contextual accuracy
- · Evaluate performance via Kaggle metrics (Subgroup AUC, BPSN AUC).



INPUT(CSV)

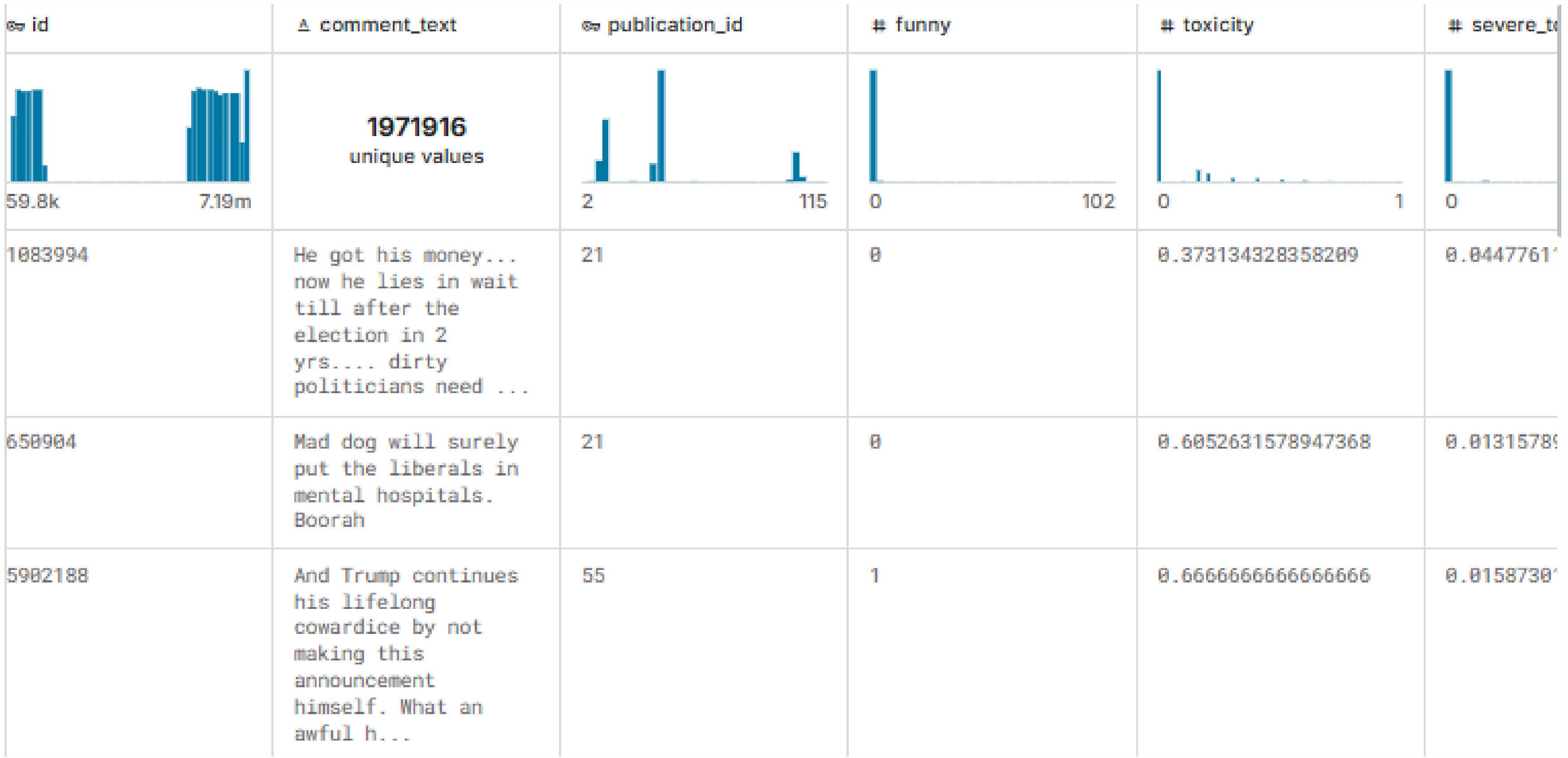
TOXICITY CLASSIFICATION MODEL

IDENTITY ATTACK MODEL

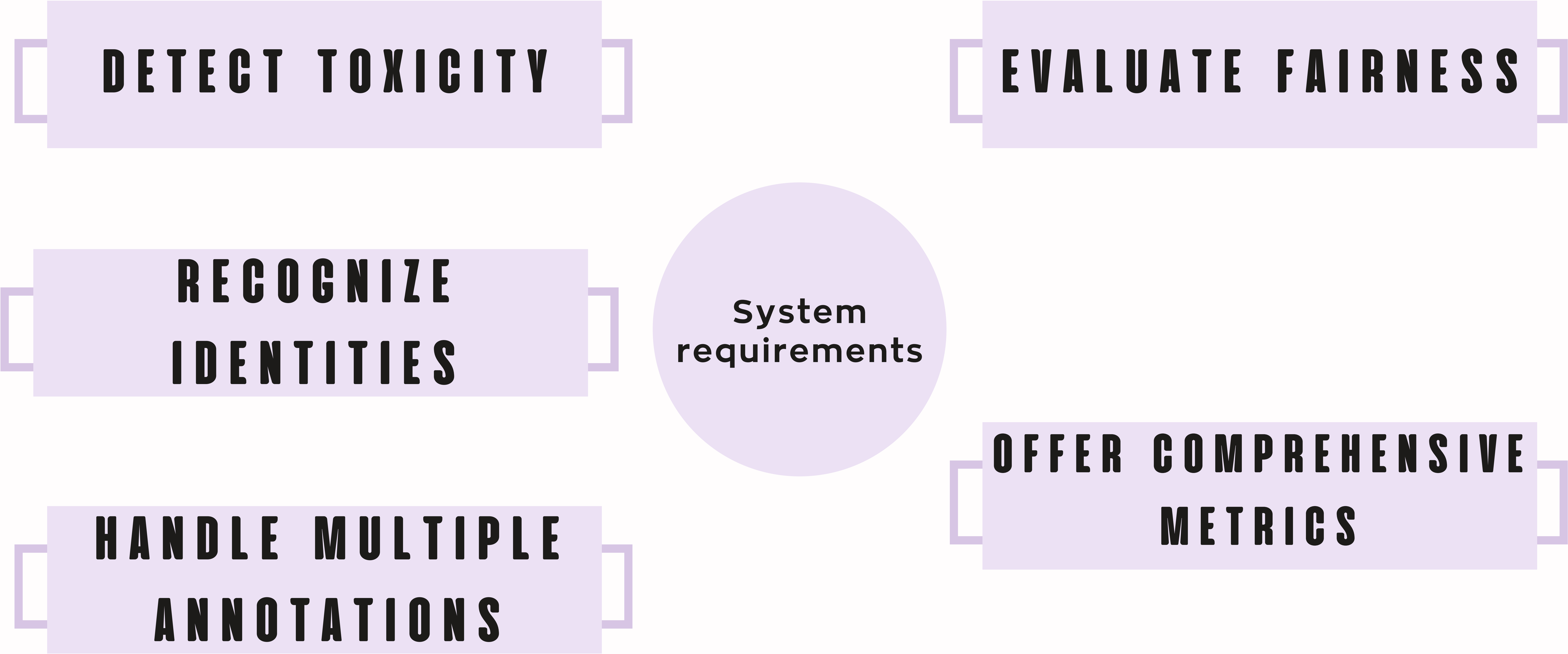
OUTPUT

DATA

train.csv: the training set, which includes toxicity labels and subgroups
test.csv: the test set, which does not include toxicity labels or subgroups
sample_submission.csv: a sample submission file in the correct format



id	comment_text	publication_id	funny	toxicity	severe_to
59.8k 7.19m	1971916 unique values	2 115	0 102	0 1	0
1083994	He got his money... now he lies in wait till after the election in 2 yrs.... dirty politicians need ...	21	0	0.373134328358209	0.0447761
650904	Mad dog will surely put the liberals in mental hospitals. Boorah	21	0	0.6052631578947368	0.0131578
5902188	And Trump continues his lifelong cowardice by not making this announcement himself. What an awful h...	55	1	0.6666666666666666	0.0158730



System requirements

DETECT TOXICITY

RECOGNIZE IDENTITIES

HANDLE MULTIPLE ANNOTATIONS

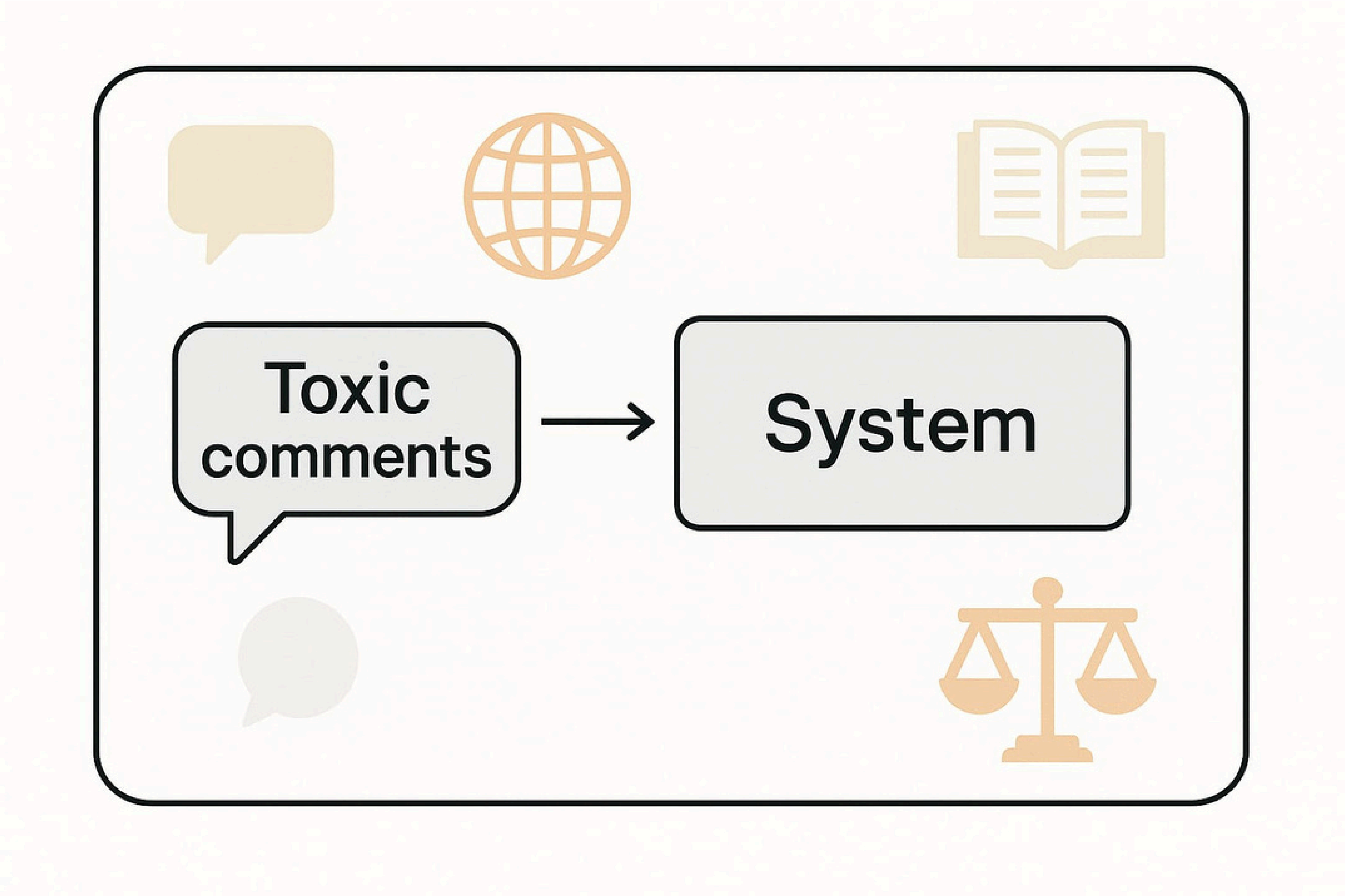
EVALUATE FAIRNESS

OFFER COMPREHENSIVE METRICS

SYSTEM ENVIRONMENT

Target users: Social platform moderators, AI ethics researchers, developers of automatic moderation systems.

External factors: linguistic changes, cultural differences, data sensitivity, legislative changes regarding freedom of expression.



CHAOTIC ATTRACTORS

ML models, especially in NLP tasks, are full of these situations.

- Language variability: comments can contain irony, sarcasm, puns, insults disguised as jokes, etc. The model often can’t capture these subtleties.
- Comments with a lot of ”noise.”
- Subjectivity of Toxicity
- Data Biases: training data may contain biases that reflect real prejudices.
- Interaction between participants: each person or team tries new things, and this means the environment is constantly changing.

SOLUTION APPROACH

MODULAR ARCHITECTURE:

Decoupled components (TextCleaner, IdentityAttackChecker) for maintainability.

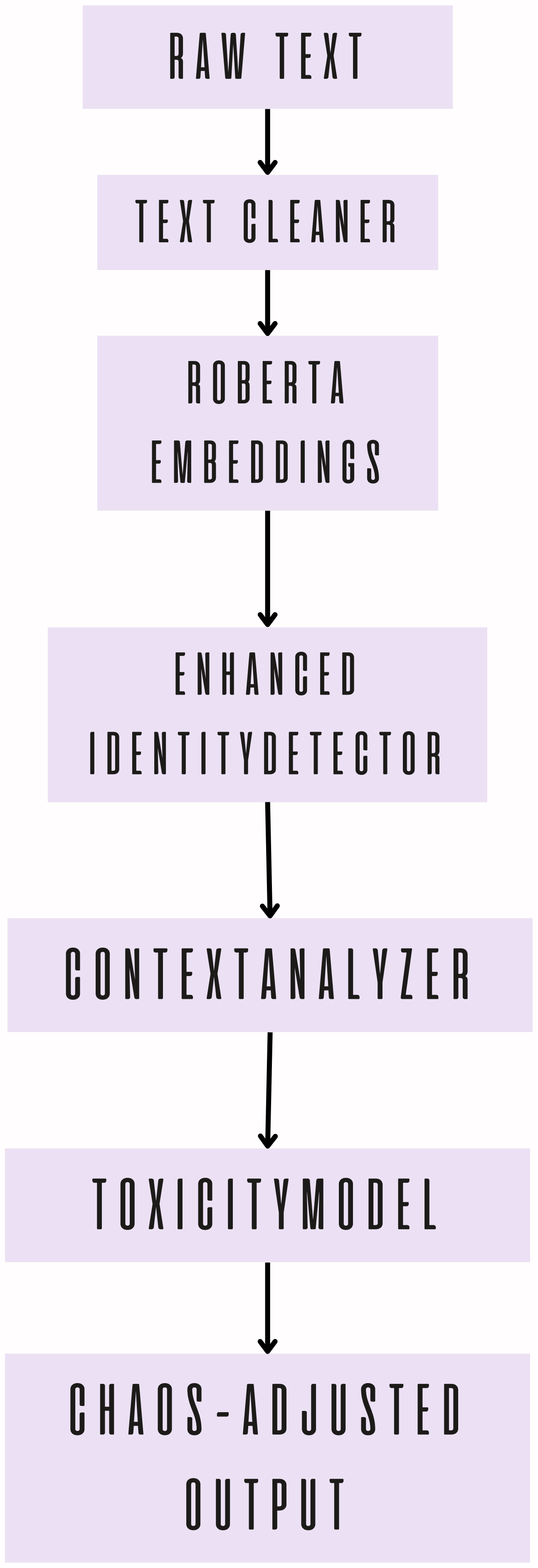
CHAOS MANAGEMENT

Specialized modules (AnnotatorWeightCalculator) to handle annotator subjectivity and (IdentityAttackChecker) detects subtle harmful patterns that are not explicitly labeled

LARGE LANGUAGE MODEL

RoBERTa embeddings

SYSTEM ARCHITECTURE



RAW TEXT

TEXT CLEANER

ROBERTA EMBEDDINGS

ENHANCED IDENTITYDETECTOR

CONTEXTANALYZER

TOXICITYMODEL

CHAOS-ADJUSTED OUTPUT

TECHNOLOGIES USED AND IMPLEMENTATION FRAMEWORK

Technology Stack

Languaje: python (selected for AI/ML ecosystem and Kaggle compatibility)
Key libraries:

- Pandas (low-level mathematical operations)
- Numpy (transformation of the train.csv, test.csv, and individual annotation files)
- Scikit-learn (implement linear regression models, such as logistic regression)

TECHNOLOGIES USED AND IMPLEMENTATION FRAMEWORK

Modeling Strategy

The modeling approach was initially based on regression models due to their simplicity and transparency. These models were integrated into a modular framework, designed with software and systems engineering principles and focused on scalability and maintainability.

DATA INGESTION MODULE

PREPROCESSING MODULE

MODELING AND EVALUATION MODULE

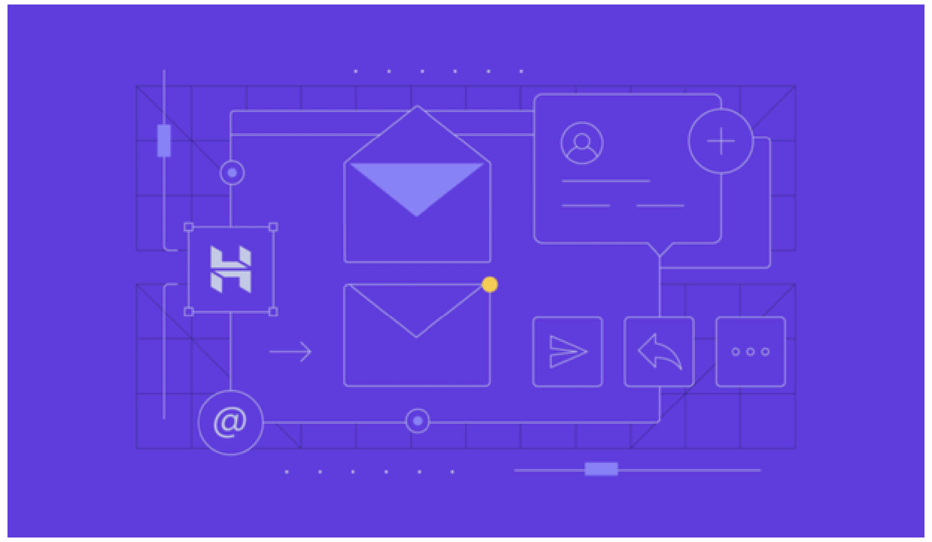
CHAOS AND SENSITIVITY MODULE

SUBMISSION MODULE

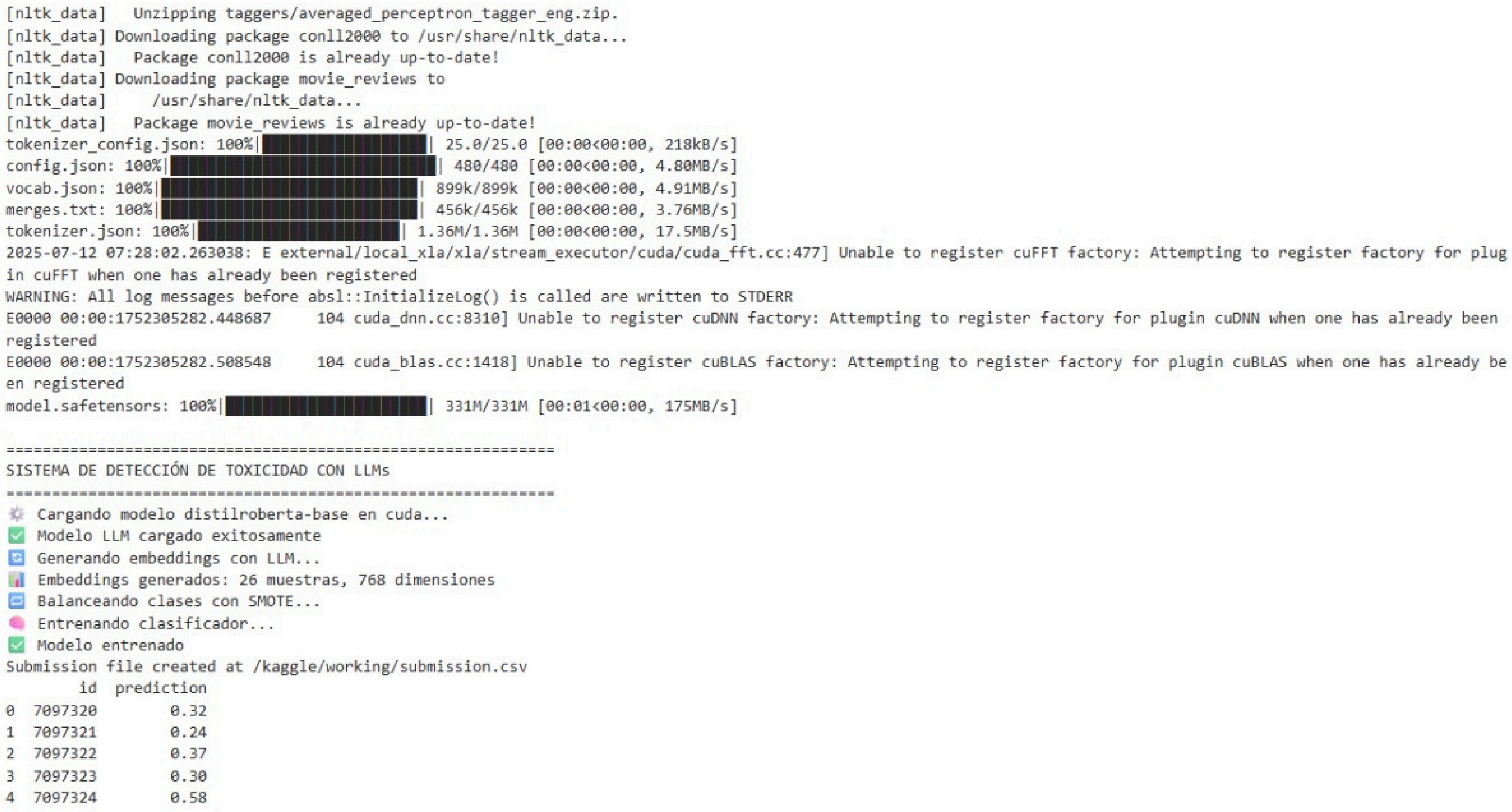
LLM INTEGRATION

Large Language Model

Models such as BERT and RoBERTa introduced the use of contextual representations, in which the meaning of a word depends on the context in which it appears. These models detect complex patterns in something as subjective as language, i.e., irony and identity attacks. In addition, they have been shown to perform better on tasks where fairness is important, since they are better at distinguishing between toxic language and identity-neutral mentions.

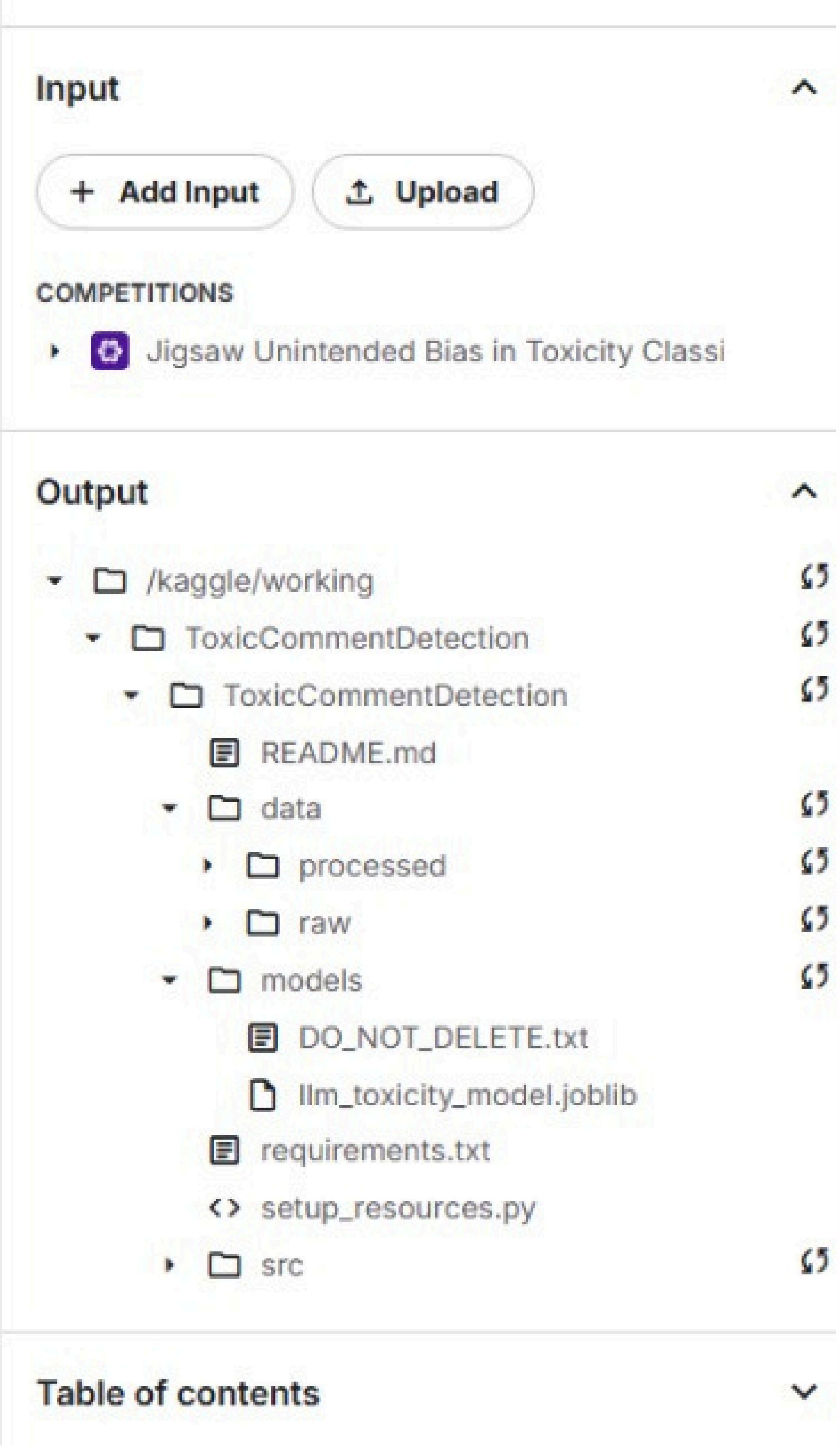


RESULTS



```
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger_eng.zip.
[nltk_data] Downloading package conll2000 to /usr/share/nltk_data...
[nltk_data]   Package conll2000 is already up-to-date!
[nltk_data] Downloading package movie_reviews to
[nltk_data]     /usr/share/nltk_data...
[nltk_data]   Package movie_reviews is already up-to-date!
tokenizer_config.json: 100%|██████████████████| 25.0/25.0 [00:00<00:00, 218kB/s]
config.json: 100%|█████████████████████████████| 480/480 [00:00<00:00, 4.80MB/s]
vocab.json: 100%|████████████████████████████| 899k/899k [00:00<00:00, 4.91MB/s]
merges.txt: 100%|████████████████████████████| 456k/456k [00:00<00:00, 3.76MB/s]
tokenizer.json: 100%|██████████████████████| 1.36M/1.36M [00:00<00:00, 17.5MB/s]
2025-07-12 07:28:02.263038: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plug
in cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1752305282.448687     104 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been
registered
E0000 00:00:1752305282.508548     104 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already be
en registered
model.safetensors: 100%|██████████████████████| 331M/331M [00:01<00:00, 175MB/s]

============================================================
SISTEMA DE DETECCIÓN DE TOXICIDAD CON LLMs
============================================================
⚙️ Cargando modelo distilroberta-base en cuda...
✅ Modelo LLM cargado exitosamente
🔄 Generando embeddings con LLM...
📊 Embeddings generados: 26 muestras, 768 dimensiones
🔁 Balanceando clases con SMOTE...
🧠 Entrenando clasificador...
✅ Modelo entrenado
Submission file created at /kaggle/working/submission.csv
        id  prediction
0  7097320        0.32
1  7097321        0.24
2  7097322        0.37
3  7097323        0.30
4  7097324        0.58
```



COMPETITIONS

- Jigsaw Unintended Bias in Toxicity Classi

```
/kaggle/working
  ToxicCommentDetection
    ToxicCommentDetection
      README.md
      data
        processed
        raw
      models
        DO_NOT_DELETE.txt
        llm_toxicity_model.joblib
      requirements.txt
      setup_resources.py
      src
```

CONCLUSIONS OF THE FINAL PROJECT

- LLMs Work Better: Context-aware models clearly outperformed linear baselines.
- Design Matters: A modular system helped us isolate failures and scale.
- Infra Matters More: A model that can’t deploy is still a failed submission.
- Future = Hybrid: LLMs should trigger only on ambiguity.

What We Learned: · Build with constraints first (RAM, runtime, internet) · Log everything: silent errors waste time · Embrace progressive upgrades, not complete overhauls

2025

THANK YOU

system analysis and design