The National School of
Artificial Intelligence
المدرسة الوطنية العليا للذكاء الاصطناعي

HACKATHON : DÉVELOPPEMENT D'UNE SOLUTION DE FILTRATION DES SUGGESTIONS D'ACTIVITÉS DE L'AUTO-ENTREPRENEURIAT

TEAM 4:

MERIEM GHORAB
SARA BENALI
AYA MOUFFOK
SERINE NOURELIMANE HACENE
IMEN NOURELHOUDA BOUDJIRA
MARWA SALI

This project presents a comprehensive and innovative approach to filtering and classifying textual data based on acceptance criteria. The dataset consists of multilingual text entries in French and Arabic, necessitating a robust strategy for semantic understanding. By leveraging advanced Natural Language Processing (NLP) techniques, we perform clustering and classification to distinguish between accepted and rejected data points, ensuring efficiency and accuracy. This work r epresents a remarkable achievement in AI-driven text processing, combining precision with scalability.

OUR OBJECTIVES

CLUSTERING

Automatically group similar sentences to identify inherent patterns

CLASSIFICATION

Train a machine learning model to categorize text entries as either "accepted" or "rejected."

MULTILINGUAL PROCESSING

Ensure that the model accurately interprets and processes multiple languages within the dataset.

APPROACH

1. DATA ACQUISITION & PREPROCESSING

We have meticulously gathered and structured our dataset, which comprises rejecting data in a dedicated file and an additional dataset containing various states (accepting, rejecting, unknown, etc.). Since the dataset initially lacked clear labels, we systematically assigned acceptance and rejection labels using Python scripts. Given the presence of multiple languages, preprocessing involved normalization, handling missing values, and ensuring consistency across all text entries. Our dedication to this step laid a strong foundation for the success of the entire project.

2. SENTENCE EMBEDDINGS

To capture the semantic essence of each sentence, we employ a multilingual Sentence-BERT model capable of generating high-quality sentence embeddings. This step ensures uniform representation across our database, allowing us to preserve contextual meaning while transforming textual data into numerical form. Our commitment to leveraging state-of-the-art NLP ensures that even nuanced linguistic differences are effectively understood by our model.

3. CLUSTERING

Using an unsupervised clustering algorithm, we group similar sentences together based on their semantic similarities. This step aids in uncovering latent structures in the dataset and ensures that classification is performed on meaningful clusters rather than isolated data points. Our clustering process brings order to complex multilingual data, enhancing the precision of subsequent classification.

4. LABELING & DATASET PREPARATION

The dataset is enriched with carefully assigned labels—"accepted" for desirable entries and "rejected" for undesirable ones. These labels serve as the foundation for supervised learning, enabling the model to differentiate between acceptable and unacceptable data with confidence. Our meticulous labeling process ensures the integrity and reliability of the dataset, empowering our AI model to make informed decisions.

5. MODEL TRAINING

A machine learning classifier is trained using the labeled embeddings. The model learns to distinguish between the two categories based on the patterns found in the training data. We utilized a comprehensive dataset containing 3,000 real-world suggestions to train the model, ensuring its robustness and functionality. Through rigorous training and evaluation techniques, we refined the classifier to deliver high accuracy and reliability.

6. CLASSIFICATION OF NEW DATA

Once trained, the classifier is deployed to assess new, unlabeled textual entries. Each new record is transformed into an embedding and classified based on its learned patterns, allowing for automated and intelligent filtering of content. This step represents t he true power of AI—providing rapid, data-driven insights with minimal human intervention.

7. WEB INTERFACE FOR DATA PROCESSING

To enhance usability, we developed a web-based interface where an admin can securely log in using a username (admin)and password(admin). Once authenticated, the admin can upload data files, triggering the AI model to process and classify the content.
The system then displays the results, offering a seamless and intuitive experience for managing and filtering textual data efficiently.
This multilingual AI-driven system is a testament to innovation, diligence, and technological
excellence. By integrating clustering, classification, and multilingual NLP techniques, we have created a powerful, scalable, and intelligent solution for filtering and categorizing textual data. Our work not only ensures accuracy but also adapts seamlessly to diverse linguistic datasets, setting a high standard in AI-driven text analysis.

Through dedication and expertise, we have built a system capable of dynamicall analyzing and filtering vast amounts of text with unparalleled precision. Additionally, our web-based platform ensure s managing and classifying data, making AI-driven text processing more accessible than ever before. This achievement reflects not only our technical proficiency but also our commitment to solving complex problems with creativity and innovation. Together, we have accomplished something truly remarkable.

---

8. OPTIMIZATION & FINE-TUNING

To enhance accuracy, hyperparameter tuning and model refinement are applied. Further optimizations, such as experimenting with different classification algorithms, fine-tuning the embedding model, and adjusting clustering parameters, are undertaken to maximize efficiency. This continuous improvement mindset ensures that our system remains cutting-edge and adaptable to evolving data challenges.

This multilingual AI-driven system is a testament to innovation, diligence, and technological excellence. By integrating clustering, classification, and multilingual NLP techniques, we have created a powerful, scalable, and intelligent solution for filtering and categorizing textual data. Our work not only ensures accuracy but also adapts seamlessly to diverse linguistic datasets, setting a high standard in AI-driven text analysis.

Through dedication and expertise, we have built a system capable of dynamicall analyzing and filtering vast amounts of text with unparalleled precision. Additionally, our web-based platform ensure s managing and classifying data, making AI-driven text processing more accessible than ever before. This achievement reflects not only our technical proficiency but also our commitment to solving complex problems with creativity and innovation. Together, we have accomplished something truly remarkable.