Fully Automated AI-Powered System for Filtering and Categorizing Startup Activity Proposals

Group 5• Feb 22, 2025

ensia x

الوكالة الوطنية
للمقاول الذاتي

Team members	Phone Number	Email
Lilia Ammar Khodja	0779989558	lilia.ammarkhodja@ensia.edu.dz
ACHOURI Anfal	0559169390	anfal.achouri@ensia.edu.dz
ARAB Sarra	0540221667	sarra.arab@ensia.edu.dz
RAHMOUNI Rahil	0549673850	rahil.rahmouni@ensia.edu.dz
Imene Nour El Houda Maarafi	0669334891	imene.nour.el.houda.maarfi@ensia.edu.dz

1. Welcome!

In an era of rapid technological advancement and entrepreneurial innovation, our project aims to create an AI-powered system that efficiently filters and categorizes startup activity proposals submitted by auto-entrepreneurs. This system seeks to identify novel business ideas while automatically excluding existing and commercial activities, fostering a more dynamic startup ecosystem.

We began with a diverse dataset containing activity descriptions in multiple languages—Arabic and French—organized into four distinct subsets: Arabic-Arabic (AA), French-French (FF), French-Arabic (FA), and Mixed (French & Arabic in the same cell). Each subset underwent comprehensive preprocessing and normalization, followed by embedding-based similarity checks to match existing activities from official lists. The result was a refined dataset of unique startup ideas, further filtered to exclude commercial activities using a carefully curated keyword list.

The validated activities were then clustered based on similarity, generating potential categories for admin review. To enhance user engagement and streamline the submission process, we developed a web-based application where users can propose their startup ideas. Admins can easily manage submissions, review categories, and upload files for batch processing.

This project exemplifies the integration of AI-driven automation with human oversight, paving the way for innovative business proposals that can contribute to economic growth and diversification.

For more Context here is the Video demo of our final product :

🎬 FULLY AUTOMATED MINISTERY ACTIVITIES FILTERING SYSTEM.…

2. Dataset Description

The dataset used in this project comprised activities and their descriptions sourced from various auto-entrepreneurs. It was characterized by its multilingual nature, with entries in Arabic, French, or a combination of both. To facilitate our analysis, we split the dataset into four subsets:

- **AA (Arabic-Arabic)**: Activities described solely in Arabic.
- **FF (French-French)**: Activities described solely in French.
- **FA (French-Arabic)**: Activities with one description in French and another in Arabic.
- **Mixed**: Entries containing both French and Arabic within the same cell.

This categorization allowed us to apply tailored preprocessing techniques for each subset.

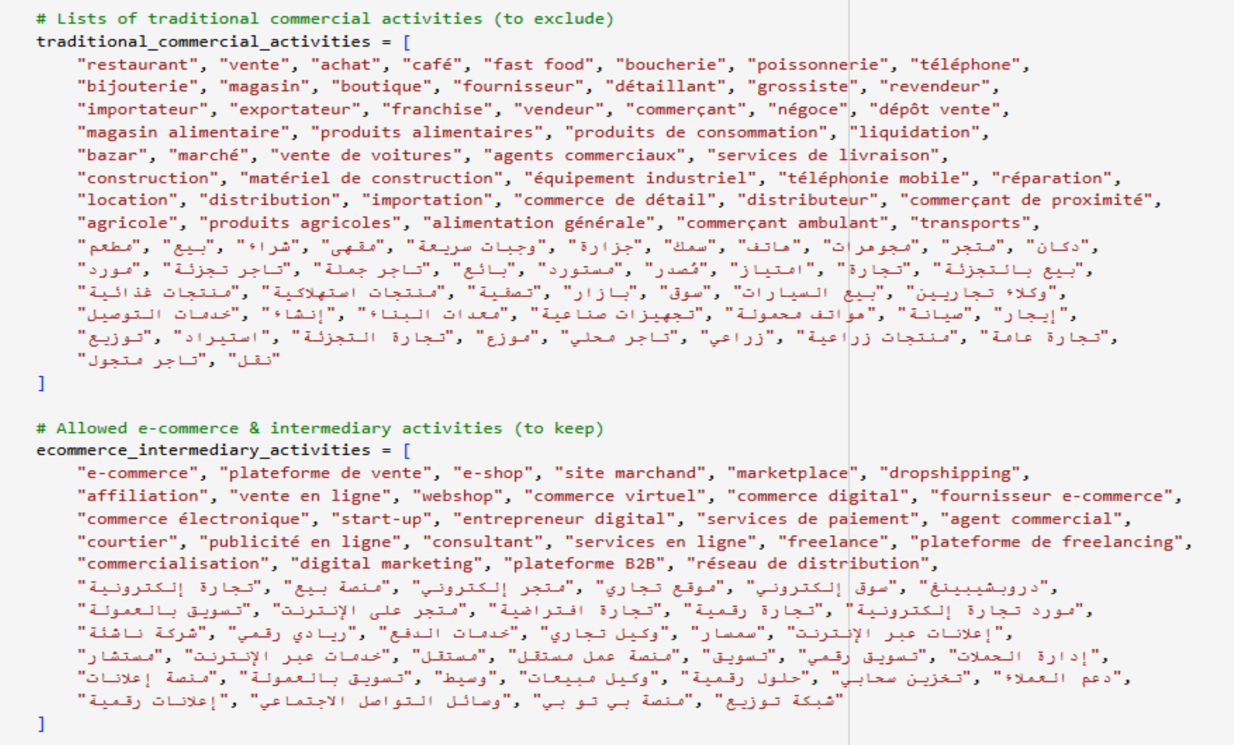
3. Methodology

The methodology employed in our project included several key steps:

3.1. Preprocessing and Normalization Each dataset subset underwent preprocessing to standardize the text format, including normalization (e.g., removing punctuation, converting to lowercase) and tokenization. This step ensured consistency across all entries, facilitating more accurate analysis.

3.2. Embedding and Similarity Matching We employed different embedding models tailored to each language subset . “**paraphrase-multilingual-MiniLM-L12-v2”** and “**Alibaba-NLP/gte-multilingual-base for the rest “.** These models allowed us to transform textual data into numerical representations, enabling effective similarity checks. By comparing the processed entries against official activity lists, we identified and removed activities already recognized by the ministry.

3.3. Commercial Activity Filtering After eliminating existing activities, we focused on refining a list of commercial keywords to filter out traditional business proposals. Exceptions, such as e-commerce-related terms, were included to avoid unjust exclusion of valid entries.

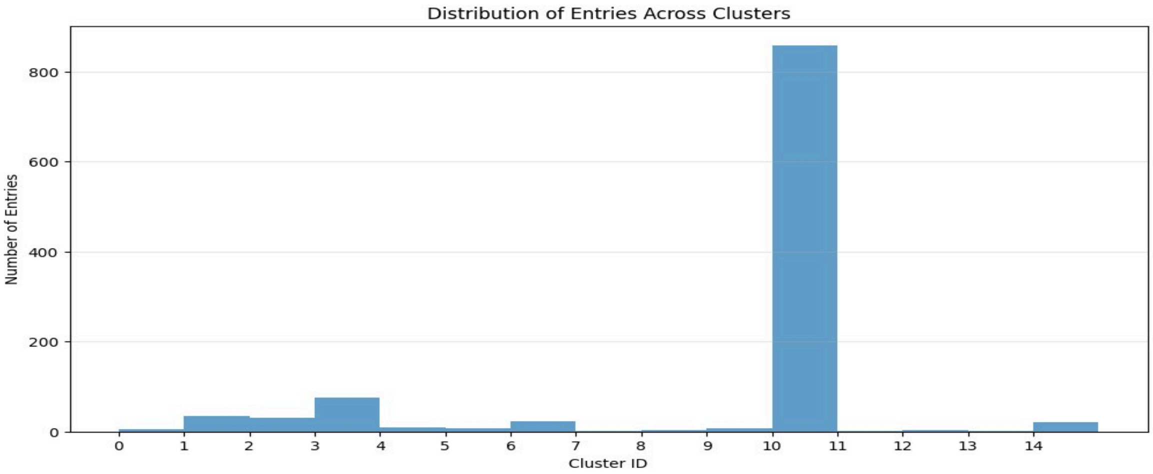


```
# Lists of traditional commercial activities (to exclude)
traditional_commercial_activities = [
    "restaurant", "vente", "achat", "café", "fast food", "boucherie", "poissonnerie", "téléphone",
    "bijouterie", "magasin", "boutique", "fournisseur", "détaillant", "grossiste", "revendeur",
    "importateur", "exportateur", "franchise", "vendeur", "commerçant", "négoce", "dépôt vente",
    "magasin alimentaire", "produits alimentaires", "produits de consommation", "liquidation",
    "bazar", "marché", "vente de voitures", "agents commerciaux", "services de livraison",
    "construction", "matériel de construction", "équipement industriel", "téléphonie mobile", "réparation",
    "location", "distribution", "importation", "commerce de détail", "distributeur", "commerçant de proximité",
    "agricole", "produits agricoles", "alimentation générale", "commerçant ambulant", "transports",
    "مطعم", "بيع", "شراء", "مقهى", "وجبات سريعة", "جزارة", "سمك", "هاتف", "مجوهرات", "متجر", "دكان",
    "مورد", "تاجر تجزئة", "تاجر جملة", "بائع", "مستورد", "مُصدر", "امتياز", "تجارة", "بيع بالتجزئة",
    "منتجات غذائية", "منتجات استهلاكية", "تصفية", "بازار", "سوق", "بيع السيارات", "وكلاء تجاريين",
    "خدمات التوصيل", "إنشاء", "معدات البناء", "تجهيزات صناعية", "هواتف محمولة", "صيانة", "إيجار",
    "توزيع", "استيراد", "تجارة التجزئة", "موزع", "تاجر محلي", "زراعي", "منتجات زراعية", "تجارة عامة",
    "تاجر متجول", "نقل"
]

# Allowed e-commerce & intermediary activities (to keep)
ecommerce_intermediary_activities = [
    "e-commerce", "plateforme de vente", "e-shop", "site marchand", "marketplace", "dropshipping",
    "affiliation", "vente en ligne", "webshop", "commerce virtuel", "commerce digital", "fournisseur e-commerce",
    "commerce électronique", "start-up", "entrepreneur digital", "services de paiement", "agent commercial",
    "courtier", "publicité en ligne", "consultant", "services en ligne", "freelance", "plateforme de freelancing",
    "commercialisation", "digital marketing", "plateforme B2B", "réseau de distribution",
    "تجارة إلكترونية", "منصة بيع", "متجر إلكتروني", "موقع تجاري", "سوق إلكتروني", "دروبشيبينغ",
    "تسويق بالعمولة", "متجر على الإنترنت", "تجارة افتراضية", "تجارة رقمية", "مورد تجارة إلكترونية",
    "شركة ناشئة", "ريادي رقمي", "خدمات الدفع", "وكيل تجاري", "سمسار", "إعلانات عبر الإنترنت",
    "مستشار", "خدمات عبر الإنترنت", "مستقل", "منصة عمل مستقل", "تسويق", "تسويق رقمي", "إدارة الحملات",
    "منصة إعلانات", "تسويق بالعمولة", "وسيط", "وكيل مبيعات", "حلول رقمية", "تخزين سحابي", "دعم العملاء",
    "إعلانات رقمية", "وسائل التواصل الاجتماعي", "منصة بي تو بي", "شبكة توزيع"
]
```

This approach ensured that only innovative startup ideas remained in the dataset.

3.4. Clustering and Categorization The filtered dataset was then subjected to clustering algorithms (AgglomerativeClustering with cosine distance) to group similar activities. This process facilitated the generation of potential category names (By an LLM groq), which were presented to an admin for review. The admin's validation process helps ensure that only relevant categories are accepted.



4. Results

The implementation of our system yielded a refined dataset of unique startup activities, alongside generated categories that reflected similar ideas. Admins are able to review and approve categories, leading to a well-organized system for managing submissions. The final output not only showcases innovative business ideas but also demonstrates the effectiveness of our AI-driven approach.

5. Web Application Development

To complement our methodology, we developed a user-friendly web-based application. This platform enables auto-entrepreneurs to submit their startup ideas through an intuitive form. Admins have access to a dashboard where they can view submissions, validate categories, and upload files for batch processing. This integration of AI with human review provides a comprehensive solution for managing startup proposals efficiently.

6. Conclusion

Our project successfully demonstrates the potential of AI technology in enhancing the startup proposal process. By combining automated filtering and categorization with human oversight, we create a robust system that not only identifies innovative business ideas but also fosters a supportive environment for aspiring entrepreneurs. Moving forward, we aim to refine our model further and explore additional features that could enhance user engagement and improve the overall experience.

7. Drive

Link to the project drive with every document and file used

ANAE_Hackathon_group5