M1SOL023 Méthodologie de la Recherche

Langue et Informatique — 2025–2026

Projet : Classification et Modèles Multimodaux

Responsables : Laurence Devillers & Nour El Houda Ben Chaabene, Sorbonne Université

Résumé du projet

L'objectif de ce projet est de conduire une expérience complète de **classification automatique** en partant d'un jeu de données choisi (texte et/ou audio). Le travail consistera à mettre en place une chaîne de traitement allant de la préparation des données jusqu'à l'évaluation des résultats, en explorant à la fois des méthodes classiques et des modèles récents de type LLM ou multimodaux.

Le projet (rapport + code) constituera la moitié de la note de contrôle continu et donnera lieu à un oral pendant la session de janvier 2026.

Objectifs

Le projet vise à :

— Mettre en place une chaîne complète de classification automatique : préparation, représentation, apprentissage, évaluation.

— Expérimenter avec des méthodes classiques (`scikit-learn`) et modernes (Flaubert/BERT pour le texte, Wav2Vec pour l'audio, combinaison multimodale).

— Comparer les performances obtenues selon les représentations et les algorithmes utilisés.

— Analyser de manière critique les résultats et proposer des pistes d'amélioration ou d'extension.

Méthodologie

Le travail attendu comprend les étapes suivantes :

1. Préparation et nettoyage du jeu de données choisi (instances X, classes y).
2. Extraction de représentations adaptées : sac de mots, CountVectorizer, embeddings (LLM), représentations audio (Wav2Vec).
3. Mise en place et entraînement de différents classifieurs (Naïve Bayes, arbres, forêts, SVM, etc.).
4. Exploration de la fusion multimodale (texte + audio, par exemple via Whisper).
5. Évaluation des performances (exactitude, précision, rappel, F-mesure, matrices de confusion).
6. Discussion critique : identification des erreurs typiques, analyse des limites, ouverture vers d'autres corpus, langues ou modalités.

Plan conseillé du rapport

Le rapport attendu (environ 10 pages hors annexes) comportera au minimum les sections suivantes :

Introduction
Présentation de la tâche, du jeu de données et des objectifs.

État de l'art
Travaux existants sur la classification automatique, l'usage de LLMs et les approches multimodales.

Corpus
Description détaillée du jeu de données choisi : taille, classes, constitution, statistiques, exemples.

Méthode
Chaîne de traitement mise en œuvre : représentations, algorithmes, paramètres, choix méthodologiques.

Résultats
Présentation chiffrée et commentée des expériences, comparaisons, analyse des erreurs.

Conclusion et perspectives
Résultats principaux, limites, pistes futures.

Annexes
Organisation du code, consignes techniques, exemples complémentaires.

Rendu attendu

— Rapport en PDF (environ 10 pages hors annexes).

— Code Python en `.ipynb` et/ou `.py`, clairement documenté.

— Rendu final : 04/01/2026 sur Moodle (code + rapport).