Lignes directrices pour la collecte de données et la conformité au droit d'auteur

Projet : Modèle de prédiction de suicidabilité chez les poètes

José Manuel Rodríguez Caballero

January 28, 2025

1 Introduction

Le présent rapport décrit les principes et les procédures pour la collecte de données nécessaires à l'entraînement d'un modèle d'apprentissage automatique capable de prédire si un poète est "suicidaire" ou non, à partir d'informations liées à ses poèmes et à ses caractéristiques démographiques.

Contrairement à certaines approches qui utilisent des techniques de fouille de texte (*text mining*), nous soulignons qu'**aucun text mining automatisé** ne sera mis en œuvre. Tous les poèmes seront **extraits manuellement** de livres empruntés en bibliothèque, afin de recueillir les variables nécessaires (*e.g.* nombre de mots, mesures de sentiment, etc.).

Le rapport explique:

- Les modalités de collecte des poèmes en accord avec le droit d'auteur.
- Les informations à extraire de chaque poème et du poète.
- Les aspects **éthiques** et **légaux** de la démarche.

2 Objectif du projet

Notre but est de développer un modèle supervisé (*machine learning*) prédictif, qui puisse estimer la probabilité qu'un poète présente un risque de suicidabilité, sur la base des informations suivantes:

- **Données démographiques et biographiques** de l'auteur (*genre*, *année de naissance*, *année de décès ou de suicide*, etc.).
- **Caractéristiques agrégées des poèmes** (ex. nombre total de mots, mesures de sentiment, diversité lexicale, etc.).

Le projet **n'exploite pas** le texte intégral des poèmes pour le stockage ou la diffusion publique, afin de respecter le droit d'auteur. Seules des mesures statistiques et métriques agrégées sont recueillies et enregistrées pour l'analyse.

3 Collecte de données

3.1 Sources des poèmes

Les poèmes sont consultés dans des livres **empruntés dans des bibliothèques**. Ainsi, chaque poème est extrait **manuellement** par l'équipe, qui relève uniquement des informations structurées:

- Titre du poème, titre du livre ou recueil.
- Nom de l'auteur, date de naissance et éventuelles autres informations biographiques (si disponibles).
- Nombre de mots, nombre de lignes, etc.
- Résultats d'analyse sémantique ou de sentiment **effectués à la main** ou via un logiciel dédié *(localement)*, sans conservation du texte intégral.

3.2 Respect du droit d'auteur

Dans le cadre de ce projet, nous nous assurons:

1. **D'obtenir légalement l'accès aux livres:**
   Les ouvrages sont consultés dans le strict respect des règles de la bibliothèque. Nous n'effectuons aucune reproduction ou numérisation massive qui serait contraire à la législation.
2. **De ne pas diffuser le texte intégral:**
   Seules des **extractions agrégées** (comptages, statistiques) seront enregistrées dans notre base de données. Les poèmes complets, même s'ils sont saisissables à la main, ne sont **pas** inclus dans les données finales.
3. **D'utiliser uniquement les mesures nécessaires:**
   Pour l'apprentissage automatique, nous ne conservons que les **indicateurs utiles** (ex. sentiment moyen, proportion de mots négatifs, etc.). Nous ne conservons pas de larges extraits ou de copies du poème susceptibles de violer le droit d'auteur.

4 Données à recueillir et formalisation

4.1 Données sur le poète

- **Identifiant unique** (numérique ou pseudonyme) permettant de relier plusieurs poèmes au même poète.
- **Nom du poète** (ou anonymisé si nécessaire).
- **Genre** (M, F, Non-binaire, *autre*).
- **Année de naissance**.
- **Année de suicide** si applicable.
- **Statut suicidaire** (oui/non) pour la classification.

4.2 Données sur chaque poème

- **Titre du poème** et **titre du livre** ou recueil.
- **Nombre de mots** (estimé ou compté).
- **Nombre de lignes** et autres métriques (ex. longueur moyenne de ligne).
- **Scores de sentiment** (ex. entre -1 et +1) ou **scores d'émotions fines** (ex. colère, joie, tristesse, exprimés en pourcentage).
- **Indicateurs de variabilité** (ex. écart-type du sentiment par ligne, *sentiment slope*).

Toutes ces variables sont suffisamment anonymes pour ne pas enfreindre le droit d'auteur, dans la mesure où elles ne permettent pas de reconstituer directement le texte intégral.

5 Exploitation des données pour la prédiction

5.1 Modèle et pipeline

1. **Constitution de la base de données**:
   Une table reliera chaque poète à ses poèmes (par *poet_id* et *poem_id*). Les champs agrégés (sentiment, etc.) seront renseignés.
2. **Séparation entraînement/test**:
   Nous divisons aléatoirement l'ensemble des poèmes en un échantillon d'entraînement et un échantillon de test.

3. **Extraction de caractéristiques**:
   Les variables pertinentes (ex. sentiment moyen, score de diversité lexicale, genre du poète, etc.) sont transformées ou encodées (éventuellement normalisées).

4. **Entraînement du classifieur**:
   Nous pourrons utiliser un algorithme de classification (Random Forest, SVM, réseau de neurones, etc.) afin de distinguer les poètes "suicidaires" de ceux qui ne le sont pas.

5. **Évaluation et interprétation**:
   Les performances seront mesurées via la précision, le rappel, la F-mesure ou encore l'AUC (surface sous la courbe ROC). Nous analyserons également les **caractéristiques les plus influentes** (importance des variables).

5.2 Aspects éthiques et limitations

- **Sujet sensible**: L'étude de la suicidabilité peut engendrer des interprétations délicates. Le modèle n'a pas vocation à poser un diagnostic médical.
- **Variabilité historique et culturelle**: Les poètes sélectionnés peuvent varier selon l'époque, la culture, et la langue. Les conclusions ne sont pas nécessairement généralisables à l'ensemble des populations.
- **Respect de la vie privée et de la mémoire**: Les données ne doivent pas diffamer ni stigmatiser des auteurs spécifiques. L'anonymisation peut s'avérer nécessaire si les poètes ou leurs ayants droit le demandent.

6 Conclusion

Ce document présente les lignes directrices pour **collecter et traiter** des informations extraites **manuellement** de poèmes empruntés en bibliothèque, sans enfreindre le droit d'auteur. Les grandes idées sont:

- **Limiter la collecte** aux variables agrégées et indispensables au modèle prédictif.
- **Ne pas reproduire** les oeuvres intégrales pour éviter toute violation de la propriété intellectuelle.
- **Privilégier la prudence** sur le plan éthique et légal, compte tenu du sujet sensible (le risque suicidaire).

Les données, une fois rassemblées, alimenteront un modèle d'apprentissage automatique conçu pour *prédire* la suicidabilité d'un poète sur la base de caractéristiques linguistiques

et démographiques. Les résultats devront être interprétés avec **caution** et ne sauraient se substituer à un diagnostic médical.

Référence Principale :

- **Stirman, Shannon Wiltsey, and James W. Pennebaker.** "Word use in the poetry of suicidal and nonsuicidal poets." *Psychosomatic medicine* 63, no. 4 (2001): 517-522.