Description du Jeu de Données, Tâche de Prévision et Lignes Directrices de Collecte

Projet : Modèle de prédiction de suicidabilité chez les poètes

José Manuel Rodríguez Caballero

February 3, 2025

Abstract

Ce document présente les jeux de données (public et privé) relatifs aux poètes et à leur niveau de "suicidabilité", ainsi que les modalités de collecte des poèmes dans le respect du droit d'auteur. Nous décrivons également la tâche de prévision, qui s'appuie sur un ensemble de **caractéristiques numériques** extraites des textes (scores d'émotions, lexiques sentimentaux, etc.). Pour certains champs comme l'orientation sexuelle (*heterosexual*), des valeurs sont manquantes et feront l'objet d'une imputation afin de ne pas exclure d'observations importantes dans le modèle prédictif.

1 Introduction

Le présent document a pour but de décrire :

- **Les jeux de données** utilisés, avec deux versions : privé et public.
- **La tâche de prévision** : modéliser la probabilité qu'un poète présente un risque de suicidabilité, en se basant sur ses caractéristiques démographiques et linguistiques.
- **La collecte** et le **traitement** des textes poétiques (extractions manuelles, création de variables agrégées) dans le respect de la législation sur le droit d'auteur.

Contrairement à certaines approches de fouille de texte (*text mining*) entièrement automatisées, nous veillons à **ne pas** reproduire l'intégralité des poèmes, qui sont protégés par le droit d'auteur. Les données textuelles brutes ne figurent **que** dans le jeu de données *privé*, lequel n'est pas diffusé. Le jeu *public*, quant à lui, ne contient que des mesures numériques extraites (sentiment, longueur moyenne des mots, etc.).

2 Description du Jeu de Données

2.1 Jeu de données privé (`case-control-long.csv`)

Ce fichier contient :

- **Données démographiques** : sexe, orientation sexuelle (hétérosexuel ou non), date de naissance, date de décès, pays d'origine, etc.
- **Textes de poèmes** : le texte intégral pour chaque poème, permettant des analyses approfondies.
- **Étiquettes de suicidabilité** : variable binaire indiquant si le poète est considéré "suicidaire" (TRUE) ou non (FALSE).

Ce jeu de données est **strictement confidentiel**, car il inclut les poèmes complets, protégés par le droit d'auteur. Il n'est utilisé *que* pour le développement du modèle au sein du groupe de recherche et sous la supervision de l'enseignant.

2.2 Jeu de données public (`case-control-clean.csv`)

Afin de respecter le droit d'auteur et de pouvoir partager une partie des informations, nous avons créé une version "nettoyée" :

- **Colonnes démographiques** : identiques (nom, sexe, dates, etc.), à l'exception des champs sensibles qui sont anonymisés ou supprimés si nécessaire.
- **Caractéristiques textuelles agrégées** :
  - **Scores émotionnels** (tristesse, joie, confusion, colère, peur, surprise, dégoût, espoir), obtenus grâce à un *dictionary-based approach*.
  - **Scores de sentiment** (`afinn`, `bing`, `nrc_ratio`).
  - **Score lexical suicidaire** (proportion de mots liés à l'idée de mort, de fin, etc.).
  - **Longueur moyenne** des mots (`mean_length`).
- **Statut suicidaire** (TRUE/FALSE).

Ici, **aucun** texte intégral de poème n'est conservé. Les valeurs manquantes (par exemple l'orientation sexuelle pour certains poètes) seront traitées via des **méthodes d'imputation**, afin d'éviter des biais importants dans l'analyse.

3 Collecte des Données et Respect du Droit d'Auteur

3.1 Sources des poèmes

- **Recueils en bibliothèque** : consultation légale et manuelle (lecture physique des livres).
- **Sites web autorisés** tels que `https://poetryarchive.org/` ou `https://www.poetryfoundation.org/`, pour récupérer **quelques** poèmes nécessaires à l'étude.

3.2 Procédure de collecte

1. **Extraction manuelle** : les poèmes sont copiés *à la main* afin d'en extraire les informations utiles (nombre de mots, sentiment, etc.).
2. **Pas de reproduction intégrale** : seule la base privée contient les textes complets, et **n'est pas** partagée publiquement.
3. **Nettoyage et agrégation** : les mesures de sentiment, d'émotions, de longueur, etc., sont calculées localement, puis conservées comme *features numériques* dans le jeu `case-control-clean.csv`.

4 Extraction des Caractéristiques Textuelles

4.1 Emotion et Sentiment Lexicons

Une **approche par dictionnaires** (*dictionary-based approach*) est mise en place pour calculer plusieurs scores :

- **Émotions de base** : tristesse, joie, confusion, colère, peur, surprise, dégoût, espoir. Chaque mot du poème est comparé à des lexiques d'émotions, puis **pondéré** selon la position de mots de négation ou d'intensification.
- **Lexiques AFINN, Bing, et NRC** :
  - **AFINN** : score de -5 à +5 pour chaque mot (colère, joie, etc.).
  - **Bing** : catégorisation binaire (positif / négatif).
  - **NRC ratio** : ratio (positif - négatif) / total.
- **Score lexical suicidaire** : proportion de mots associés à la mort ou au désespoir (`despair, suicide, death, ...`).

4.2 Longueur moyenne des mots

Pour chaque poème, on calcule également la taille moyenne des tokens. Cette variable peut être indicative du style poétique ou du registre de langue.

4.3 Gestion des données manquantes

Pour la variable `heterosexual` (orientation sexuelle), certaines valeurs sont absentes. Nous appliquerons des **méthodes d'imputation statistique** (e.g. *multiple imputation by chained equations* - MICE) afin de **ne pas** exclure ces poètes des analyses. Dans la mesure où l'orientation peut être pertinente pour étudier d'éventuels biais socioculturels, il importe de compléter ces données prudemment, sans introduire d'erreurs majeures.

5 Tâche de Prévision

5.1 Objectif principal

L'objectif est de **prédire la suicidabilité d'un poète** (variable binaire `suicidal = TRUE/FALSE`) à partir des informations suivantes :

- **Données démographiques** (sexe, orientation — éventuellement imputée, date de naissance, pays, etc.).
- **Caractéristiques textuelles** (scores d'émotions, mesures de sentiment, longueur moyenne, etc.).

5.2 Modélisation

- **Régression logistique hiérarchique** à trois niveaux :
  1. **Pair cas-témoin** : un poète suicidaire (cas) et un poète non suicidaire (témoin), appariés selon la période, le genre, etc.
  2. **Niveau poète** : agrégation de plusieurs poèmes par auteur.
  3. **Niveau poème** : chaque texte constitue une sous-unité d'observation (scores de sentiment, etc.).
- **Autres approches supervisées** : Random Forest, SVM, réseaux de neurones, etc., pour comparer les performances.
- **Gestion des données manquantes** : intégration d'un schéma d'imputation (MICE ou équivalent) pour la variable `heterosexual`, afin de réduire les biais de sélection.

5.3 Pipeline

1. **Prétraitement** : imputation des données manquantes, normalisation des variables si nécessaire.

2. **Séparation Entraînement/Test** : division aléatoire du jeu public (ou privé) en ensembles de modélisation et de validation.

3. **Entraînement du modèle** : régression logistique hiérarchique (ou autre algorithme).

4. **Évaluation** : précision, rappel, F-mesure, AUC (ROC). Interprétation des coefficients ou importance des variables.

6 Aspects Éthiques et Limitations

- **Sujet sensible** : La prédiction de la suicidabilité chez un individu (même un poète décédé) doit être manipulée avec prudence ; la santé mentale n'est pas réductible à de simples variables quantitatives.

- **Risques de mauvaise interprétation** : Les utilisateurs pourraient faire des généralisations hâtives ou transformer ces analyses en jugement de valeur sur la vie d'un auteur.

- **Données manquantes** : L'imputation, bien que utile, reste imparfaite et peut introduire des incertitudes. Les résultats doivent être interprétés en tenant compte de ces biais potentiels.

- **Droit d'auteur** : La non-diffusion du texte intégral demeure impérative pour respecter la législation en vigueur. Seules les mesures agrégées sont publiques.

7 Conclusion

Dans ce document, nous avons présenté :

- Les deux versions du jeu de données, privé (avec textes intégraux) et public (avec features agrégées).

- La **méthodologie de collecte**, fondée sur une extraction manuelle et minutieuse, respectant le droit d'auteur.

- Les **caractéristiques textuelles** dérivées (émotions, sentiment, longueur, etc.).

- L'importance de l'**imputation des valeurs manquantes** (particulièrement pour l'orientation sexuelle) afin de garder un maximum de poètes dans l'analyse.
- Le **modèle de prévision** envisagé (régression logistique hiérarchique) et les possibles alternatives (Random Forest, SVM, etc.).

Cette approche vise à comprendre si la production poétique peut renseigner sur un risque de suicidabilité, tout en **préservant la confidentialité** des oeuvres et le respect dû aux auteurs. Les résultats, quel que soit le taux de réussite du modèle, doivent être interprétés avec **prudence** et ne sauraient se substituer à un diagnostic médical ou psychiatrique.

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.

Projet sur GitHub :

Texte : `https://github.com/josephcmac/STT-7335_texte`

Code : `https://github.com/josephcmac/STT-7335_code`