Table des matières

1 Introduction

Lors de la création d'un dataset à partir de différentes sources, l'un des principaux problèmes consiste à rendre les données suffisamment homogènes - le niveau de suffisance dépendant des objectifs pour lesquels le dataset est construit. Dans le cas des données textuelles, cela s'applique non seulement à un certain nombre d'éléments quantitatifs ou facilement reconnaissables – longueur du texte, langue, etc. – mais aussi à leur propre signification.

Les pages suivantes présentent une approche possible du *pre-processing* des données textuelles collectées à partir de diverses sources, qui permet également de vérifier l'homogénéité du sens des textes contenus dans l'ensemble de données.

2 Les données

Les données concernent le monde *incel* italien. Les incels - abréviation de "involuntary celibates" - sont les membres d'une certaine communauté en ligne. Ils se caractérisent par le fait qu'il s'agit principalement d'hommes blancs hétérosexuels qui se décrivent comme incapables d'avoir un partenaire romantique ou sexuel bien qu'ils en aient envie.

En particulier, les données ont été collectées à partir de trois sources différentes grâce à des techniques de *data scraping* :

- groupe sur Telegram "Azione Incel"
- groupe sur Telegram "FDB"
- forum en ligne "il forum dei brutti" [1]

Les données sont donc constituées des messages postés sur les deux groupes télégrammes et des messages postés sur le forum.

Le dataset à partir duquel commence notre analyse est déjà le résultat de l'union des données de ces différentes sources et contient trois colonnes : id, text, file_label. "id" est un code numérique que nous pouvons ignorer pour le moment. "text" est la chaîne de texte qui contient le contenu du message et "file_label" est une étiquette qui indique la source d'où provient le texte.

1. "le forum des moches" en italien

3 Le processus

Le prétraitement de ces données est divisé en deux étapes principales. La première concerne des aspects plus simples : l'élimination des données nulles et la vérification de la longueur des textes en excluant les valeurs aberrantes. La deuxième concerne le sens de ces textes, afin de vérifier l'homogénéité de leur contenu. Aux fins du présent article, il convient de se concentrer uniquement sur la deuxième phase.

L'objectif final est de vérifier que les textes du dataset, bien que provenant de sources différentes, ont des significations similaires. Pour ce faire, nous avons d'abord transformé les phrases en un vecteur numérique - ce que l'on appelle "embedding" dans le domaine du traitement automatique des langues. Cette transformation a été effectuée à l'aide d'un modèle de sentence embeddings dérivé du modèle de langue BERT. L'architecture de sentence-BERT et ses détails techniques sont présentés dans cet article [2].
En particulier, le modèle utilisé est *sentence-bert-base-italian-xxl-uncased*[3], qui est dérivé du modèle *bert-base-italian-xxl-uncased*. Il s'agit d'un modèle BERT entraîné sur des données italiennes provenant des corpus OPUS et OSCAR.

Pour entrer dans les détails techniques, les embeddings ont été obtenus à l'aide de la librairie python Sentence Transformer. Les résultats ont ensuite été enregistrés dans une matrice numpy de la forme (95442, 768). À ce stade, les techniques de réduction dimensionnelle suivantes ont été utilisées pour visualiser ces données : analyse en composantes principales (PCA) et UMAP. La visualisation de ces données à travers ces technique nous permet de voir s'il existe des *cluster* particuliers en fonction de la source des données. Les représentations graphiques des réductions dimensionnelles sont présentées ci-dessous. La couleur des points dépend de la source des données.

2. Nils Reimers et Iryna Gurevych, *Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks*, août 2019, URL : `https://arxiv.org/abs/1908.10084v1` (visité le 20/04/2024).

3. *nickprock/sentence-bert-base-italian-xxl-uncased · Hugging Face*, janv. 2024, URL : `https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased` (visité le 20/04/2024).

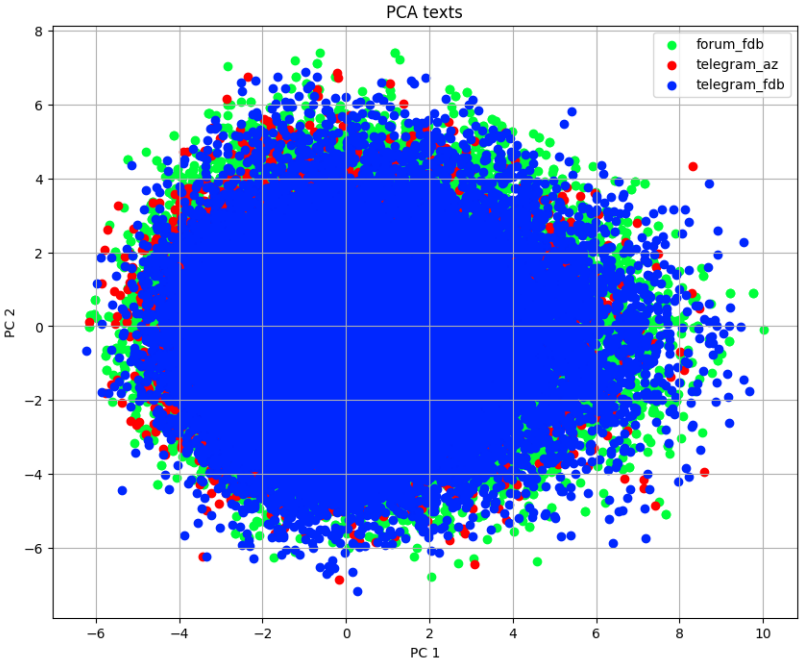

FIGURE 1 – Plot des données réduites par PCA

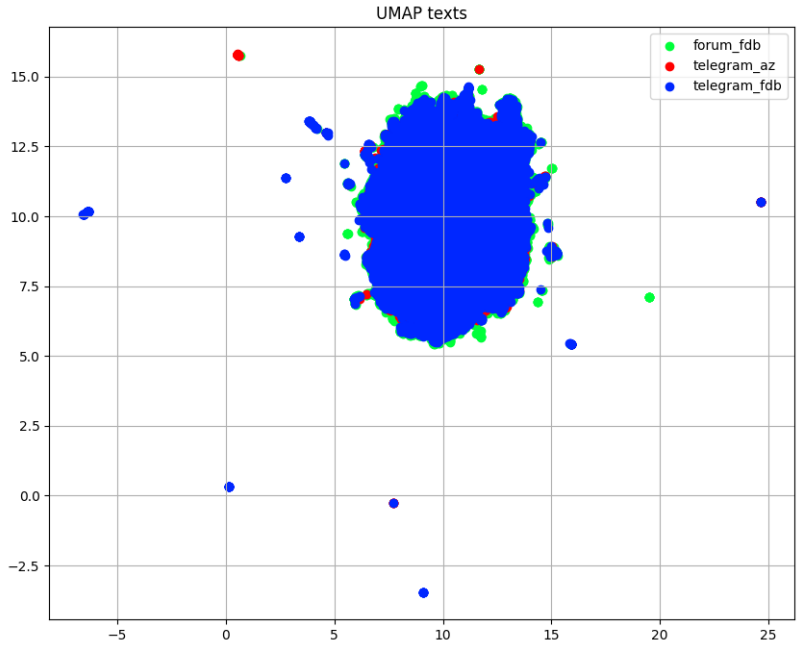

FIGURE 2 – Plot des données réduites par UMAP

Comme on peut le voir, les graphiques ressemblent à une multitude de points sans *pattern* particuliers. Tant dans le cas de la PCA que dans celui de l'UMAP, il n'y a pas de cluster en fonction de la couleur. Cela montre que les textes ont une signification homogène quelle que soit la source.

Références

nickprock/sentence-bert-base-italian-xxl-uncased · Hugging Face, janv. 2024, URL : https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased (visité le 20/04/2024).

Reimers (Nils) et Gurevych (Iryna), *Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks*, août 2019, URL : https://arxiv.org/abs/1908.10084v1 (visité le 20/04/2024).