ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Francesco Paolo Savatteri

Avril 2024

Pre-processing de données textuelles collectées de plusieurs sources

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.

Tout le code utilisé est disponible en ligne à cette adresse : `https://github.com/savaij/devoir_TAL`

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

1. "le forum des moches" en italien

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.

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).

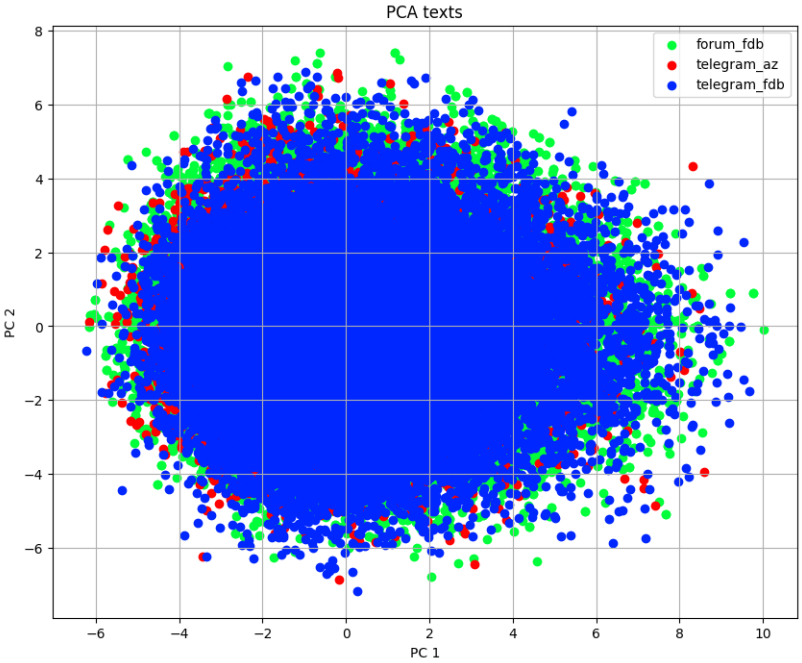

FIGURE 1 – Plot des données réduites par PCA

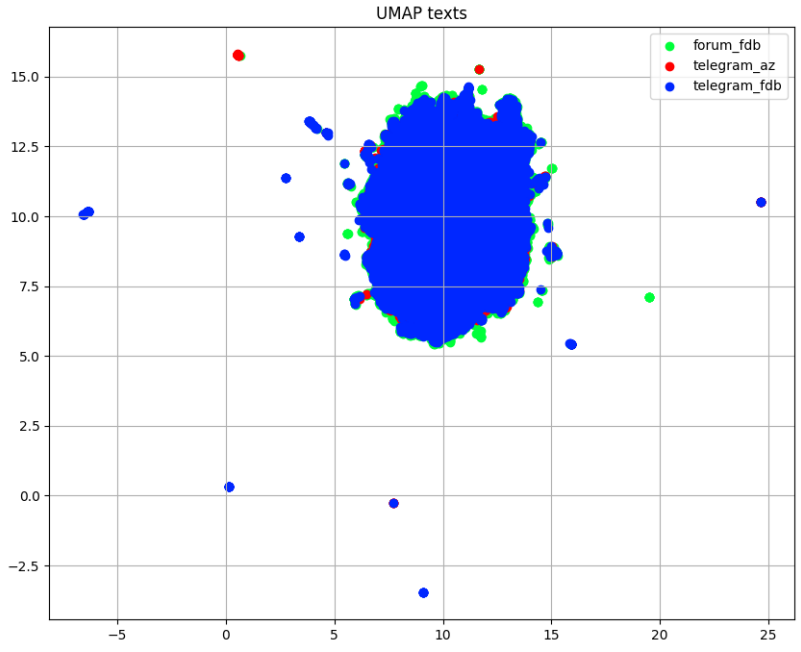

FIGURE 2 – Plot des données réduites par UMAP

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 leur source.

4 Le contrôle

Pour être sûrs que nos résultats sont fiables - et que ce n'est pas, par exemple, le modèle linguistique qui ne fonctionne pas - nous devons effectuer des tests de contrôle. Il s'agit essentiellement de comparer les textes Incel avec des textes qui n'ont rien à voir avec eux, pour comprendre comment cela affecte leur visualisation.

Dans ce cas, on a utilisé un jeu de données contenant des résumés d'articles du journal italien "ilpost" [4]. Il a été créé par le *Applied Recognition Technology Laboratory* (Arte-Lab), un laboratoire de recherché rattaché à l'Université de l'Insubrie (Varese, Italie). En particulier, on a utilisé la partie "train" du dataset, qui correspond 35.201 titres. À l'aide du même modèle sentence-BERT utilisé précédemment, on a créé des embedding. Ensuite, les embeddings que nous venons de créer ont été fusionnés avec ceux du jeu de données incel de départ et nous avons appliqué les techniques de réduction dimensionnelle mentionnées précédemment, PCA et UMAP. Les représentations graphiques des résultats sont présentées ci-dessous [5]. Les points verts sont les phrases de contrôle et les points rouges sont les phrases incel.

4. *ARTeLab/ilpost · Datasets at Hugging Face*, oct. 2022, URL : `https://huggingface.co/datasets/ARTeLab/ilpost` (visité le 21/04/2024).

5. Pour la représentation des points obtenus avec UMAP, une librairie graphique différente a été utilisée pour rendre la visualisation plus claire.

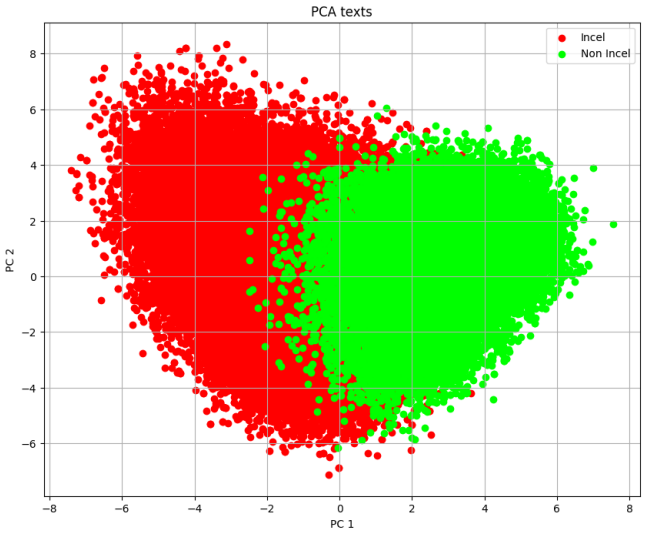

FIGURE 3 – Plot des données réduites par PCA

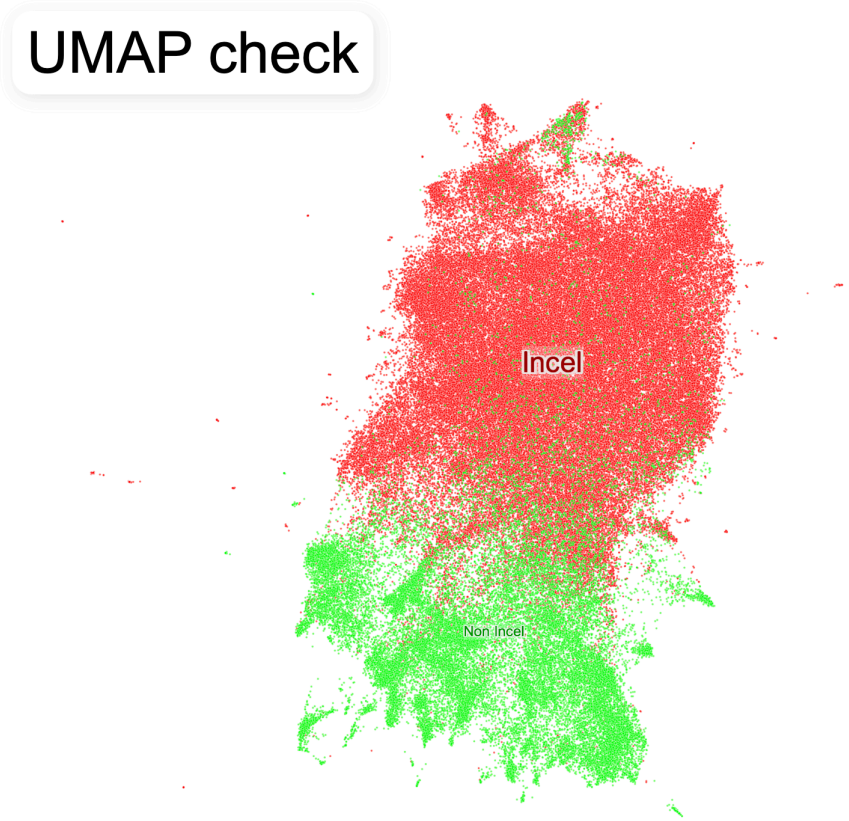

FIGURE 4 – Plot des données réduites par UMAP

Comme on peut le voir, les deux groupes peuvent être identifiés assez facilement à partir des deux graphiques. Cela montre globalement que le modèle de langue utilisé fonctionne et que, par conséquent, le chevauchement presque complet des points dans les figures 1 et 2 peut être considéré comme une preuve de l'homogénéité du sens des phrases dans le dataset de départ.

5 Les limites

L'approche présentée dans ces pages est incomplète pour plusieurs raisons. Tout d'abord, la méthode du groupe de contrôle décrite dans la section précédente n'est utile qu'à un niveau général et intuitif. Pour comparer les résultats de manière plus rigoureuse, il existe des méthodes mathématiques qui permettent de mesurer le niveau de *clustering* des données. En outre, la performance d'un modèle sentence-BERT peut elle-même être calculée en utilisant des jeux de données déjà existants, même en italien, afin d'obtenir des scores précis. L'approche décrite dans ces pages n'est donc utile que pour une analyse préliminaire des données dans les cas où un niveau de précision particulièrement élevé n'est pas requis.

En outre, un autre aspect n'est pas pris en compte : la présence de valeurs aberrantes. Comme on peut le voir dans la figure 2, certains points sont très éloignés des autres (il y en a aussi dans le cas de la PCA, bien qu'ils ne soient pas visibles dans le graphique). Cela suggère qu'il pourrait s'agir de valeurs aberrantes. Une approche plus complète devrait identifier et décider comment traiter ces valeurs. Une méthode possible, par exemple, consiste à calculer un embedding moyen et à déterminer la distance de chaque texte par rapport à la moyenne, afin d'éliminer ensuite les valeurs trop élevées – la définition de "trop élevé" pouvant être trouvée de différentes manières.

Références

ARTeLab/ilpost · Datasets at Hugging Face, oct. 2022, URL : https://huggingface.co/datasets/ARTeLab/ilpost (visité le 21/04/2024).

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).