**Nos données**

Nous utilisons le dataset Sentimental Liar construit à partir du dataset Liar.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Liar DataSet | | | | | | | | | | | | |
| ID | Label | | Statement | | SUBJECT | | SPEAKER | SPEAKER JOB | | STATE INFO | PARTY AFFILIATION | |
| Id du statement | Label de vérité avec 6 valeurs possibles :  -TRUE   * MOSTLY TRUE * BARELY TRUE * HALF TRUE * FALSE * PANTS ON FIRE | | Déclaration du politique, souvent une seule phrase, parfois coupé pour garder seulement l’information utile.  Exemple :  "When undocumented children are picked up at the border and told to appear later in court ... 90 percent do not then show up." | | Sujet(s) de la conversation | | Nom et Prénom de l’interlocuteur | Métier de l’interlocuteur | | Etat dans lequel l’auteur exerce son métier | Partie politique de l’interlocuteur | |
| Liar DataSet | | | | Sentimental Liar DataSet | | | | | | | |
| Label Count | | CONTEXT | | Sentiment\_Score | | Emotions Scores | | | Sentiment | | |
| Compte le nombre de messages de l’auteur dans les différents Labels | | Endroit où la déclaration a été dite. | | Score de sentiment général calculé avec l’API NLP Google.  Un score négatif indique une émotion négative, sinon une émotion positive. | | Scores d’émotion entre 0 et 1 pour 5 émotions calculés avec l’API NLP IBM.  Les 5 émotions sont :   * Colère * Peur * Joie * Tristesse * Dégout | | | \_\_NEG\_\_ ou \_\_POS\_\_ en fonction du score de sentiment | | |

Nous avons aussi essayé d’ajouter un peu plus d’information avec un autre dataset donnant d’avantages d’informations sur les politiques :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Twitter\_username | Account\_start\_time | Sex | Birthplace | Birthday | Age | Instagram\_username |
| Pseudo twitter du politicien | Date de création du compte twitter | Sexe du politicien | Pays de naissance du politicien | Date de naissance du politicien | Age du politicien | Pseudo instagram du politicien si existant (vide si inexistant) |

Nous relions les données grâce au nom des politiciens.

**Quelques informations sur nos données**

* Nombre de messages : 11515

**Nombre de mots**

* Nombre de mots différents dans les déclarations (statement) : 14089
* Nombre de mots moyens des déclarations (statement) : 17.8 (écart type de 7.7)
* Nombre de mots différents moyens des déclarations (statement) : 16.396 (écart type de 6.3)

Même chose avec les radicaux et une liste de stopwords :

* Nombre de mots moyens des déclarations (statement) : 9.337 (écart type de 3.87)
* Nombre de mots différents moyens des déclarations (statement) : 9 (écart type de 3.56

**Les Auteurs**

* Nombre d’auteurs différents dans le dataset Sentimental Liar: 3125
* Nombre d’auteurs mis en relation avec notre second dataset : 544
* Nombre de messages par auteur en moyenne : 3.6848 (de variance 15.7)
* Nombre d’auteurs avec au moins 5 messages : 378 (184 trouvé dans le second dataset) avec 7533 messages
* 939 Républicain, 854 Démocrates, 938 sans parties et quelques autres
* Nombre de job différents : 1267
* Nombre de personnes sans job renseigné : 1144
* Job le plus représenté : State Senator avec 50 personnes
* Nombre de job représenté par au moins 5 personnes : 34
* Moyenne du nombre de personne par job : 1.56 (écart-type de 3.59)

**Les labels**

* Nombres de Label false/half-true/mostly-true/true/barely-true/pants-fire: 2254/2361/2212/1844/1890/954
* Nombre de déclarations avec sentiment négatif/positif : 8560/1541

**Les scores**

* Moyenne des différents scores d’émotions (colère/peur/joie/dégout/tristesse) :

0.42789406, 0.15821795, 0.13586807, 0.17648124, 0.20195123

**Les sujets**

* Nombre de sujets différents : 142 dont 135 avec au moins 5 déclarations
* Nombre moyen de messages par sujet : 175 (écart type de 235)

**Avantages de nos données**

* Beaucoup d’information sur les auteurs.
* On a les métiers des utilisateurs, très utile pour la notation de compétence
* Les déclarations ne gardent que des informations utiles
* Messages annotés avec des différents niveaux de vérités => intéressant car nous voulons aussi des niveaux différents de notation d’information
* Des scores d’émotions sont disponibles, intéressant car les émotions sont marqueurs de fausses informations.
* Information sur le contexte, peut être intéressant pour voir les médias pouvant être source utile d’information.
* Présence des sujets des messages, intéressant la aussi pour la notation de compétence pour séparer les messages dans différents domaines.
* Le partie politique peut permettre une contextualisation des opinions des différents auteurs (deux personnes en accord sur un sujet est plus intéressant s’ils sont dans des parties politiques différents).
* Plusieurs messages pour un même auteur

**Désavantages de nos données**

* Pas de lien entre les messages (sauf l’auteur), ce qui rend complexe la mise en contradiction de messages.
* Pas de dates pour les messages.
* Déclarations courtes => peu de données dans les messages.
* Perte d’une partie de l’information avec la suppression éventuelle d’une partie des messages.
* Ensemble d’auteurs limités (seulement des politiques, donc pas de faux comptes par exemple, ce qui réduit l’intérêt d’une note utilisateur).
* Contexte trop détaillé ce qui empêche de faire une séparation par contexte (même si reste possible).

**Autres Spécificités des données**

* Sources multiples