PLDAC - Cotation d'information textuelle

Luc STERKERS

Dao THAUVIN

1. Table des matières

Introduction

Aujourd'hui les réseaux sociaux possèdent une place importante dans notre rapport aux autres et à l'information. Selon un sondage de 2019 réalisé par la Fondation Jean-Jaurès [7], les réseaux sociaux représentent la principale source d'information de 46% des Français.es de moins de 35 ans. La non-régulation et le mécanisme de diffusion de proche en proche de ces réseaux rend ces derniers particulièrement favorables pour le développement et la diffusion de fausses nouvelles. Il est donc aujourd'hui important de mettre en place des méthodes permettant d'appréhender la véracité de ces informations.

L'objectif de ce projet est d'ouvrir le champ de recherche sur la cotation d'information.
Le but de la cotation d'information n'est pas de vérifier si une information est vraie ou fausse dans l'absolu, à la différence des travaux dit de « fact checking ». Il s'agit plutôt de la mise en place d'algorithmes utilisant certaines composantes propres aux données (le parti politique d'un auteur, le sujet de l'information, …) pour attribuer un score de confiance à des informations.
La cotation d'information est un domaine encore peu développé, il n'existe donc pas de base de données spécialisées.

Dans un premier temps nous avons donc essayé de trouver une base de données intéressante pour cette tâche. Nous explorerons ensuite deux critères de cotation d'information basé sur l'auteur du texte que nous explorons. Dans un second temps, nous étudierons comment les informations d'un auteur peuvent être utiles pour connaître la véracité d'une information.
Dans un troisième temps, nous allons définir deux méthodes utilisant à la fois l'information du métier de l'auteur et le sujet de l'information. La première méthode basée sur la probabilité de la vérité d'une information sachant son sujet et le métier de son auteur. La deuxième, basée sur une approche sémantique de leurs intitulés.

Etat de L'art

La plupart des travaux réalisé aujourd'hui sur la véracité de l'information se concentre sur une approche de « fact-checking ». Une autre approche populaire de la question est l'approche linguistique qui se concentre sur l'apprentissage de motif dans la langue pour essayer de classifier des fake-news. C'est le cas dans le papier de Abhishek Koirala [2] écrit en 2020 où l'on étudie l'application de réseaux de neurones récurrents dans la classification de fake-news sur le covid.

Dans l'article *Providing Web Credibility Assessment Support* [3], les auteurs présentent une grande diversité de méta données que l'on peut utiliser pour renforcer les performances de nos algorithmes. Ils évoquent notamment l'utilisation des informations de l'auteur ou le type de site internet sur lequel l'article a été trouvé.

Cette approche multicritère a été mis en pratique dans l'article *Automatic Fact-Checking Using Context and Discourse Information* [1]. Dans cette article les auteurs classifient des informations orales retranscrites provenant du débat de 2016 entre Hilary Clinton et Donald Trump. Ils comparent alors la performance de différents algorithmes de détection de fake-news sur la retranscription du débat en enrichissant leurs entrées avec les éléments qui l'entoures, comme la réaction physique de l'auteur ou l'hilarité du public. Ils mettent ainsi en avant un gain net de performance de leurs algorithmes en utilisant des données enrichies plutôt qu'en utilisant uniquement l'information textuelle.

Dans l'article *Information Evaluation* [6], ces auteurs présentent une approche multicritère plus globale de la véracité de l'information. Ils mettent notamment en avant un ensemble de critère à étudier en s'appuyant sur le raisonnement humain instinctif vis-à-vis d'une nouvelle information. Qui l'a écrite ? L'information est-elle vraisemblable ? L'information fait-elle partie d'un consensus ? L'auteur est-il compétent ?

Démarche

Aujourd'hui, le problème se rapprochant le plus de la cotation d'information est celui de la détection de fake-news dont l'approche la plus populaire est celle dite de « fact-checking ». Une approche de recherche d'informations qui consiste à extraire une information d'un texte et la comparer avec des informations d'une source extérieur (wikipédia, journaux fiable, …).

Seulement cette approche possède certaines limites : L'extraction d'informations dans un texte est une tâche difficile à réaliser, le processus de recherche est lent et couteux donc difficilement applicable sur des énormes flux de données comme Tweeter ou Facebook. Cette approche dépend aussi de la confiance que l'on accorde aux parties d'internet que l'on utilise dans la validation de l'information, de plus l'information n'est pas toujours vérifiable.

Ce projet se place dans le problème plus large de la cotation d'information. Peu d'articles ont abordés ce domaine. Nous essayerons donc d'ouvrir ce domaine en explorant différentes approches.

Nous nous sommes appuyés sur le résonnement instinctif humain vie à vie d'une nouvelle information et l'ouvrage *Information Evaluation* [6]. Dans lequel ses auteurs mettent en avant une approche de l'évaluation d'une nouvelle information en 4 points :

Le premier point est la fiabilité de l'auteur. Il s'agit simplement de la confiance que l'on peut avoir dans une source : diffuse t'elle souvent des informations mensongères ? Fait elle partit d'un groupe qui diffuse souvent des informations erronées ?

Le second point est la compétence de l'auteur. Comme dit le proverbe : « Sutor, ne supra crepidam » (Cordonnier, pas au-delà de la sandale). Plus l'auteur est proche professionnellement du sujet abordé, plus une information est crédible. Un médecin sera toujours mieux placer pour parler d'une maladie qu'un commerçant.

Le troisième point est la plausibilité d'une information, il s'agit de pouvoir établir une probabilité apriori de la véracité d'une information. Ce point-ci est très difficile à analyser car la plausibilité est subjective.

Le dernier point est la crédibilité d'une information. C'est le consensus autour d'une information. Plus l'information est largement diffusée, plus elle a de chance d'être vraie.

Dans ce projet, nous nous concentrons uniquement sur les 2 premiers points.

Base de données

Nous aimerions une large base de données d'informations étiquetés par différents dégrées de vérité pour pouvoir vérifier que nos méthodes donnent des résultats cohérents. Nous aimerions en plus, avoir un maximum d'information sur les auteurs de ces textes affins de pouvoir définir des scores approximatifs des critères de compétence et fiabilité basé sur ces informations.

Nous nous penchons plus spécialement sur les bases de données de « fact-checking » disponibles en grande quantité et diversité sur internet. Ces bases de données possèdent comme caractéristique commune de posséder un ensemble de texte étiqueté selon leur véracité. Néanmoins la plupart des bases de données de fact-checking sont anonymisés.

Les données de la base « Liar, Liar Pants On Fire » [4] que nous avons utilisé dans ce projet proviennent du site POLITIFACT.COM. C'est un site de vérification d'informations provenant des médias et dirigé par des journalistes. Nous utilisons une version enrichie des données [5] ajoutant notamment des scores de sentiment en utilisant le Google NLP API pour ajouter un score de sentiment et le IBM NLP API pour ajouter des scores pour différentes émotions.

Voici un aperçu de ce que l'on peut trouver dans la base de données :

On définit $v = \{\text{true, mostly-true, barely-true, half-true, false, pants-fire}\}$
Une citation x, est composée de :

- $l \in v$: Un label étiqueté manuellement.
- $\mathbf{d}$: Le texte de la citation constitué de termes de $n = |d|$ termes $\mathbf{d} = (t_1, t_2, ..., t_n)$.
- $s \subset Sujets$: Un ensemble de sujets tel que $|s| > 0$.
- $c \subset Contextes$: Un ensemble de contextes tel que $|c| > 0$.
- $sc \in [-1, 1]$: Un score de sentiment général.
- $ls \in \{NEG, POS\}$: Un label de sentiment.
- $sent \in [0, 1]^5$: Des scores de sentiments ($sent = (sent_{anger}, sent_{fear}, sent_{joy}, sent_{disgust}, sent_{sadness})$).
- $a \in Author$: Un auteur.

Un auteur a, est composé de :

- n : Un nom
- $j \in Job$: Le job de l'auteur.
- $s \in State$: L'état d'origine de l'auteur.
- $p \in Party$: L'affiliation politique de l'auteur.

Le principal intérêt de ces données est son nombre important d'attributs associés à un texte. Notamment, la présence d'un contexte et d'un sujet pour chaque message qui sont des données très intéressants dans notre cas.

Les contextes sont peu nombreux (142 pour 11515 messages) et ont un grand nombre de données associées (en moyenne 175 messages par sujets avec un écart type de 235). On peut donc facilement réaliser une probabilité conditionnelle en utilisant cet attribut.
L'attribut métier d'un auteur est également un élément très intéressant. Il permet notamment d'analyser la compétence d'un auteur sur un domaine en particulier.
Il est aussi intéressant de connaitre le partie politique de l'auteur pour contextualiser l'opinion d'un auteur sur un sujet.
Il est aussi important de remarquer que les textes sont parfois coupés pour seulement garder l'affirmation de l'auteur, ce qui donne des citations courtes. Cela permet de réduire le bruit dans le texte même s'il est possible qu'une partie de l'information soit perdu.
Exemple d'une citation coupée:
« When undocumented children are picked up at the border and told to appear later in court ... 90 percent do not then show up. »
En moyenne une phrase contient 17.8 mots (avec un écart type de 7.7) dont 16.396 mots différents. En utilisant une liste de stopwords on passe à 9.337 (avec un écart type de 3.87) mots avec 9 mots différents en moyenne ce qui est relativement faible.
Tous les messages de notre base de données possèdent un score de vérité attribué manuellement dont voici les répartitions :

FIGURE 4 : REPARTITION DES DIFFERENTS SCORES SUR L'ENSEMBLE DES CITATIONS

Cette notation est pertinente par sa non-binarité. En effet, au lieu d'avoir une binarisation vraie et faux comme c'est le cas pour bon nombre de bases de données, on a ici 6 labels possibles que l'on peut mettre sur une échelle : True, Mostly-True, Half-true, Barely-true, False, Pants on fire.
On peut également binariser ces scores dans le cas où l'on voudrait établir un classifieur binaire avec deux classes : la classe positif (half-True, Mostly-True, True) et la classe négative (Pants on fire, False, Barely true).
On peut également séparer cette échelle en trois catégories : les labels positifs (True, Mostly-True), les labels neutres (Barely-true, Half-True) et les labels négatifs (False, Pants on fire).
Cette notation nous permet donc une grande liberté dans les méthodes pour attribuer un score de crédibilité à un message ou à un auteur.

Toutefois on peut noter certains points faibles de cette base données.
En premier lieu, les politiciens sont surreprésentés dans notre base de données.

Un autre problème est le manque de normalisation des jobs et des contextes. Cela rend difficile l'analyse de ces attributs, la plupart n'étant reliés qu'a un seul auteur ou texte.

Le nombre de messages par auteur est aussi relativement faible, avec une moyenne de 3.6848 (d'écart type 3.96). On a seulement 378 auteurs avec plus de 5 messages (avec 7533 messages associés), ce qui est très faible par rapport à l'ensemble des données et rend difficile l'étude des messages par rapport aux auteurs.

Fiabilité de l'auteur

Les auteurs de notre base de données possèdent 4 attributs : un nom, un nom de métier, un état de naissance et un parti politique d'affiliation. On veut établir une fiabilité à priori sur l'auteur, savoir si on peut lui faire confiance ou non uniquement avec ses informations. On établit donc des probabilités conditionnelles sur la valeur de vérité des différents messages selon les attributs des auteurs.

0. Définition

On s'intéresse à la description de la fiabilité de l'auteur.

La fiabilité a pour objectif d'utiliser la question « Qui propose l'information ? » pour donner une indication sur la véracité de l'information.

On peut donc naïvement définir la fiabilité d'un auteur comme étant la probabilité de vérité d'une information sachant un auteur.

$$\boldsymbol{A} \in \boldsymbol{Auteur}, \qquad \boldsymbol{V} \in \boldsymbol{vérité}$$
$$\boldsymbol{P(V|A)}$$

Une première approche est donc simplement d'effectuer les probabilités conditionnelles pour les différents auteurs. On peut même complexifier cette probabilité avec l'hypothèse qu'un auteur à une certaine stratégie dans sa manière de mentir et donc chercher à établir la probabilité :

$$\boldsymbol{A} \in \boldsymbol{Auteur}, \qquad \boldsymbol{Vi} \in \boldsymbol{vérité}$$
$$(\boldsymbol{Vi\ est\ la\ valeur\ de\ vérité\ du\ i-ème\ message\ de\ l'auteur\ A})$$

$$\boldsymbol{P(V_i|A, V_0, V_1, \dots, V_{i-1})}$$

Seulement, cette approche est impossible avec notre base de données actuelle. Nos données ne sont pas datées, ce qui nous empêche toute analyse temporelle. La majorité des auteurs n'ont qu'un seul texte associé et la moyenne du nombre de données par auteurs est de 3. Cette contrainte réduit considérablement la pertinence d'une analyse sur les auteurs.

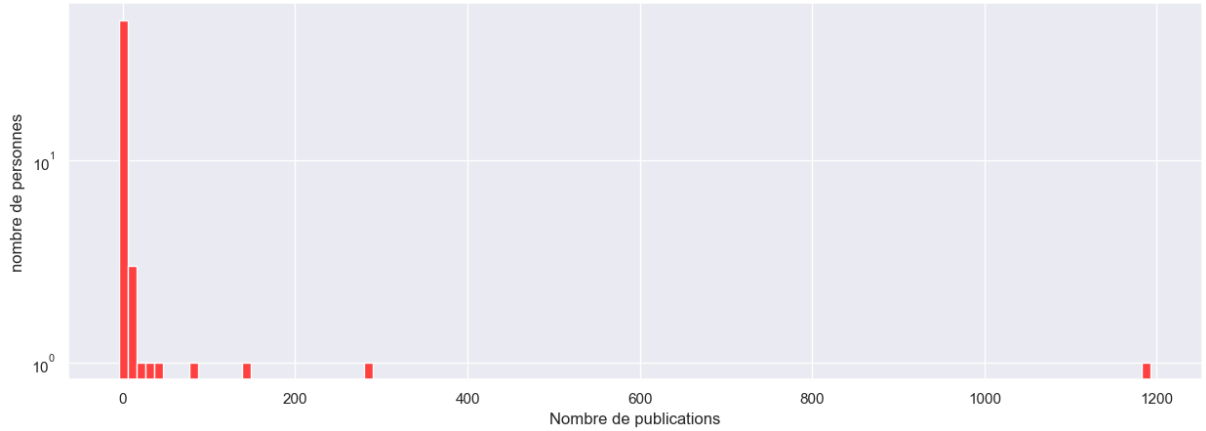


FIGURE 1 : NOMBRE D'AUTEUR PAR NOMBRE DE PUBLICATION

Pour pallier à ce problème, on regroupe les auteurs ensembles selon leurs caractéristiques communes. On met donc en place la probabilité suivante :

$$\boldsymbol{G\ un\ ensemble\ d'auteur}$$
$$\boldsymbol{P(V|G)}$$

Dans les parties suivantes, nous étudions donc les probabilités conditionnelles des différentes valeurs de vérité selon plusieurs regroupements possibles d'auteur.

1. Probabilité de vérité par partis politiques

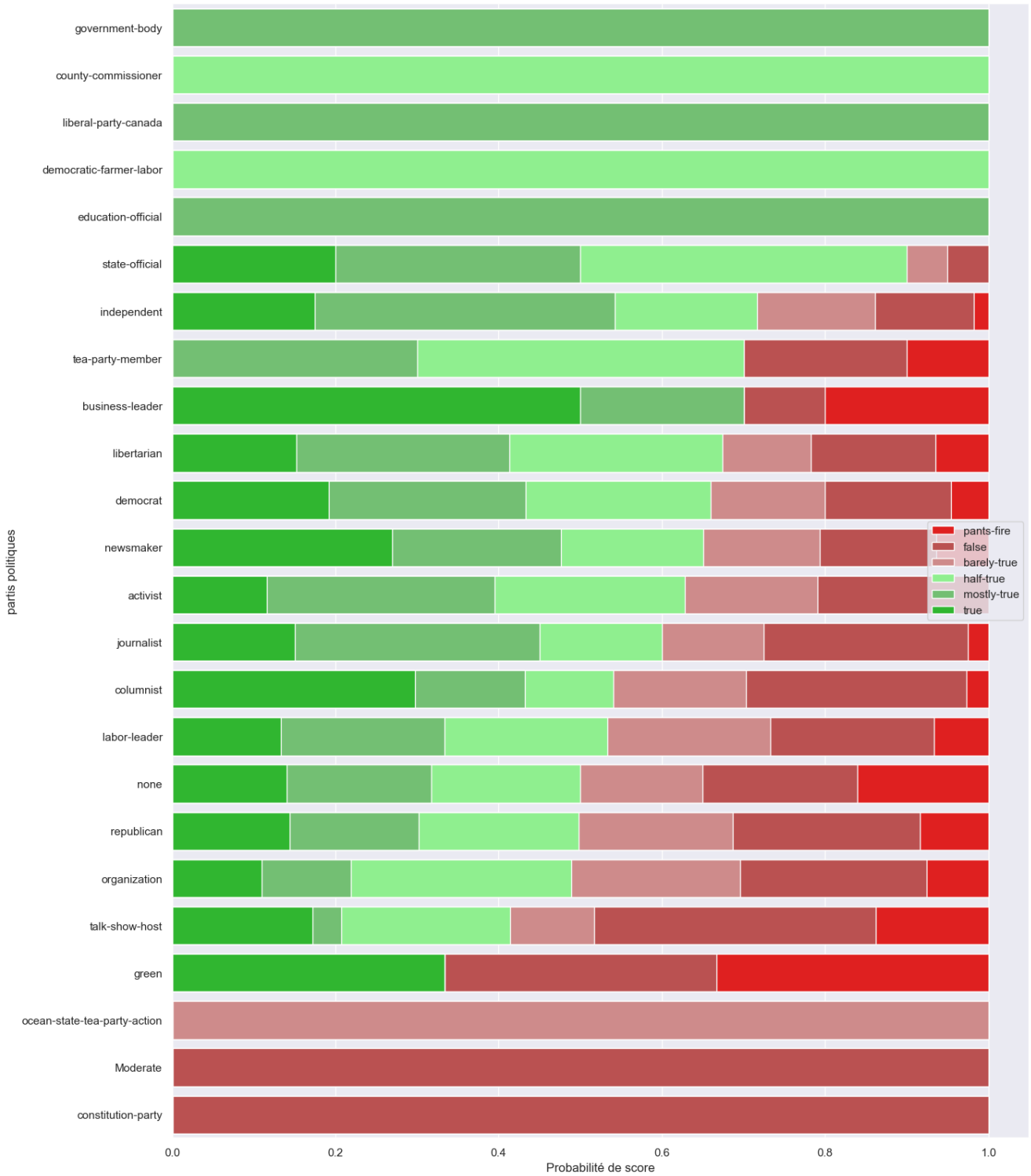


FIGURE 5 : REPARTITION DES DIFFERENTS SCORES PAR PARTI POLITIQUE.

Sur la figure ci-dessus, on peut observer les probabilités des différentes valeurs de vérité des informations selon le parti politique de l'auteur. On peut remarquer que ce groupement n'est pas parfait car certains partis politiques ont peu d'auteurs associés (2 ou 3) ce qui se traduit par une probabilité certaine pour une valeur de vérité.

2. Probabilité de vérité par état de naissance

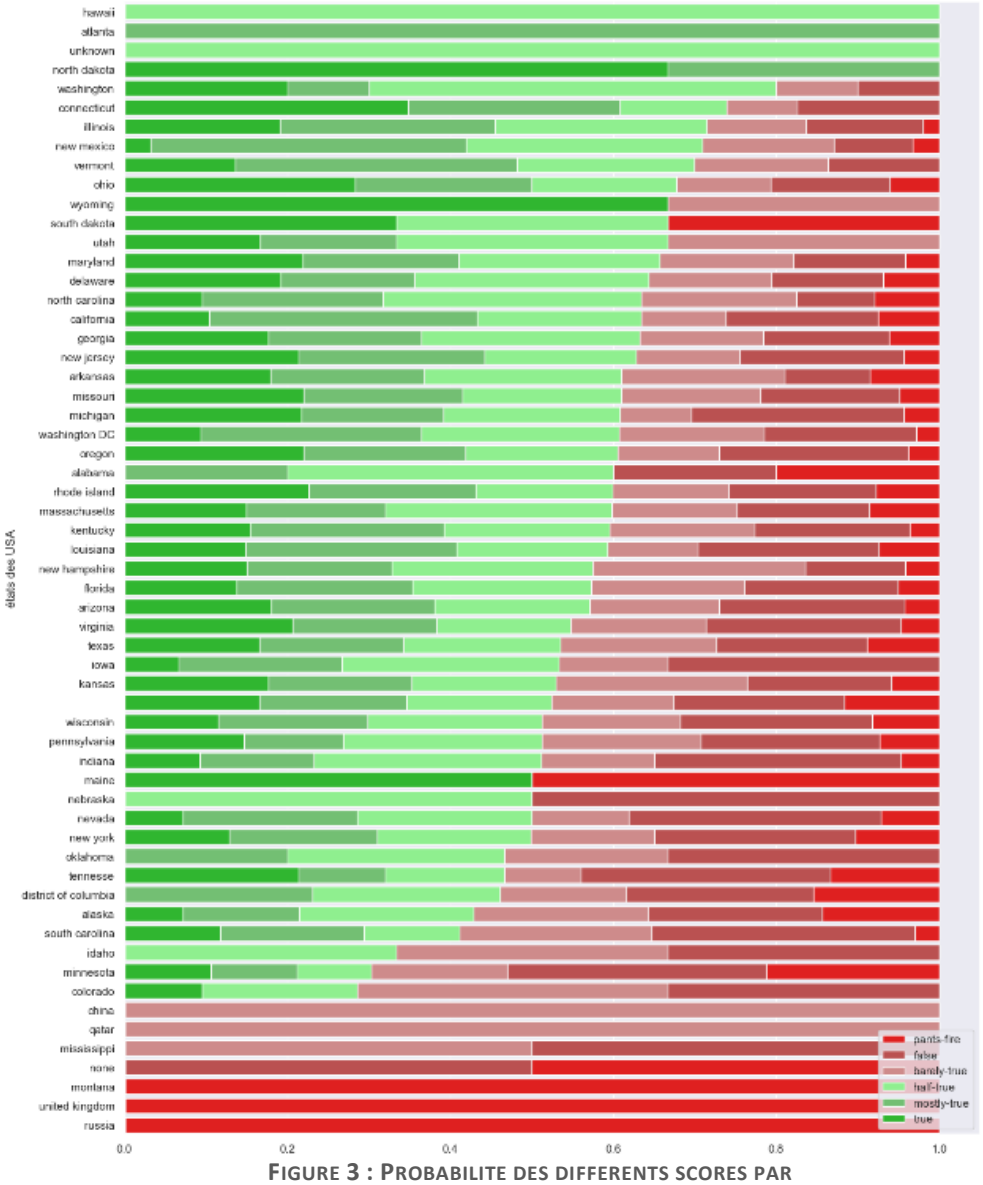


FIGURE 3 : PROBABILITE DES DIFFERENTS SCORES PAR ETAT DE NAISSANCE (EN GRAND FORMAT DANS L'ANNEXE).

Comme les partis politiques, les états de naissances nous donnent également des probabilités conditionnelles qui nous sembles intéressantes. On remarque également peu de messages pour certains états. On peut également se poser la question de la pertinence de l'utilisation de l'état de naissance comme critère de fiabilité. En effet, cet usage pose une question éthique : peut-on vraiment juger la fiabilité d'un auteur par son lieu de naissance ?

3. Développement et limites de cette approche.

Les probabilités sur un regroupement d'auteur est donc un moyen simple de réaliser une approximation de la fiabilité à priori à un auteur. Avec plus de données, on peut développer cette approche en effectuant différents groupements pour établir une probabilité conditionnelle plus complexe : $\mathbf{P}(\mathbf{V}|\ \mathbf{G_1}, \mathbf{G_2}, \ldots)$. Il faut cependant penser chaque probabilité et leur utilisation dans notre méthode d'un point de vue éthique.

Compétence de l'auteur

Une information nous paraît d'autant plus vraisemblable que son auteur est proche professionnellement du sujet abordé. Un médecin nous paraît mieux placer qu'un mécanicien pour parler de médecine. Intuitivement, on a la notion que certaines professions sont plus crédibles que d'autres sur un sujet donné. Par réciprocité, on peut estimer que les métiers qualifiés pour parler d'un sujet sont ceux qui ont la plus forte probabilité de donner des informations vraies sur un sujet donné. On peut donc décrire la compétence d'une profession sur un sujet comme étant :

$$\boldsymbol{V \in Labels - positif, J \in Job, S \in Sujet}$$

$$\boldsymbol{P(V|J,S)}$$

On pose également l'hypothèse suivante : Si un métier n'évoque jamais ou presque un sujet, alors ce métier n'est pas compétent pour parler du sujet. Cette hypothèse nous permet d'exploiter le manque de messages pour un couple sujet-job.

1. Clustering

Un premier problème auquel nous faisons face est l'annotation non normalisé et très spécifique des intitulés des jobs. On compte 1185 intitulés de métier différents pour 2909 auteurs identifiés. Sans compter le fait qu'un grand nombre d'auteurs n'ont pas de métier renseigné. On a donc en moyenne 2.45 auteurs par métier ce qui ne nous permet pas d'établir des probabilités conditionnelles fiables en utilisant cet attribut.

FIGURE 4 NOMBRE DE METIER PAR NOMBRE DE PERSONNES AFFILIES AUX METIERS

'abc news', 'abc news chief health medical editor', 'abc news chief white house correspondent', 'acting director independent texans', 'activist group', 'actor', 'actor director', 'actor director activist', 'advertising executive', 'adviser'

ECHANTILLON DE METIERS PRESENTS DANS NOTRE BASE DE DONNEES

On se fait la réflexion que beaucoup de métiers nous semblent sémantiquement proches. Comme 'abc news' et 'abc news chief health medical editor'. Nous avons donc choisi d'effectuer un regroupement sémantiquement des noms de métiers.

a) Vectorisation

On commence par prétraiter les intitulés des métiers pour enlever les ponctuations inutiles, les chiffres, les majuscules et corriger un certain nombre de fautes d'orthographe.
On utilise ensuite le modèle Word-To-Vec *glove-wiki-gigaword-300* de la bibliothèque python gensim, pour représenter chaque terme des intitulés de métier en vecteur de taille 300.
On calcule ensuite la moyenne des vecteurs de mots d'un métier pour établir un unique vecteur de dimension 300 le représentant.

Intitulé du métier	'advocacy group senior citizens'	'ceo san antonio water system'	'former ruth chris steak house ceo'
Termes les plus proches de la représentation vectorielle de l'intitulé avec leur similarité cosinus.	[('group', 0.763512134552002), ('groups', 0.7134451270103455), ('advocacy', 0.69323194026947), ('citizens', 0.6904720664024353), ('senior', 0.6352942585945129), ('members', 0.60796028375625), ('organization',0.6017993092536), ('organizations',0.5950316786752), ('citizen', 0.5779377818107605), ('leaders', 0.5716676712036133)]	[('san', 0.7379541993141174), ('francisco', 0.645039856484), ('antonio', 0.637013256542), ('diego', 0.58427006006084), ('water', 0.5547513365745544), ('california', 0.54230087995527), ('system', 0.5387765169143677), (',', 0.5308255553245544), ('.', 0.5155636668205261), ('systems', 0.51551419496535)]	[('former', 0.6239941716194153), ('chris', 0.610386312007904), ('house', 0.5977050065994263), ('executive', 0.54552167654037), ('ceo', 0.5294105410575867), ('smith', 0.5237122774124146), ('friend', 0.5199733376502991), ('johnson', 0.507293283993616), ('ruth', 0.5070990920066833), ('miller', 0.4945620000362396)]

CONFRONTATION DES INTITULES DE METIER ET LEUR REPRESENTATION VECTORIELLE.

On observe ci-dessus, diffèrent noms de métier et les termes les plus proches de leur représentation vectorielle. On observe que la majorité des termes présents dans le nom du métier sont également dans les 10 plus proches termes de son vecteur associé. Cette représentation nous semble donc pertinente car elle semble garder l'information de l'intitulé. On pourrait toutefois enlever les localisations géographiques présentes dans certain métier. Un élément qui ne nous semble pas pertinent pour traiter sémantiquement un métier dans le cadre de la mise en place d'un critère de compétence.

b) Clusterisation

On teste ensuite, avec nos métiers vectorisés, plusieurs méthodes de clusterisation : MeanShift, SpectralClustering et Kmeans. Pour effectuer un regroupement qui nous semble pertinent

On aimerait avoir une distribution de la taille des clusters la plus uniforme possible pour avoir un maximum de message par cluster dans le but d'effectuer des probabilités plus fiables sur les clusters obtenus, sans déséquilibre des données.

L'algorithme MeanShift produit, dans notre cas, une centaine de cluster avec un unique élément à l'intérieur et un unique cluster possédant 98% des intitulés de jobs. On ne privilégie donc pas cette méthode par rapport à la distribution que l'on recherche.

Les méthodes Kmeans et SpectralClustering possèdent un attribut indiquant le nombre de clusters à effectuer. Cela nous permet de faire varier la sensibilité de notre modèle. On remarque que plus on augmente le nombre de clusters à effectuer et plus on augmente la capacité à de petits groupes de se former.

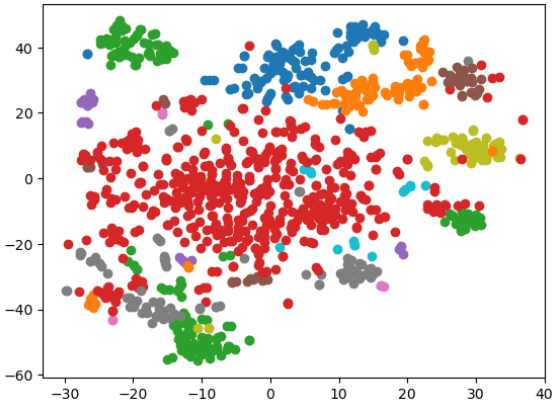


FIGURE 5 TSNE DU CLUSTERING DES METIERS AVEC L'ALGORITHME SPECTRAL CLUSTERING

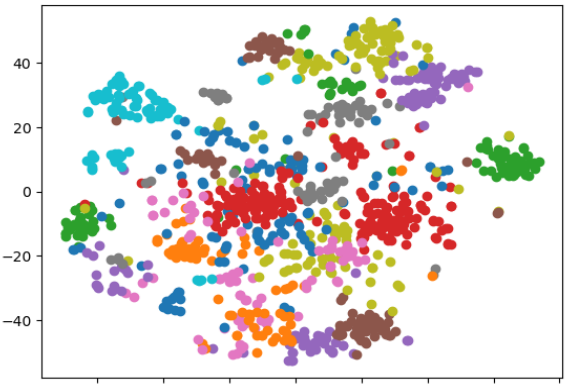


FIGURE 6 TSNE DU CLUSTERING DES METIERS AVEC L'ALGORITHME KMEANS

On observe ci-dessus les clusters formés par l'algorithme Spectral Clustering et K-means. On remarque que l'algorithme Spectral Clustering possède un cluster dominant, ce que l'on veut éviter. On préfère donc K-means qui nous donnes des clusters plus petits et de tailles plus uniformes.

Après avoir parcouru les différents métiers, on construit un apriori manuel sur les différents groupes à former :

['politician', 'journalist', 'host', 'administration', 'teacher', 'scientist', 'businessman', 'communication', 'activist lobbyist group trade financial', 'ceo', 'policeman', 'owner', 'doctor health', 'army', 'religious', 'justice attorney', 'artist']

Pour augmenter les performances de la clusterisation de l'algorithme K-means, on introduit donc cet apriori sur les différents groupes à former en initialisant l'algorithme avec la vectorisation de ces mots.

Une fois les groupes formés, il est difficile de savoir à quoi chaque groupe correspond, quel est le thème en commun des différents éléments du groupe. Pour étiqueter un groupe, on lui attribue le terme le plus proche du vecteur moyen de ces éléments.

Ainsi les groupes aprioris ci-dessus deviennent après exécution de l'algorithme kmeans :

['republican', 'editor', 'television', 'secretary', 'teacher', 'university', 'businessman', 'computer', 'president', 'director', 'officer', 'owner', 'physician', 'retired', 'broadcaster', 'county', 'musician']

On peut remarquer que certains clusters n'ont pas changé de sens ('teacher'-> 'teacher'), certains ont dérivé (politician -> républican).

Ces dérivations sémantiques semblent liées à nos données. Le métier 'army' dérive en 'retired'. Une raison pour expliquer cette dérive est que les militaires en fonction n'ont souvent pas le droit de parler de sujets politique. La majorité des militaires présents dans notre base de données serait donc des militaires retraités. Cela pourrait expliquer la dérive sémantique du centre du cluster.
Il serait intéressant de tester cette approche sur d'autre base de données pour voir si les résultats changent.

c) Résultats

On a plus de 130 sujets différents, pour pouvoir effectuer un affichage lisible de la matrice conditionnelle entre les métiers et sujets. On effectue la même opération précédente de clustering sur les différents sujets. On choisit arbitrairement de former 60 clusters de sujet.

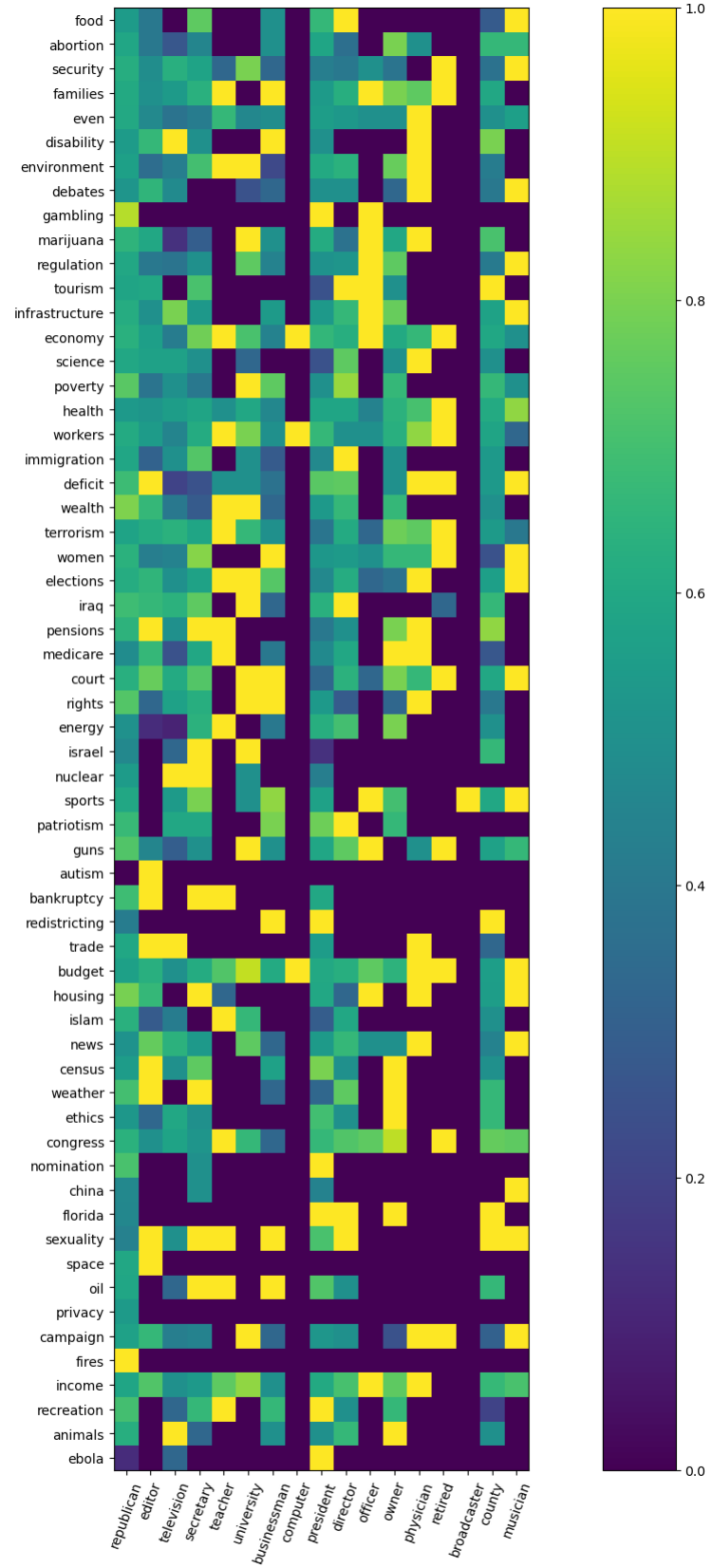


FIGURE 7 : PROBABILITE D'AVOIR UNE ETIQUETTE 'TRUE' OU 'MOSTLY-TRUE' SELON LE CLUSTER DE JOB ET LE CLUSTER DU SUJET

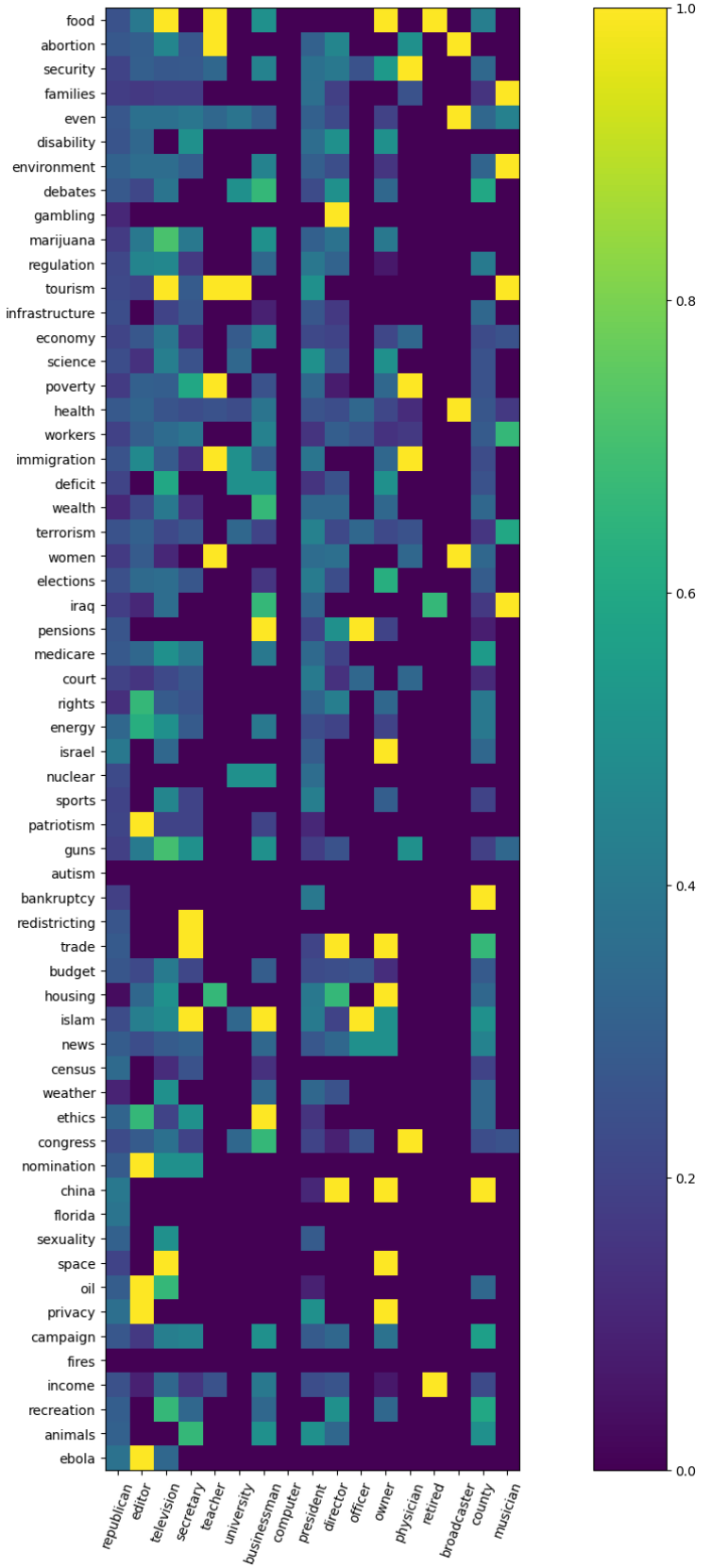


FIGURE 7 : PROBABILITE D'AVOIR UNE ETIQUETTE 'PANTS-FIRE' OU 'FALSE' SELON LE CLUSTER DE JOB ET LE CLUSTER DU SUJET

Sur ces matrices, on observe des fortes probabilités pour des couples job-sujet qui nous semble étonnantes. Comme la forte probabilité de vérité entre le métier 'directeur' et le sujet 'cuisine' ou celle entre 'university' et 'bankruptcy'.

Cette observation est à mettre en perspective avec les caractéristiques de notre base de données. Comme il s'agit d'une base de données initialement prévu pour effectuer du fact-checking. Les informations ont l'air d'avoir été sélectionnés pour former un rapport 50/50 de labels de vérité positives et négatives.

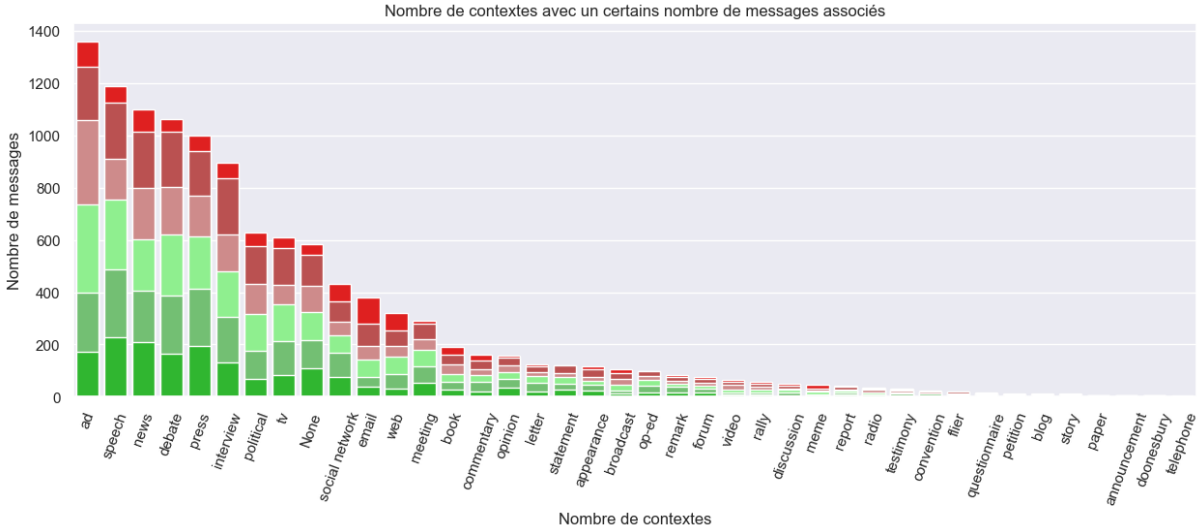


FIGURE 8 DISTRIBUTION DES VALEURS DE VERITE DES MESSAGES PAR CONTEXTES

On peut observer sur la figure ci-dessus qu'un nombre important d'informations de notre base de données proviennent de publicités. On peut également noter que la majorité des informations viennent de sources journalistiques (tv, interview, press, news, …) ou d'évènements politiques (speech, debate, …). Donc on peut mettre en doute la représentativité des auteurs de notre base de données dans leur métier. Cela pourrait limiter l'utilisation des probabilités formés avec cette base de données sur d'autres données.

2. Rapprochement Sémantique

Nous avons vu avec la méthode précédente, que les biais de nos données pourraient être un obstacle à la création d'un critère fiable.

Nous explorons donc une nouvelle approche plus indépendante de nos données pour tenter de définir un score de compétence plus fiable qu'avec la méthode précédente.

On fait l'hypothèse qu'il existe une corrélation entre la compétence d'un métier sur un sujet et la similarité sémantique entre l'intitulé d'un métier et du sujet.
Nous utilisons cette hypothèse pour réaliser cette nouvelle méthode. Nous réaliserons donc une distance entre la vectorisation sémantique des sujets et des métiers. Nous vérifierons qualitativement et quantitativement ce rapprochement.

a) Vectorisation

On utilise la même méthode que dans la partie précédente, pour chaque intitulé de métier ou sujet, on transforme chacun de ces termes en vecteur de dimension 300 en utilisant le modèle Word-To-Vec glove-wiki-gigaword-300 de la bibliothèque python gensim, puis une moyenne est réalisée pour obtenir un unique vecteur de taille 300 pour le métier/sujet.

b) Similarité

Nous avons testé les distances euclidienne et cosinus pour établir une distance sémantique.
Pour analyser les résultats, nous affichons pour différents sujet et métiers, les éléments (qui correspondent aux métiers pour un sujet et les sujets pour un métier) les plus proches. Ainsi un élément proche de notre job/sujet sera écrit plus gros qu'un autre éloigné (ici on affiche les 200 éléments les plus proches).
Il s'agit en grande partie d'une analyse subjective et basé sur une petite partie des données, elle ne permet donc pas de tirer de conclusions.
Cette étude permet tout de même de voir si la méthode semble cohérente dans ses résultats.
Pour le sujet « abc news week », les métiers associés les plus proches sont :

FIGURE 9 : WORDCLOUD PAR DISTANCE EUCLIDIENNE

FIGURE 8 : WORDCLOUD PAR DISTANCE COSINUS

On observe que ce sujet en rapport avec les news a été associé aux métiers de journalistes en priorité.
Ces résultats nous semblent cohérent.
Passons maintenant à la même analyse pour les métiers.
Par exemple le métier « éducation group » :

FIGURE 11 WORD CLOUD AVEC LA DISTANCE EUCLIDIENNE

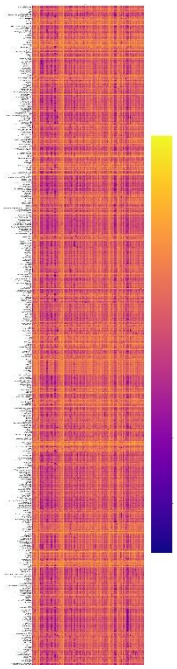
FIGURE 10 WORD CLOUD AVEC LA DISTANCE COSINUS

Les résultats nous semblent cohérents ici aussi.
Durant nos expériences, nous avons observé l'apparition de « Fact » en premier plan dans la plupart des WordClouds avec la distance euclidienne.
En effet cela pourrait s'expliquer par sa norme. Elle est la plus faible de tous nos sujets. Ainsi pour un métier avec une norme faible, sa distance avec ce sujet sera plus faible que les autres sujets avec une norme plus importante. Il semble donc plus intéressant d'observer la distance cosinus ici pour éviter ce possible biais.
Cette méthode semble pertinente d'un point de vue qualitatif et subjectif pour quantifier la compétence d'un métier sur un sujet.

c) Comparaison des résultats avec les données.

Ici nous analyserons en détail les résultats obtenues par cette deuxième méthode, en les comparants avec notre base de données. Nous étudions donc ainsi, la corrélation entre la véracité des messages d'un sujet et métier avec la distance sémantique de ce sujet et ce métier.
Et également, la corrélation entre le nombre de message pour un sujet et un métier et la distance sémantique entre ce sujet et ce métier.
Bien que, comme nous l'avons dit précédemment, les résultats pourraient varier avec d'autres données compte tenu des possibles biais de nos données et l'indépendance forte de l'approche avec les données.

Voici la matrice des distances obtenue sur les données avec cette méthode sur la base de données :



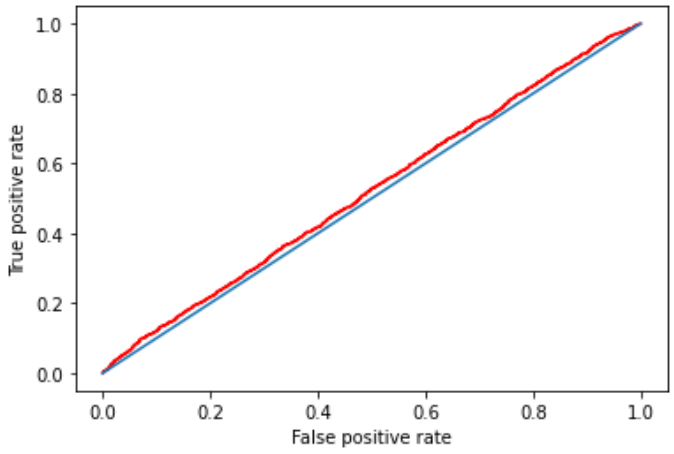
MATRICE DES DISTANCES

On observe que certains métiers et sujets ont une distance plus élevée que la moyenne pour tous les éléments, par exemple les sujets « ébola » ou « bipartisanship » et les métiers « restaurateur » ou « géologiste ».
A l'inverse, on n'observe pas de métiers ou sujets ayant une distance plus faible que la moyenne pour tous les éléments.
Une première comparaison que l'on peut faire avec cette matrice est la comparaison avec les labels de vérité de notre base de données pour vérifier que cette méthode donne bien une indication sur la vérité des messages.
Mais cela pose une difficulté majeure, il s'agit de comparer une distance et un label de vérité qui ne sont pas dans la même unité de mesure.
Pour pouvoir réaliser cette comparaison, nous réaliserons tout d'abord une binarisation des valeurs de vérités en 2 classes :

- Négative correspondant aux labels : barely-true, false, pants on fire
- Positive correspondant aux labels: true, mostly-true, half-true.

Puis pour obtenir une unique classe pour un couple metier-sujet, on utilise la classe majoritaire (Positive ou Négative) parmi l'ensemble de message avec ce job et sujet. Les couples sujet-job ayant autant de données de classe positive que négative ou n'ayant aucunes données associées seront ignorées car ils sont difficilement interprétables.

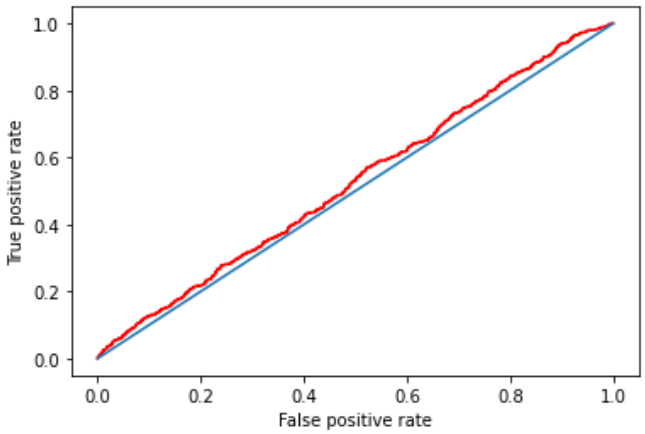
Le problème est qu'une distance n'est toujours pas comparable à un attribut binaire, c'est pour cela que nous utiliserons une courbe ROC pour les comparer.
Nous faisons donc l'hypothèse que pour un couple sujet-job, plus la distance sémantique est faible, plus il est probable que la classe majoritaire soit la classe positive.
Ainsi nous ajoutons progressivement les couples sujet-job dans l'ordre croissant des distances dans la classe positive pour créer notre courbe.
Voici la courbe ROC obtenue en rouge et en bleu la courbe attendue pour un ajout aléatoire des données :



COURBE ROC

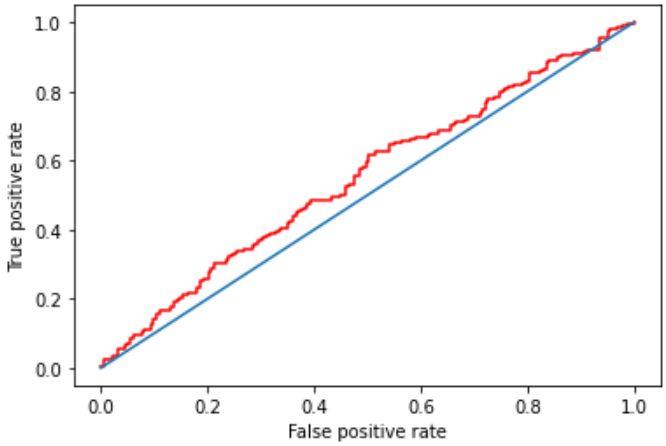
On observe ici que le critère semble n'apporter aucune information sur la vérité du message, en effet on observe une courbe quasi identique à l'aléatoire ici.
Mais ce résultat pourrait être biaisé car on observe un déséquilibre des données. Beaucoup de métiers ne sont présents qu'une seule fois dans la base (pour un seul message), ce qui n'est pas suffisant pour avoir un résultat cohérent.
Une solution pour obtenir de meilleurs résultats est d'augmenter le nombre de données pour ces métiers ou réaliser un clustering des métiers comme dans la partie précédente. Le problème est que cela demande de trouver une façon de mettre en place un rapprochement sémantique entre un sujet et un cluster de métiers.

La façon la plus simple et qui est adopté ici est de réduire le problème en ignorant une partie des données n'ayant pas assez de messages représentants. Nous le faisons ici en ajoutant un seuil de 20 messages minimum pour qu'un job ou un sujet soit pris en compte (il reste 123 sujets sur 140 et 49 métiers sur 1061, on a donc 6027 couples possibles) :



COURBE ROC AVEC SEUIL DE 20 MESSAGES PAR AUTEUR OU SUJET

Le résultat reste très proche de l'observation précédente et cela se confirme en augmentant le seuil à 100 messages par exemple (il reste 12 métiers et 58 sujets, on a donc 696 couples possibles dont 616 couples positifs, ce qui est faible pour avoir une mesure déterminante) :

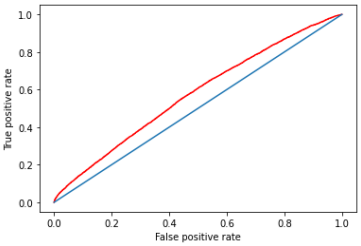


COURBE ROC AVEC SEUIL DE 100 MESSAGES PAR AUTEUR ET SUJET

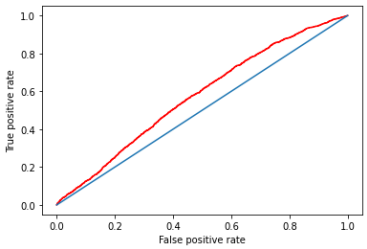
Les résultats sont légèrement meilleurs mais cela ne permet pas de conclure à un résultat meilleur que l'aléatoire. Ce résultat pourrait s'expliquer par le manque d'apriori sur les auteurs et les sujets.
Comme on peut le voir sur l'histogramme dans l'annexe page 26, il existe un apriori non négligeable sur les sujets dans notre base qui n'est pas pris en compte par notre critère.

Une seconde approche intéressante pour tester la pertinence de la méthode sans avoir de labels de vérités (notamment utile si on veut tester l'algorithme sur d'autres bases de données) est de faire l'hypothèse que si un métier est compétent pour un sujet alors les personnes de ce métier parleront davantage de ce sujet (et donnerons plus souvent de bonnes informations sur celui-ci).

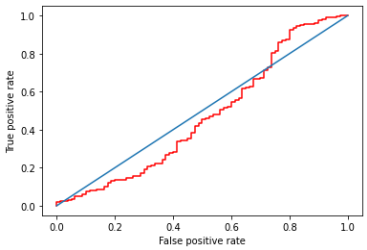
Avec cette hypothèse, on peut faire la comparaison de notre matrice de distance avec la matrice des présences simultanés d'un sujet et un job (valant 1 s'il existe un message sur le sujet d'un auteur de ce job, 0 sinon), ici nous utilisons un seuil d'un message pour que la classe de notre couple soit 1 mais on pourrait augmenter ce seuil.
On devrait s'attendre à trouver des messages sur les jobs et sujets ayant des distances sémantiques les plus faibles.
On peut alors réaliser la même expérience que précédemment avec des courbes ROC :



1 SANS SEUIL MINIMUM



2 AVEC UN SEUIL DE 20 MESSAGES



3 AVEC UN SEUIL DE 100 MESSAGES

Avec cette hypothèse, on observe que notre critère est légèrement meilleur que l'aléatoire sur les 2 premières courbes, mais les résultats sont moins concluants pour la dernière courbe, mais comme indiqué précédemment cela pourrait être du au manque de données.

d) Conclusion

Les résultats semblent intéressants d'après l'analyse qualitative effectuée mais cela n'est pas confirmer par notre analyse quantitative. Les données semblant être collecté pour obtenir une proportion équitable des différentes valeurs de vérités, Il est tout à fait possible que notre méthode se valide mieux sur des données réelles.
Bien sûr, cette méthode a beaucoup de défauts. Elle dépend principalement de l'intituler du sujet et du job. Si ces derniers sont mal renseignés les résultats pourraient en être impactés.
De plus, comme vu dans la partie Erreur ! Source du renvoi introuvable. nos valeurs sont très éloignées des valeurs de vérité. Comme dit dans la partie précédente, cela pourrait être due au fait que certains sujets et jobs sont plus enclins à créer de la fausse information. Cet apriori n'étant pas pris en compte dans notre méthode.
On peut aussi noter que l'utilisation de Word-To-Vec limite aussi la taille des sujets et jobs pris en compte, cette vectorisation étant connu pour ne pas être efficace sur des textes de tailles dépassant la dizaine de mot.
Cette méthode a aussi ces avantages. Même si un sujet ou un métier n'a jamais été rencontré auparavant, on peut leur associé un score de compétence, ce qui n'était pas possible avec l'approche précédente. Ainsi il n'est même pas nécessaire de réaliser une normalisation des noms tant que notre vectorisation est robuste (gère les fautes de frappes, les nouveaux mots, …).
De plus, cette indépendance aux nombres de données pour les différents sujets et jobs permet d'éviter l'impact des possibles biais de notre base de données sur nos résultats.
Pour finir, notre résultat est une distance, ce qui rend cette méthode difficilement comparable ou combinable avec d'autres critères.

3. Extensions et Approfondissements possibles

a) Information apriori

Comme nous l'avons vu précédemment, le second critère semble manquer d'un élément important pour réellement représenter les vérités des messages, l'information apriori sur les sujets et les métiers. Il serait donc intéressant d'ajouter une étape pour prendre en compte un apriori sur ceux-ci.

b) Prise en compte du vocabulaire

Il est aussi intéressant de remarquer que les méthodes présentées n'utilisent pas le contenu des messages et s'appuie complètement sur les intitulés des sujets et métiers.
Il serait donc intéressant d'utiliser le vocabulaire propre aux différents sujets et métiers pour enrichir leurs définitions et rendre les critères plus robustes.
Par exemple, en utilisant un classifieur linéaire (en utilisant un classifieur par sujet en one vs all et de même pour les métiers) pour trouver des poids pour les différents mots pour chaque sujet et métier. On pourrait alors utiliser

ces vecteurs de poids pour réaliser le clustering des sujets et métiers pour le premier critère. Pour le second, on peut changer notre hypothèse de départ et réaliser une distance entre ces vecteurs de poids directement plutôt que sur les noms d'intitulés vectorisés (même si on perd l'avantage de la vectorisation sémantique : un nouveau mot ne sera pas pris en compte dans la vectorisation).
Un problème qui pourrait survenir avec cette méthode est qu'elle nécessite un nombre important de messages pour chaque sujet et métier. De plus, cette approche rend plus difficile la prise en compte de nouveaux métiers et sujets. En contre parti, cette approche permet d'attribuer aux nouveaux messages un sujet ou un métier. Ce qui permet alors d'utiliser le critère pour des données n'ayant pas de sujets associés aux textes ou de métiers associés aux auteurs.

Conclusion et travaux futurs

Dans ce projet nous avons étudié et exploré des pistes dans le domaine de la cotation d'information.

Nous avons étudié comment mettre en place une mesure de confiance d'un auteur en créant des probabilités conditionnelles sur les groupes dans lequel il s'inscrit. Nous avant mis en avant certain problème éthiques qui peuvent apparaitre avec cette approche.

Nous nous sommes intéressés à la façon de mettre en place un score de compétence pour un auteur. Nous avons exploré deux approches : La première étant une approche probabiliste. Qui semble limitée par notre base de données. La seconde est une approche sémantique, qui semble limité par le manque d'apriori sur les couples job/sujet.

Dans des futurs travaux, il serait intéressant de valider la pertinence de nos méthodes sur d'autres données. On pourra alors vérifier si les points et les hypothèses mis en avant par notre étude sont véritablement pertinents dans le domaine de la cotation d'information.

Nous avons aussi observé la difficulté de tester la pertinence des approches de cotation d'information. Il serait donc intéressant de développer des méthodes de test de ces approches.

On pourrait également approfondir les recherches sur d'autres approches des notions de fiabilité et compétence ou approfondir les méthodes que nous avons évoqué.

Il serait également pertinent de développer les autres points que nous avons évoqués sans les traiter : la cotation de la plausibilité et crédibilité d'une information.

Un dernier point qui peut être étudier est la mise en commun de ces différentes approches pour créer un premier modèle de cotation d'information.

Bibliographie

[1] P. Atanasova *et al.*, « Automatic Fact-Checking Using Context and Discourse Information », *J. Data and Information Quality*, vol. 11, nº 3, p. 12:1-12:27, mai 2019, doi: 10.1145/3297722.

[2] A. Koirala, *COVID-19 Fake News Classification with Deep Learning*. 2020. doi: 10.13140/RG.2.2.26509.56805.

[3] S. Aggarwal, H. Van Oostendorp, Y. R. Reddy, et B. Indurkhya, « Providing Web Credibility Assessment Support », in *Proceedings of the 2014 European Conference on Cognitive Ergonomics*, New York, NY, USA, sept. 2014, p. 1-8. doi: 10.1145/2637248.2637260.

[4] W. Y. Wang, « "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection », *arXiv:1705.00648 [cs]*, mai 2017, Consulté le: févr. 23, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1705.00648

[5] B. Upadhayay et V. Behzadan, « Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification », *arXiv:2009.01047 [cs, stat]*, oct. 2020, Consulté le: févr. 22, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2009.01047

[6] Philippe Capet et T. Delavallade, Éd., *Information Evaluation*, 1re éd. John Wiley & Sons, Ltd, 2013. Consulté le: janv. 23, 2021. [En ligne]. Disponible sur: http://onlinelibrary.wiley.com/doi/10.1002/9781118899151

[7] Noémie Bonnin, « INFO FRANCEINFO. Les réseaux sociaux première source d'info en ligne chez les personnes sensibles aux théories du complot », *Franceinfo*, févr. 18, 2019. https://www.francetvinfo.fr/internet/reseaux-sociaux/info-franceinfo-les-reseaux-sociaux-premiere-source-d-info-en-ligne-chez-les-personnes-sensibles-aux-theories-du-complot_3191963.html (consulté le mai 27, 2021).

Annexe

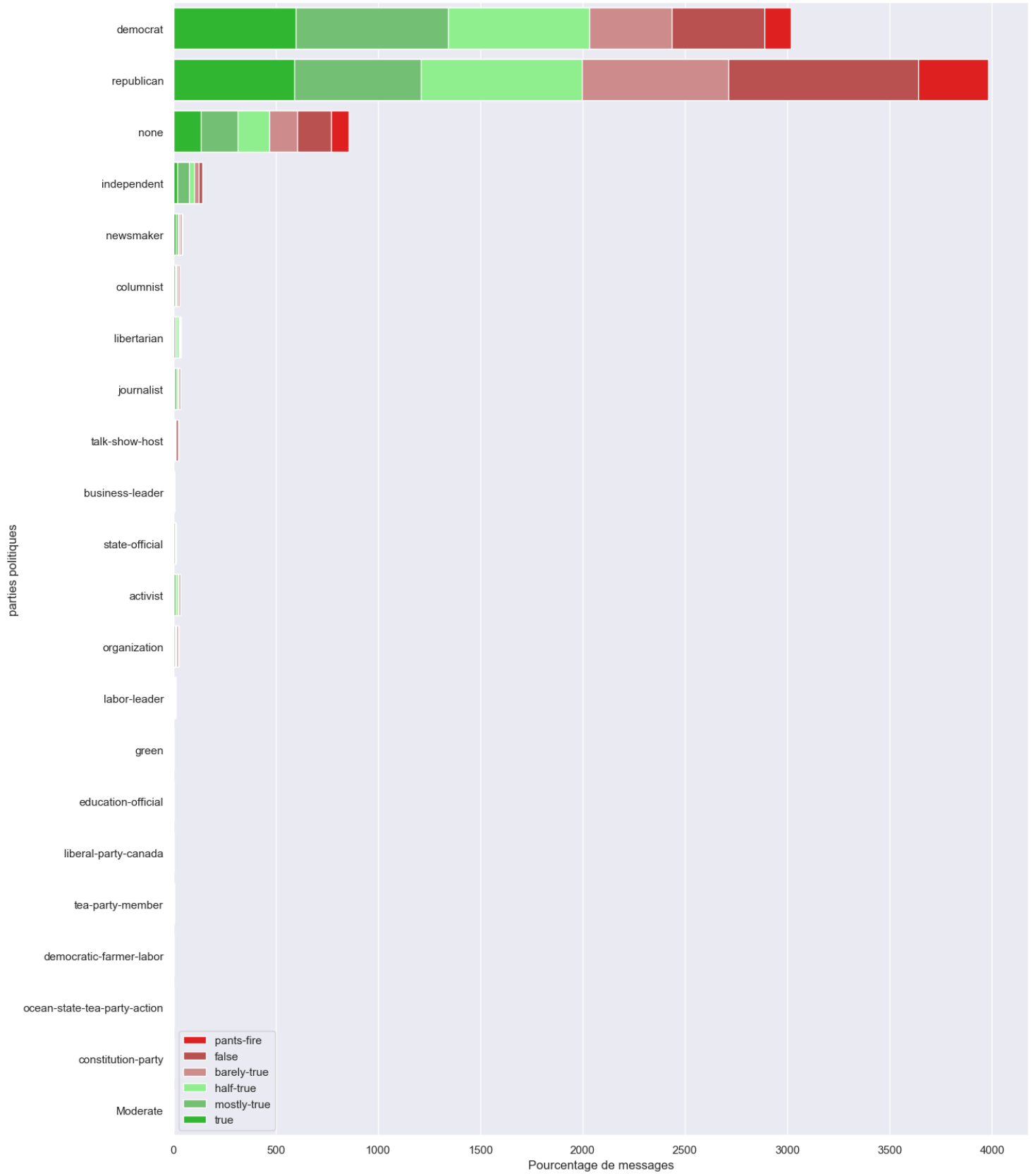


FIGURE 12 : HISTOGRAMME DES LABELS DE VERITE PAR PARTIS POLITIQUES

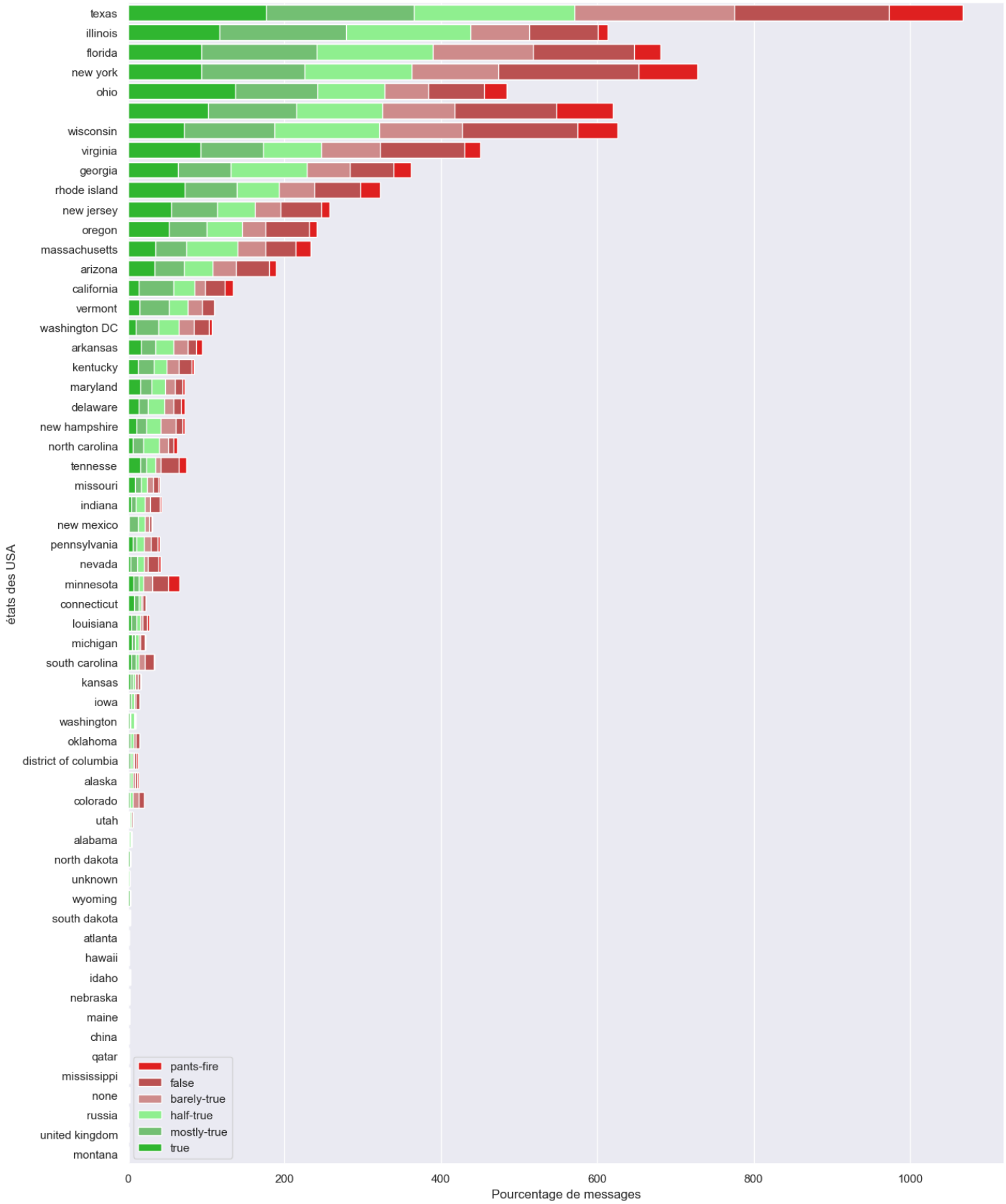


FIGURE 15 : HISTOGRAMME DES LABELS DE VERITE PAR ETAT DE NAISSANCE.

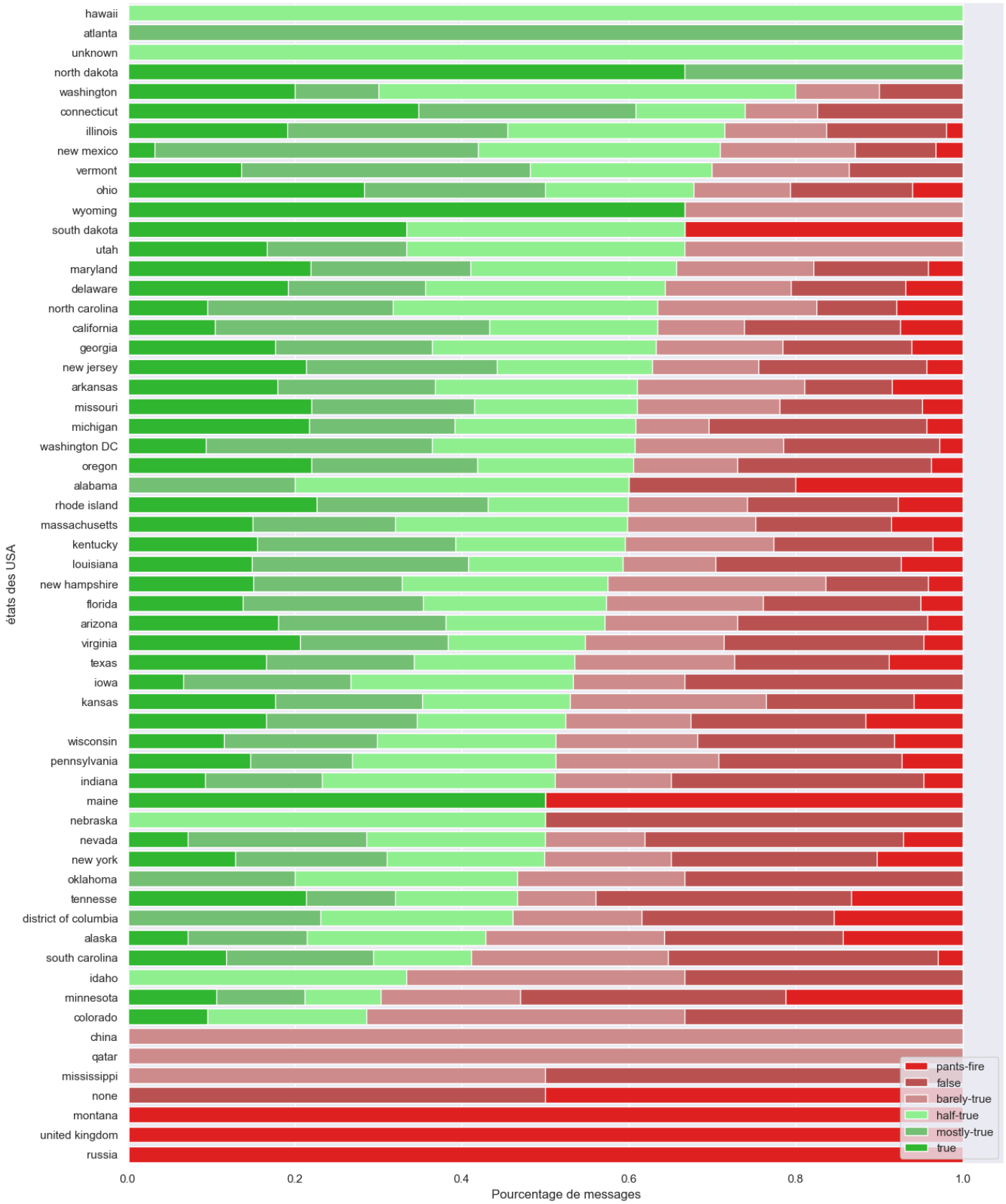


FIGURE 16 : PROBABILITE DES LABELS DE VERITE PAR ETAT DE NAISSANCE.

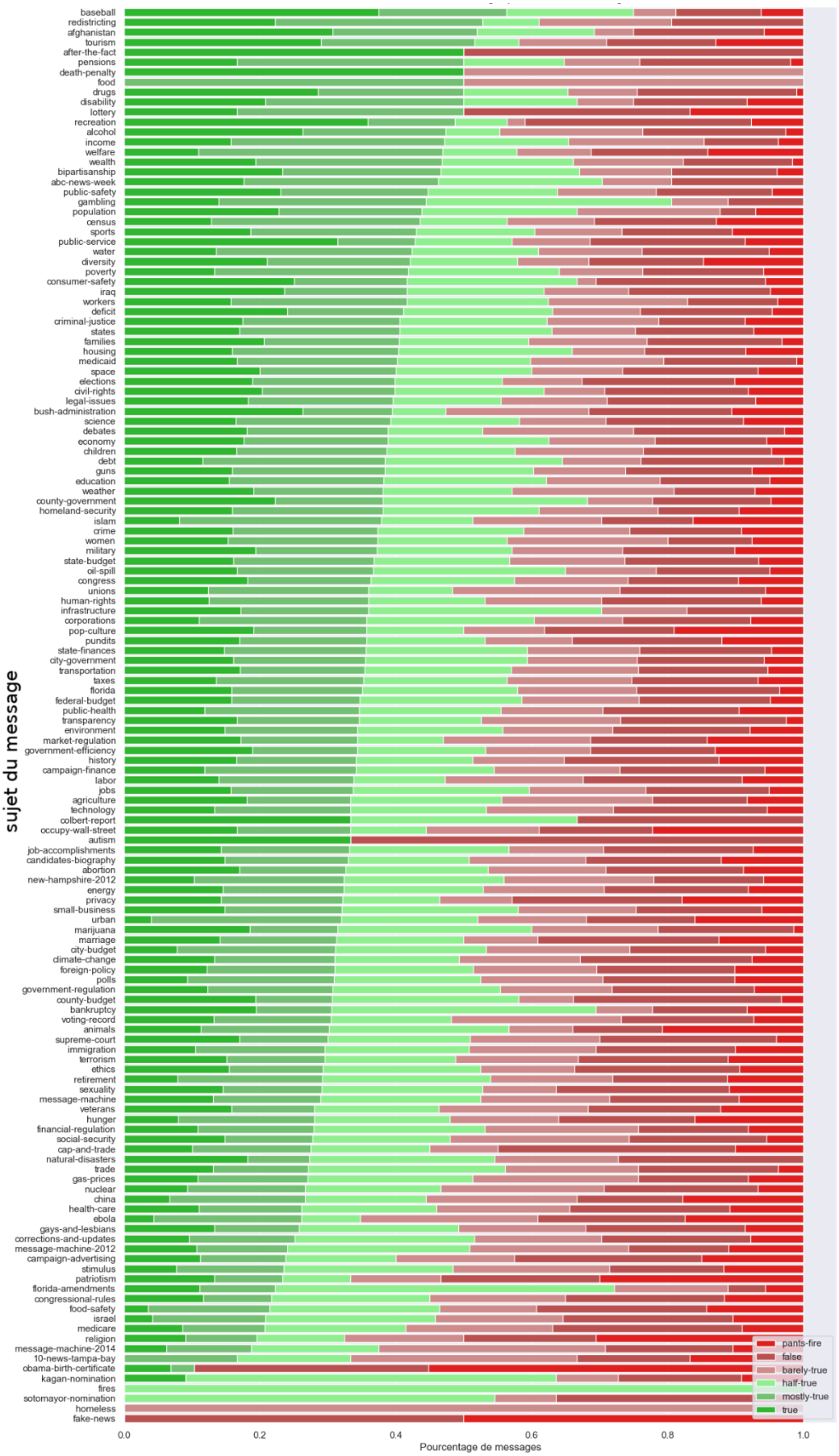


FIGURE 17 : PROBABILITE DES LABELS DE VERITE PAR SUJET.