Université de Lille

Faculté des Sciences Economiques, Sociales et des Territoires

Devoir Introduction à l'Analyse de Données Textuelles

« Analyse des caractéristiques linguistiques d'articles de presse traitant des avantages et des risques liés à l'utilisation de l'Intelligence Artificielle. »

Elauïne BERNARD

15 mai 2023

« Analyse des caractéristiques linguistiques d'articles de presse traitant des avantages et des risques liés à l'utilisation de l'Intelligence Artificielle. »

Plan

1- Mise en contexte

Dans le cadre du cours d'Introduction à l'Analyse de Données Textuelles, l'évaluation consiste à présenter un dossier de travail effectué sur les avis des internautes concernant un film ou sur des articles de presses. L'objectif principal de ce travail est de mettre en application les analyses du cours et d'essayer d'utiliser à bon escient les instructions et les packages R exposés en cours.

En ce sens, nous présentons ce travail portant sur le thème de l'Intelligence Artificielle. Plus précisément, notre sujet consiste en une analyse des caractéristiques linguistiques de 12 articles de presse traitant des avantages et des risques liées à l'utilisation de l'Intelligence Artificielle.

2- Problématique

Le Parlement européen définit l'intelligence artificielle comme tout outil utilisé par une machine afin de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité. L'Intelligence Artificielle, couramment appelée IA, est l'un des sujets les plus prisés actuellement dans le monde. Ce qui est tout à fait compréhensible, compte tenu de son implication grandissante dans les différents aspects de notre vie. On attribue à l'Intelligence Artificielle de nombreux progrès dans les domaines de la santé, le commerce, l'agriculture, le transport, la finance, le marketing, l'éducation, etc. Malgré les nombreux éloges, l'Intelligence Artificielle fait l'objet de nombreux critiques négatifs.

Plusieurs articles de presses font des parallèles entre les avantages et les risques liés à l'utilisation de l'intelligence artificielle. Bon nombre d'entre eux traitent des avantages et risques liés à l'utilisation de l'intelligence artificielle dans un domaine particulier. La lecture de plusieurs de ces articles nous a donné l'impression qu'il existe des similitudes de vocabulaires intéressantes entre ces articles malgré le fait qu'ils parlent de domaines d'activité différents. Ainsi, nous avons décidé d'analyser les caractéristiques linguistiques d'une douzaine d'articles de presse qui traitent des avantages et des risques liés à l'utilisation de l'Intelligence Artificielle dans les domaines de la santé, l'agriculture, l'éducation et la finance. De manière spécifique nous allons essayer de répondre à ces deux questions : **Quelles sont les noms, les pronoms personnels et les verbes les plus utilisés pour parler des avantages et risques liés à l'utilisation de l'intelligence artificielle ? Les mots utilisés dans ces articles varient-ils en fonction du domaine d'activité** ?

3- Méthodologie du travail

Pour répondre à notre problématique nous avons partagé le travail en trois grandes parties :

- Dans la première partie du travail nous avons constitué un corpus de textes avec tous les articles. Ensuite, ce corpus est traité de manière à devenir exploitable.
- Dans la deuxième partie du travail nous avons procédé à une analyse globale de l'ensemble du corpus. Pour cela, nous avons d'abord construit un lexique général et ensuite nous avons étudié les mots importants découlant de notre sujet.
- Dans la troisième partie du travail, nous avons procédé à une comparaison de sous-corpus des domaines d'activité. Pour cela, nous avons créé un sous-corpus de textes pour chaque domaine d'activité et comparé les mots communs et les mots différenciants les domaines d'activité.

4- Construction et Traitement du Corpus

Dans cette partie du travail nous allons tout d'abord construire le corpus de texte avec les articles de presses. Ensuite nous allons traiter le corpus de texte crée.

4.1- Construction du Corpus

Pour construire notre corpus, 12 articles de presse rédigés en français ont été récupéré par copier-coller depuis des sites web vers Word. Chaque article a été enregistré dans un document Word. Pour chacun des 12 documents certaines modifications ont été effectué dans les articles sur Word : suppression des images, suppressions des liens web, séparation des mots collés et transformation de chaque phrase de l'article en paragraphe. Nous avons également classé les 12 articles en 4 catégories en fonction du domaine d'activité sur lequel le sujet de l'article est focalisé. Ces 4 catégories sont : agriculture, santé, éducation et finance.

Par la suite ces 12 articles ont été importé sur R dans un dataframe. Ce dataframe constitue notre corpus de texte. Et Il contient 613 lignes. Chaque ligne étant une phrase d'un article.

Lien tableau donnant les détails sur les articles

Lien script R création dataframe du corpus

4.2- Traitement du Corpus

Nous avons effectué quelques traitements sur le corpus de textes crée. Ces traitements consistent principalement à mettre les textes en minuscule, enlever les ponctuations, mettre au singulier certains mots liés à la problématique, enlever les blancs inutiles et enlever les mots outils. Soulignons que nous avons enlevé les pronoms personnels de la liste des mots outils.

Lien script R traitement du corpus

5- Analyse globale de l'ensemble du corpus

Dans cette partie nous allons procéder à une analyse globale du corpus contenant les 12 articles. Tout d'abord, nous allons construire un lexique global. Ensuite, nous allons faire une représentation graphique du lexique. Puis, nous allons faire une brève commentaires des mots les plus fréquents du lexique. Finalement, nous allons faire une étude des relations entre des mots qui sont liés à notre problématique.

5.1- Construction d'un Lexique globale

Nous avons construit, à partir du corpus de texte créé et traités, un tableau de lexique des mots de l'ensemble des 12 articles. Une fois créé sur R, ce lexique a été exporté et visualisé sur Excel. Cette visualisation sur Excel avait pour objectif de vérifier l'existence de mots collés dans le lexique. Nous avons trouvé environ 50 mots collés dans le lexique et nous avons effectué les traitements nécessaires. Finalement, nous avons abouti à un lexique contenant 3450 mots différents.

Lien script R construction du lexique global

Lien extrait lexique global

5.2- Représentation graphique du Lexique global

Nous avons importé le lexique traité sur R à nouveau et réaliser un graphique nuage de points pour représenter les 300 mots les plus fréquents du lexique global.

Graphique 1- Nuage des mots les plus fréquents du lexique global

D'après le graphique 1 le mot « données » est le mot le plus fréquent du lexique.

Lien script R nuage de mots lexique global

5.3- Commentaires sur les mots les plus fréquents du Lexique global

Dans cette partie nous allons faire quelques commentaires sur les mots qui sont très fréquents dans le lexique global et qui sont liés à notre problématique. Il s'agit des mots suivants : données, intelligence, artificielle, technologie, numérique, algorithme, apprentissage, risque, avantage, être, pouvoir, devoir, elles, nous, elle.

- Le mot le plus fréquent est « ***données*** ».Il est mentionné 107 fois. Les données constituent l'élément de base de l'Intelligence Artificielle. Ce sont les données qui permettent à un système d'intelligence artificielle d'apprendre et de reproduire les comportements humains. Alors dès qu'on parle de l'intelligence artificielle, les données sont l'un des premiers éléments auxquels on fait référence. Nous nous attendions à ce que le mot « données » soit parmi les plus fréquents.
- Les mots « ***Intelligence*** » et « ***Artificielle*** » sont respectivement les 3ème et 5ème mot les plus utilisés. Ils sont respectivement mentionnés 71 et 62 fois. Ce qui est tout à fait compréhensible vue que les articles ont pour thématique de fonds l' intelligence artificielle.
- Les mots « ***technologie*** », « ***numérique*** », « ***algorithme*** », « ***apprentissage*** » apparaissent au moins 36 fois. Ce que l'on peut comprendre, vue ces mots sont en général utilisés pour décrire le fonctionnement de l'intelligence artificielle.
- Le mot « ***risque*** » est le 7ème mot le plus fréquent du lexique. Il apparait 58 fois. Alors que le mot « ***avantage*** » apparait seulement 18 fois. Il est aussi important de souligner que certaines fois le mot « risque » est employé pour parler d'un avantage de l'intelligence artificielle. A titre d'exemple, nous avons extrait cette phrase du corpus de textes :

  [21] " L'IA pourrait ainsi contribuer à améliorer la rapidité, la qualité et la pertinence de la sélection des données déclarées et leur transmission à l'autorité concernée, en signalant ou en « auto-corrigeant » les anomalies éventuelles (erreurs, champs vides, etc.) ce qui pourrait avoir un impact sur la mesure du **risque** prudentiel des établissements concernés."

- Les verbes « ***être*** », « ***pouvoir*** » et « ***devoir*** » sont les plus fréquents du lexique. Les verbes « être » et « devoir » sont souvent employé dans la même phrase. Et assez souvent, ces trois verbes sont utilisés dans le corpus pour parler des capacités de l'intelligence artificielle ou pour donner des consignes sur comment tirer des avantages en utilisant l'intelligence artificielle et se protéger des risques liés à l'utilisation de l'intelligence artificielle. A titre d'exemple, nous avons extrait ces 3 phrases du corpus de textes :

  [55] " L'intelligence artificielle **peut** permettre de gérer rapidement d'immenses quantités de données de façon beaucoup plus efficace qu'un humain, à condition qu'on sache bien l'utiliser."
  [71] " Chaque information **doit** d'abord **être** mise en forme pour être correctement analysée par le système."
  [78] "La cybersécurité **doit être** au rendez-vous"

- Les pronoms « ***elles*** », « ***nous*** » et « ***elle*** » sont les plus fréquents du lexique. Dans le lexique « elle » apparait 38 fois, « nous » apparait 25 fois et « elles » apparait 20 fois dans le lexique. Certaines fois « nous » est utilisé pour montrer que tout le monde est concerné par l'intelligence. D'autres fois « nous » est utilisé pour parler d'un groupe qui expérimente l'utilisation de l'intelligence artificielle. Certaines fois, « elle » est utilisée pour parler de l'intelligence artificielle et « elles » pour parler des données. A titre d'exemple, nous avons extrait ces 3 phrases du corpus de textes :

  [9] "La règlementation permet d'ailleurs d'ores et déjà le recours à certaines utilisations de l'IA, comme c'est le cas des assistants bancaires et du conseil robotisé que **nous** avons déjà rencontrés."

  [22] " **Nous** avons des hackers éthiques à nos côtés ainsi que des systèmes de surveillance intelligents, capables de détecter les comportements suspects avant une attaque massive."

```
[166] "\" Pour avoir une bonne intelligence artificielle, il faut qu'elle fonctionne
avec un bon algorithme."

[170] " A contrario, lorsque les données sont précises et bien formulées, elles peuv
ent même avoir leur place dans un essai clinique."
```

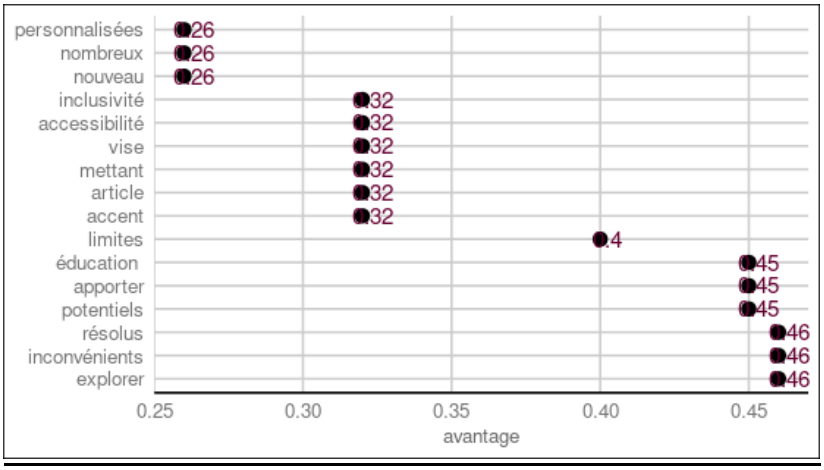
Lien script R extraction phrase avec un mot spécifique

5.4- Etude des relations entre les mots

Dans cette partie, nous allons étudier les relations entre les mots liés à notre problématique. Dans un premier temps nous allons visualiser les mots ayant une association d'au moins 25% à avantages et les mots ayant une association d'au moins 25% à risque dans les articles. Dans un deuxième temps nous allons calculer les corrélations entre des mots se référant à avantage, risque, intelligence artificielle, certains verbes et le pronom nous.

5.4.1- Visualisations des mots associés à avantage et des mots associer à risque

Graphique 2- Visualisation des mots associés à avantage



D'après le graphique 2, il y a 16 mots qui ont une association de 25% au moins avec le mot « avantage ». Les mots « résolu », « inconvénients » et « explorer » sont les plus associés au mot « avantage ». Ils ont une association de 46% avec le mot « avantage ». Chacun de ces 3 mots est mentionné une seule fois dans le corpus. Et le mot « avantage » est dans cette unique phrase du corpus. Cela pourrait expliquer le niveau élevé d'association.

Dans les sous-titres des articles sur l'éducation l'expression « avantages et limites dans l'éducation » est utilisé plusieurs fois. Cela explique le fait que l'association entre limites et avantage soit 40% et l'association entre éducation et avantage soit 45%. Il en est de même des mots « personnalisées », « inclusivité » et « accessibilité » qui sont utilisé au moins deux fois avec le mot « avantage » dans des sous-titres des articles sur l'éducation.

Graphique 3- Visualisation des mots associés à risque

D'après le graphique 3, il y a 53 mots qui ont une association de 25% au moins avec le mot « risque ». Les mots « spécifiques » et « maitrise » sont les plus associés au mot « risque ». Ils ont une association de 38% avec le mot « risque ». Ces mots sont utilisés assez souvent pour mettre l'accent sur le mot « risque », surtout lorsqu'il s'agit de gérer les risques. On trouve dans le corpus des expressions comme : « maîtriser les risques spécifiques », « risques spécifiques ».

Parmi les mots ayant une association de 33% avec le mot « risque », on retrouve les mots « risque », « ia », . Certaine fois le mot « risque » est utilisé deux fois dans la même phrase. On trouve certaine fois le mot « ia » dans la même phrase que le mot « risque » pour spécifier que l'on parle des risques liés à l'utilisation de l'IA.

Le graphique montre également la présence de mots que l'on emploie couramment pour parler des risques liés à l'utilisation de l'IA. Ils ont une association avec le mot « risque » allant de 27% à 33%. A titre d'exemple nous pouvons citer ces mots du graphique : « règlementation », « gouvernance », « terrorisme », « prudentielle », « divulgation », « criminelle »

Lien script R visualisation associations avec avantage et avec risque

5.4.2- Calcul des corrélations entre des mots

Certaine fois pour parler de l'intelligence artificielle, de ses avantages et de ses risques on n'emploie pas spécifiquement les mots « intelligence », « artificielle », « avantage », « risque ». On emploie des mots qui font référence à eux. Voilà pourquoi nous avons décidé de vérifier si c'est le cas dans notre corpus. Nous allons calculer les corrélations entre des mots qu'on utilise pour faire référence aux mots « intelligence », « artificielle », « avantage », « risque ». Également nous allons étudier les relations entre ces mots et les trois verbes les plus fréquents du lexique. Nous allons considérer les corrélations de 10% au moins.

Pour les mots faisant référence à l'intelligence artificielle, nous avons considéré les mots suivants : données, numérique, technologie, automatique, algorithme, apprentissage, révolution, futur, décision. Pour les mots faisant référence à avantage, nous avons considéré les mots suivants : mieux, intelligence, améliorer, aider, opportunité, valeur, possibilité. Pour les mots faisant référence à risque, nous avons considéré les mots suivants : défis, éthique, transparence, confidentialité, cyberattaque.

- Corrélations entre des mots référant à avantage et à intelligence artificielle :

La corrélation est de 11% entre améliorer et artificielle, 11% entre algorithme et décision, 13% entre automatique et technologie, 14% entre technologie et possibilité, 18% entre valeur et numérique, 39% entre apprentissage et automatique. D'après ces chiffres nous pouvons avancer que certaines fois les articles peuvent parler des avantages de l'utilisation de l'intelligence artificielle sans mentionner le mot « avantage ». Un exemple du corpus avec le mot « possibilité » :

```
[3] " En rendant l'éducation plus accessible et inclusive, la technologie de l'IA peut aide
r à garantir que tous les élèves ont la possibilité d'apprendre et de réussir."
```

- Corrélations entre des mots référant à risque et à intelligence artificielle :

La corrélation est de 10% entre cyberattaque et risque, 10% entre numérique et transparence, 18% entre artificielle et éthique, 21% entre défis et futur, 21% entre éthique et confidentialité, 27% entre données et confidentialité. D'après ces chiffres nous pouvons avancer que certaines fois les articles peuvent parler des risques de l'utilisation de l'intelligence artificielle sans mentionner le mot « risque ». Un exemple du corpus avec « éthique » et « confidentialité » :

```
[3] "Préoccupations éthiques liées à la confidentialité des données et aux préjugés"
```

- Corrélations entre des mots référant à avantage et les verbes être, pouvoir et devoir :

La corrélation est de 15% entre améliorer et peut, 21% entre peut et être, 24% entre valeur et être, 34% entre doivent et être. D'après ces chiffres nous pouvons avancer que le verbe « pouvoir » est utilisé pour parler des avantages qu'il est possible de tirer de l'IA sans mentionner le mot comme « avantage ». Un exemple du corpus avec le mot « améliorer :

```
[5] " Cela peut contribuer à améliorer les résultats des élèves et à réduire la charge de t
ravail des enseignants."
```

- Corrélations entre des mots référant à risque et les verbes être, pouvoir et devoir :

La corrélation est de 14% entre éthique et doivent, 16% entre éthique et être, 21% entre confidentialité et doivent. D'après ces chiffres nous pouvons avancer que le verbe « devoir » est utilisé pour parler des risques liés à l'utilisation de l'IA sans mentionner le mot « risque ». Un exemple du corpus avec le mot « confidentialité » :

```
[7] " Les établissements d'enseignement doivent prendre des mesures pour garantir que la co
nfidentialité des données et les considérations éthiques reçoivent l'attention voulue et qu
e des garanties appropriées sont mises en place pour protéger les données des étudiants et
prévenir toute conséquence involontaire de l'utilisation de l'IA."
```

Lien script R corrélation entre des mots référant à intelligence artificielle, avantage, risque, et verbes

6. Comparaison de sous-corpus

Dans cette partie du travail nous allons comparer le sous-corpus des domaines d'activité. Tout d'abord, nous allons créer les sous-corpus des domaines d'activité. Ensuite, nous allons créer le lexique des domaines d'activités. Finalement, nous allons comparer les articles des différents domaines d'activité.

6.1- Création du sous-corpus des domaines d'activité

Pour cela nous avons créé 4 sous-corpus. Un sous-corpus est créé par domaine d'activité. On a ainsi le sous-corpus agriculture, le sous-corpus éducation, le sous-corpus finance et le sous-corpus santé. Chaque sous-corpus contient 3 articles. Ces sous-corpus sont créés sur R.

Lien script R création sous-corpus

6.2- Création du lexique des domaines d'activité

L'objectif ici est de créer un lexique contenant une colonne pour chaque domaine d'activité. Nous avons utilisé les sous-corpus des domaines d'activité pour construire le lexique des domaines d'activité.

D'après le lexique des domaines créé, le mot « numérique » est le plus fréquent pour le sous-corpus agriculture, le mot « enseignants » est le plus fréquent pour le sous-corpus éducation, le mot « données » est le plus fréquent pour le sous-corpus finance et le sous-corpus santé.

Lien script R construction du lexique des domaines d'activité

Lien extrait du lexique des domaines d'activité

6.3- Comparaison des domaines d'activité

Pour comparer les articles des différents domaines nous allons nous baser sur les mots qu'ils ont en commun et les mots qui les différencient.

- Représentation graphique des 100 premiers mots communs entre les domaines

Graphique 4 : Nuage de mots communs aux 4 domaines d'activité

D'après le graphique 6, on peut voir que le mot « données » est le premier mot commun aux différents domaines. Ce mot apparait 12 fois dans le sous-corpus agriculture, 24 fois dans le sous-corpus éducation, 43 fois dans le sous-corpus finance et 28 fois dans le sous-corpus santé. Le verbe « être » est le premier verbe que les sous-corpus ont en commun. Et Le mot « risque » apparait commun l'un des premiers mots qu'ils ont en commun.

Lien script R nuage de mots communs

- Représentation graphique des 200 premiers mots différenciants les domaines

Graphique 5 : Nuage de mots différenciant les 4 domaines d'activité

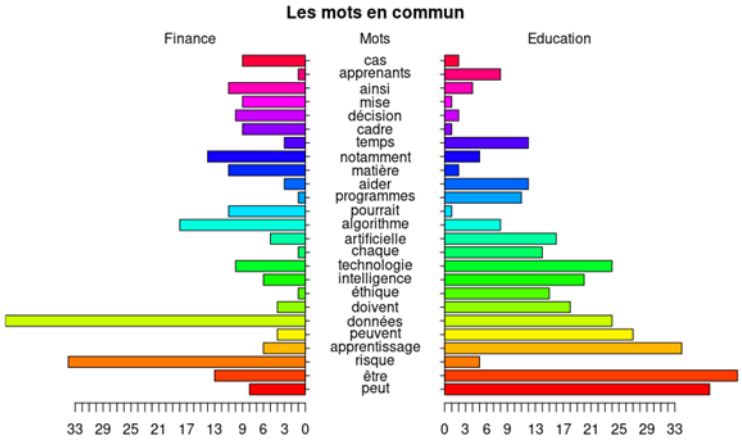
Le graphique 5 montre que :

- Les mots « ferme » et « tech » distinguent le sous-corpus agriculture des trois autres.
- Les mots « apprentissage », « élèves », « fournir » distinguent le sous-corpus éducation des trois autres.
- Les mots « données », « risque », « crédit » et « banque » distinguent le sous-corpus finance des trois autres
- Les mots « santé », « patients », « soins » et « médecin » distinguent le sous-corpus santé des trois autres

Lien script R nuage de mots de différences

- Comparaison entre sous-corpus Finance et sous-corpus Education

Graphique 6 : Les 25 premiers mots communs entre Finance et Education



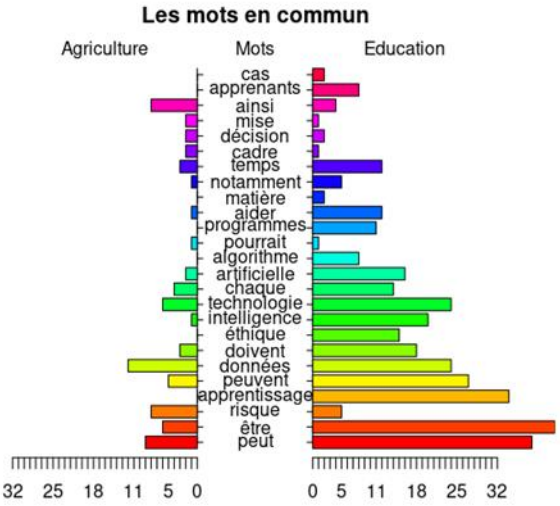
Le graphique 6 montre que parmi les 25 premiers mots communs entre les domaines éducation et finance il y a des mots liés à notre sujet : « décision », « algorithme », « technologie », « données », « apprentissage », « éthique », « risque », « intelligence », artificielle ». Toutefois, ces mots ne sont pas forcément utilisés dans le même contexte pour les 2 domaines. Par exemple, dans le domaine finance les mots « apprentissage » et « programmes » font surtout référence à l'intelligence artificielle. Alors que dans le domaine

éducation, ils renvoient à des pratiques scolaires. Dans le domaine éducation le mot « risque » fait référence surtout aux risques liés à l'utilisation de l'intelligence artificielle. Alors que dans le domaine finance, il renvoie surtout à des risques de crédit. Certains de ces mots sont employés plus fréquemment dans un domaine que d'autre. Par exemple, le mot « apprentissage » est utilisé 8 fois plus dans le domaine éducation que dans le domaine de la finance et le mot « risque » est utilisé 8 fois plus dans le domaine éducation que dans le domaine de la finance. Ces 2 domaines ont en commun les verbes « être », « pouvoir », « devoir » et « aider ».

Lien script R graphique mots communs Finance et Education

- Comparaison entre sous-corpus Agriculture et sous-corpus Education

Graphique 7 : Les 25 premiers mots communs entre Agriculture et Education

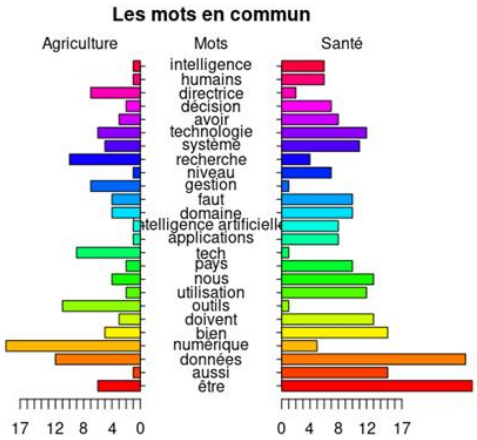


Le graphique 7 montre que parmi les 25 premiers mots communs entre les domaines agriculture et éducation il y a des mots liés à notre sujet : « décision », « algorithme », « technologie », « intelligence », « artificielle », « risque », « données », « éthique ». La plupart de ces mots sont utilisés plus fréquemment dans le domaine éducation que dans le domaine agriculture. Le mot « risque » est l'unique mot lié à notre sujet commun entre agriculture et éducation qui soit le plus utilisé dans le domaine de l'agriculture. Ces 2 domaines ont en commun les verbes « être », « pouvoir », « devoir » et « aider ».

Lien script R graphique mots communs Agriculture et Education

- Comparaison entre sous-corpus Agriculture et sous-corpus Santé

Graphique 8 : Les 25 premiers mots communs entre Agriculture et Santé

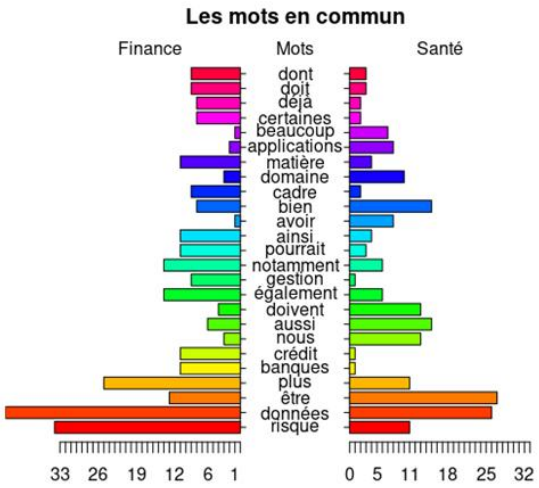


Le graphique 8 montre que parmi les 25 premiers mots communs entre les domaines agriculture et santé il y a des mots liés à notre sujet : « décision », « intelligence », « artificielle », « numérique », « technologie », « données ». La plupart de ces mots sont plus fréquents dans le domaine santé que dans le domaine agriculture. On peut remarquer que parmi les premiers communs entre agriculture et santé il n'y a aucun qui fait directement référence au mot « risque ». Ces 2 domaines ont en commun les verbes « être », « devoir », « avoir », « falloir » et « aider ». Ils ont aussi en commun le pronom personnel « nous ».

Lien script R graphique mots communs Agriculture et Santé

- Comparaison entre sous-corpus Finance et sous-corpus Santé

Graphique 9 : Les 25 premiers mots communs entre Finance et Santé



Le graphique 9 montre que parmi les 25 premiers mots communs entre les domaines agriculture et santé il y a des mots liés à notre sujet : « risque » et « données ». On peut remarquer qu'il y a très peu de mots liés à notre problématique qui soient parmi les premiers mots communs entre finance et santé. On peut remarquer également que parmi les premiers communs entre agriculture et santé il n'y a aucun qui fait directement référence au mot « risque ». Ces 2 domaines ont en commun les verbes « être », « devoir », « pouvoir » et « avoir ». Ils ont aussi en commun le pronom personnel « nous ».

Lien script R graphique mots communs Finance et Santé

7. Conclusion

Dans le cadre de ce travail notre objectif consistait à analyser les caractéristiques linguistiques de 12 articles de presse traitant des avantages et des risques liés à l'utilisation de l'Intelligence Artificielle dans les domaines de la santé, l'agriculture, l'éducation et la finance.

Dans la première partie de notre travail nous avons créé un lexique pour l'ensemble des 12 articles, analyser le lexique et étudier les relations entre les mots liés à notre sujet. L'analyse de ce lexique a permis de faire les constats suivants : le lexique contient 3450 mots différents ; le mot « données » est le mot le plus fréquent ; les verbes « être », « pouvoir » et « devoir » sont les plus fréquents du lexique ; les pronoms « elles », « nous » et « elle » sont les plus fréquents du lexique ; le mot « risque » est 3,22 fois plus fréquent que le mot « avantage ». L'étude des relations entre les mots a révélé que :16 mots ont une association de 25% au moins avec le mot « avantage » ; 53 mots ont une association de 25% au moins avec le mot « risque » ; certaines fois les articles peuvent parler des avantages et des risques liés à l'utilisation de l'intelligence artificielle sans mentionner le mot « avantage » et le mot « risque », des mots comme « améliorer », « éthique », « confidentialité » sont employés de préférences.

Dans la deuxième partie du travail nous avons comparer les 4 domaines d'activités pour lesquels les avantages et les inconvénients liés à l'utilisation de l'intelligence artificielle ont été traité. La comparaison a montré d'une part que les mots « données », « risque », « technologie » et « être » sont parmi les premiers mots que les domaines ont en commun. D'autre part la comparaison a montré que les articles se diffèrent par des mots propres aux domaines. Par exemple le mot « crédit » pour la finance, le mot « médecine » pour la santé, le mot « élève » pour l'éducation et le mot « ferme » pour l'agriculture. On peut remarquer certaine fois que le contexte de l'emploi de certains mots liés à notre sujet peut varier d'un domaine à un autre. Par exemple l'emploi du mot « risque » en finance et en éducation. Il est à souligner aussi que la fréquence de l'emploi de certains mots peut varier largement d'un domaine à un autre. Par exemple le mot « données » est surtout fréquent en santé et en finance, le mot « éthique » en éducation et le mot « numérique » en agriculture.

8. Annexes

Tableau 1 : Liste des articles de l'études

Titre Article	Site	Date de publication	Nombre de mots Avant	Nombre de mots Après	Domaine Activité
Les opportunités et les risques rattachés au développement du numérique dans l'agriculture	L'Usine Nouvelle	1 mars 2022	1063	1028	Agriculture
Drones, tracteurs autonomes La nouvelle révolution de l'agriculture	L'Express	28 septembre 2022	1385	1337	Agriculture
Comment le numérique pourrait bien modifier en profondeur l'agriculture	Sciences et Avenir	2 mars 2022	1109	1034	Agriculture
L'intelligence artificielle peut-elle remplacer les enseignants ?	Management & Data Science	11 mars 2023	1430	1232	Education
L'IA dans l'éducation : avantages et limites	Gobookmart	11 février 2023	1390	1299	Education
L'intelligence artificielle, des enjeux pour l'éducation et la formation	Formiris	3 décembre 2021	1874	1627	Education
Vers une utilisation responsable de l'intelligence artificielle dans la finance	Mind Fintech	9 novembre 2022	2305	1867	Finance
Intelligence artificielle dans le secteur bancaire et financier	Les Echos	3 mars 2021	2931	2831	Finance
Intelligence artificielle dans le secteur bancaire et financier	Revue Banque	7 octobre 2022	339	336	Finance
Santé Globale : les défis à relever pour utiliser l'IA en médecine	Sciences et Avenir	25 novembre 2022	853	838	Santé
Les freins à l'IA en médecine sont davantage humains que techniques	Sciences et Avenir	6 mai 2022	1761	1636	Santé
Rapport mondial sur l'intelligence artificielle (IA) appliquée à la santé et six principes directeurs relatifs à sa conception et à son utilisation	OMS	28 juin 2021	1217	1189	Santé

Tableau 2- Extrait lexique global

mot	freq
données	107
être	88
intelligence	71
plus	69
artificielle	62
peut	62
risque	58
technologie	52
augmenté	49
élèves	43
algorithme	42
peuvent	42
utilisation	42
santé	40

Liste des packages

```
library(tm)
library(qdap)
library(qdapTools)
library(readtext)
library(SnowballC)
library(wordcloud)
library(ggplot2)
library(ggthemes)
library(plotrix)
```

Script R Création dataframe du corpus

```
Text <- read_docx('AGRI_Express_2022.docx')
N <- length(Text)
titre <- rep("AGRI_Express_2022",N)
Doc1 <- data.frame(titre,Text)
Text <- read_docx('AGRI_Sciences_et_Avenir_2022.docx')
N <- length(Text)
titre <- rep("AGRI_Sciences_et_Avenir_2022",N)
Doc2 <- data.frame(titre,Text)
Text <- read_docx('AGRI_Usine_Nouvelle_2022.docx')
N <- length(Text)
titre <- rep("AGRI_Usine_Nouvelle_2022",N)
Doc3 <- data.frame(titre,Text)
```

```
Text <- read_docx('EDU_Formiris_2021.docx')
N <- length(Text)
titre <- rep("EDU_Formiris_2021",N)
Doc4 <- data.frame(titre,Text)
Text <- read_docx('EDU_Gobookmart_2023.docx')
N <- length(Text)
titre <- rep("EDU_Gobookmart_2023",N)
Doc5 <- data.frame(titre,Text)
Text <- read_docx('EDU_Management_DS_2023.docx')
N <- length(Text)
titre <- rep("EDU_Management_DS_2023",N)
Doc6 <- data.frame(titre,Text)
Text <- read_docx('FINC_Les_Echos_2021.docx')
N <- length(Text)
titre <- rep("FINC_Les_Echos_2021",N)
Doc7 <- data.frame(titre,Text)
Text <- read_docx('FINC_Mind_Fintech_2022.docx')
N <- length(Text)
titre <- rep("FINC_Mind_Fintech_2022",N)
Doc8 <- data.frame(titre,Text)
Text <- read_docx('FINC_Revue_Banque_2022.docx')
N <- length(Text)
titre <- rep("FINC_Revue_Banque_2022",N)
Doc9 <- data.frame(titre,Text)
Text <- read_docx('SANTE_2_Sciences_et_Avenir_2022.docx')
N <- length(Text)
titre <- rep("SANTE_2_Sciences_et_Avenir_2022",N)
Doc10 <- data.frame(titre,Text)
Text <- read_docx('SANTE_OMS_2021.docx')
N <- length(Text)
titre <- rep("SANTE_OMS_2021",N)
Doc11 <- data.frame(titre,Text)
Text <- read_docx('SANTE_Sciences_et_Avenir_2022.docx')
N <- length(Text)
titre <- rep("SANTE_Sciences_et_Avenir_2022",N)
Doc12 <- data.frame(titre,Text)
Doc <- rbind.data.frame(Doc1, Doc2, Doc3, Doc4, Doc5, Doc6, Doc7, Doc8, Doc9, Doc10, Doc11, Doc12)
nrow(Doc)
```

Script R traitement du corpus

```
# 2.1- Créer un vecteur qui contiendra les phrases des articles
Phrase<-Doc$Text
# 2.2- Dictionnaire de mots outils
# 2.2.1- Création d'un dictionnaire de mots outils
mots_outils <- stopwords("french")
mots_outils
#2.2.2- Enlever des mots-outils du dictionnaire
# Nous allons enlever les pronoms personnels du dictionnaire.
pronoms <- c("nous")
M <- length(pronoms)
N <- length(mots_outils)
mots_outils2 <- mots_outils
# imbrication d'une boucle
for(i in 1:N) {
  for(j in 1:M) {
    if(mots_outils[i]==pronoms[j]) {
      mots_outils2[i] <- ""
      j <- M+1
    }
  }
}
mots_outils2

# 2.3- Apauvrir les textes
# 2.3.1- Mettre les textes en minuscules
Phrase <- tolower(Phrase)
Phrase[400]
# 2.3.2- Enlever les mots outils
Phrase <- removeWords(Phrase,mots_outils2 )
Phrase[400]
## 2.3.3 Remplacer les ponctuations par des blancs
enlever <- c("," , "!" , "?" , "." , ":" , ";" , "/", "'", "’", "…", "«", "»")
remplacer <- c(" " , " " , " " , " " , " " , " " , " ", " ", " ", " ", " ", " ", " ", " ")
Phrase <- multigsub(enlever, remplacer, Phrase)
Phrase[400]
# 2.3.4- Remplacer certains mots au pluriel par leur singulier
# 12 mots très lié au problématique
pluriel <- c("intelligences" , "artificielles" , "risques" , "technologies" , "numériques" , "algorithmes" , "apprentissages", "décisions", "informations", "éthiques", "avantages", "systèmes")
singulier <- c("intelligence" , "artificielle" , "risque" , "technologie" , "numérique" , "algorithme" , "apprentissage", "décision", "information", "éthique", "avantage", "système")
```

```
Phrase <- multigsub(pluriel, singulier, Phrase)
```

Script R Création lexique global

```
# transformation de vecteur Phrase en table
y <- data.frame(doc_id=seq(1:nrow(Doc)), text=Phrase)
# l'objet corpus
corpus_global <- SimpleCorpus(DataframeSource(y), control = list(language = "fr"))
# Enlever les nombres en chiffres
corpus_global <- tm_map(corpus_global, removeNumbers)
# Enlever la ponctuation
corpus_global <- tm_map(corpus_global, removePunctuation)
# Enlever les blancs inutiles
corpus_global <- tm_map(corpus_global, stripWhitespace)
# le tableau lexical entier global
tdm_global <-TermDocumentMatrix(corpus_global, control = list(encoding="latin1"))
# on le transforme en objet matrice pour faire des calculs de fréquences
tdm_global.mat <-as.matrix(tdm_global)
# Obtenir la dimension de la matrice
dim(tdm_global.mat)
# Obtenir la fréquence de chaque mot
term.freq <- rowSums(tdm_global.mat)
# Création de la table du lexique globale
lexic_global <-data.frame(mot=names(term.freq), freq=term.freq)
# on sauve le lexique dans un classeur excel
write.csv2(lexic_global, "lexic_global.csv", fileEncoding="latin1",row.names = F)
```

Script R Nuage de mots lexique global

```
# importer lexique global traité
lexfreq<- read.csv2(file="lexic_global1.csv", encoding="latin1" )
# on passe en minuscules, plus lisible dans un nuage de mots
lexfreq$mot <- tolower(lexfreq$mot)
pal <- brewer.pal(8, "Purples")
pal <- pal[-(1:4)]
wordcloud(lexfreq$mot,lexfreq$freq, max.words = 300, random.order=FALSE, colors=pal)
```

Script R extraction textes avec un mot spécifique

```
# Extraction de textes avec le mot risque
```

```
RISQUE <- grep("risque", Doc$Text, ignore.case=TRUE)
NN <- length(AVANTAGE)
lire2 <- rep("text",NN)
for(i in 1:NN) {
  j <- RISQUE[i]
  lire2[i] <- Doc$Text[j]
}
lire2
```

9. Webographie

Supports de cours

https://www.cnil.fr/fr/intelligence-artificielle/intelligence-artificielle-de-quoi-parle-t-on

Liens des 12 articles