RAPPORT DE PROJET INF473G

Analyse du discours autour du changement climatique dans la presse

9 juin 2024

Ivain GUITTARD
Martin BEAUFILS

ÉCOLE POLYTECHNIQUE

IP PARIS

TABLE DES MATIÈRES

1 Introduction **3**

2 Présentation des données et outils **3**

3 Première approche **3**

4 Deuxième approche **4**

5 Visualisations et interprétations **4**
5.1 Localisation 6
5.2 Événements 7
5.3 Produits 8
5.4 Graphe 9
5.4.1 Communauté liée au débat politique américain 10
5.4.2 Communauté liée aux industries 11
5.4.3 Deux dernières communautés 12

6 Exploitation de WikiData pour étudier les principaux pollueurs **12**
6.1 Graphe concernant les pollueurs 12
6.2 Mention des pollueurs dans les articles 14

7 Conclusion, pistes d'amélioration et de poursuite du projet **16**

1 INTRODUCTION

Au cours de ce projet, nous avons mis en œuvre les techniques apprises lors du cours **INF473G** pour analyser des bases de données d'articles de presse. L'objectif de ce travail est d'examiner la manière dont le réchauffement climatique est abordé dans les médias. Nous avons cherché à identifier les thèmes principaux et les acteurs du changement climatique dans les articles de presse.

Grâce à la restructuration des articles sous forme de graphes, ce travail a permis de repérer les tendances et les lacunes dans la couverture médiatique des journeaux étudiés entre 2016 et 2020.

2 PRÉSENTATION DES DONNÉES ET OUTILS

L'étude qui suit s'est essentiellement basée sur deux bases de données d'articles, un classement des principaux industriels polluant, la bibliothèque python spaCy, et deux modèles d'analyse de texte de ClimateBERT.

La première base de données est composé d'environ 2 Go d'articles publiés par CNBC (que vous pouvez retrouver ici). C'est une chaîne de télévision américaine spécialisée dans les actualités financières et économiques.

La seconde base de données, All the news 2.0 (que vous pouvez retrouver ici) est composée d'environ 2.7 millions d'articles (environ 8 Go) parus entre 2016 et 2020 dans des publications américaines.

Le classement des princpaux pollueurs est issu d'un article de The Guardian (que vous pouvez retrouver ici). Grâce à Wikidata, nous avons étoffé les informations à propos de certains pollueurs afin d'en étudier les activités.

Enfin, en termes d'outils, nous avons utilisé la bibliothèque spaCy. spaCy est une bibliothèque de traitement du langage naturel (NLP) en Python. Elle offre des fonctionnalités pour le traitement de texte telles que la tokenization, le part-of-speech tagging, la lemmatisation, la reconnaissance d'entités nommées, et l'analyse syntaxique. Nous avons également utilisé deux modèles ClimateBERT. ClimateBERT est une série de modèles spécialisés dans l'analyse des textes liés au changement climatique. Ces modèles, basés sur BERT, sont entraînés sur des données climatiques spécifiques pour offrir une meilleure compréhension et analyse des textes dans ce domaine. Nous avons utilisé un modèle détectant si un texte est en lien avec le changement climatique et un autre indiquant le sentiment du texte vis-à-vis du changement climatique. Nos principaux codes sont disponibles https://github.com/martinbfls/modalINF473G.

3 PREMIÈRE APPROCHE

Notre première approche consiste à nous baser sur notre liste des plus gros pollueurs. L'objectif est ainsi de rechercher parmis les articles de nos bases de données ceux mentionnant ces plus gros pollueurs. La première difficulté fut de trouver le moyen de traiter autant de volume de données (environ 10 Go). Après avoir essayé sans succès d'utiliser les Notebook GoogleColab, nous avons choisi d'utiliser les machines de l'école plus performantes.

Après avoir retenu les articles mentionnant les plus gros pollueurs, nous avons réalisé la reconnaissance d'entité et de triplés de la forme (sujet, prédicat, objet) à l'aide de spaCy afin de cosntruire un graphe représentatif de ces articles.

Cependant, une fois les noeuds et triplés construits il ne restait que très peu de noeuds reliés aux pollueurs ce qui rendait finalement cette approche peu pertinente. En effet, il nous était alors impossible, de mettre en

lumière des liens pertinents entre les pollueurs et le changement climatique. Nous avons donc choisi de changer d'approche qui nous a permis d'obtenir des résultats exploitables.

4 DEUXIÈME APPROCHE

La deuxième approche utilise comme point de départ le modèle de ClimateBERT détectant les texte à propos du changement climatique. Au cours de cette approche nous avons uniquement utilisé la base de données All the news 2.0.

Nous avons tout d'abord utilisé le modèle climate_related de ClimateBERT. Ce modèle permet de détecter si un texte est en lien avec le changement climatique ou non. Ce modèle est calibré pour traiter des textes de 512 tokens maximum et est assez couteux en termes de temps de calcul. C'est pourquoi, compte tenu du temps qu'il nous restait, nous avons choisi d'appliquer ce modèle uniquement sur les titres des articles de la base de données All the news 2.0.

Ensuite, nous avons procédé à une reconnaissance d'entités et de triplés en utilisant la bibliothèque spaCy. Nous avons ainsi pu obtenir un début d'un graphe. La taille des fichiers étant de l'ordre de la vingtaine de megaoctets. Le nombre d'arcs et de noeuds était de l'ordre de la centaine de milliers. Afin, d'obtenir un graphe plus épuré et plus facilement représentable nous avons utilisé PostgreSQL afin de traiter les données. Nous nous sommes tournés vers cet outil pour des raisons de performances. En effet, les requêtes par Neo4j étaient beaucoup trop longues à s'exécuter.

Lors de la reconnaissance d'entités, spaCy associe pour chaque entité un type. Nous avons donc choisi de nous intéresser à 7 types d'entités en particuliers : PERSON, ORG, NORP, LOC, PRODUCT, EVENT et WORK OF ART. Pour chaque type, nous avons classé les entités en fonction de leur degré. Nous avons ensuite gardé les entités les plus importantes de manière à ce que le sous ensemble de ce type représente 75% des occurences du type dans la base de données globale.

Ainsi, en ne gardant que les triplés impliquant ces entités, nous avons une base de données restreinte représentative des principaux centres d'intérêts des articles en liens avec le changement climatique de notre base de données initiale. Cela nous a permis d'obtenir un graphe représentable avec environ 500 noeuds pour environ 18 000 arcs. Enfin, nous avons également appliqué l'algorithme de Louvain sur ce graphe restreint pour en détecter les communautés.

5 VISUALISATIONS ET INTERPRÉTATIONS

Après avoir détecté dans le titre les articles en rapport avec le changement climatique, nous avons appliqué un autre modèle de ClimateBERT permettant de le sentiment vis à vis au changement climatique dans le texte. Les textes sont classifiés en trois catégorie. Ils peuvent associer le changement à une opportunité, à un risque ou peuvent être neutres. La difficulté fut que ces modèles sont calibrés pour des textes de 512 tokens maximum donc, comme expliqué précédement, nous les avons appliqué uniquement sur les titres.

Figure 1 – Répartition des résultats des modèles ClimateBERT

Ainsi, nous remarquons une faible proportion d'articles en lien avec le climat dans la base de données All the news 2.0. De plus, la majeure partie de ces articles ont une position neutre vis-à-vis du changement climatique ou le considère comme une opportunité. Ce sujet ne semble donc finalement que peu traité relativement à l'importance qu'il va avoir dans nos modes de vie. Les articles de notre base de données ne paraissent donc pas, à première vue, participer de manière active à mettre en garde contre les risques du changement climatique.

Afin d'avoir une compréhension plus fine des articles retenus nous avons étudié la répartition des noeuds par types d'entités selon leurs degré.

5.1 Localisation

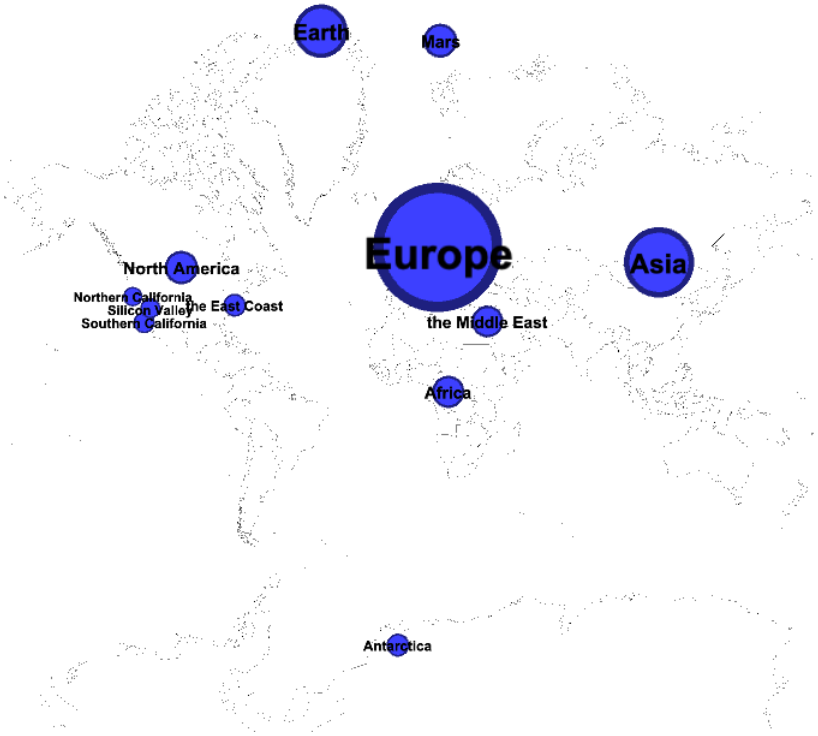

Figure 2 – Principales localisations mentionnées dans les articles en lien avec le changement climatique

Ce graphe, réalisé avec Gephi, où les nœuds représentent les principales localisations mentionnées dans les articles, montre que la taille des nœuds est proportionnelle au nombre d'occurrences de chaque localisation. Le fond du graphe est un planisphère, permettant de visualiser la répartition géographique des mentions.

La base de données utilisée est principalement concentrée sur le monde occidental, ce qui est compréhensible étant donné que les articles proviennent de All the news 2.0 : ils sont issus de publications américaines. Cela reflète une certaine limitation dans la couverture géographique de la presse américaine, focalisée sur des régions spécifiques. Un élément notable de cette analyse est la présence de mentions de Mars, indiquant, dès ce stade de notre analyse, une déconnexion des tendances climatiques de la presse avec des enjeux climatiques directs et concrets.

5.2 Événements

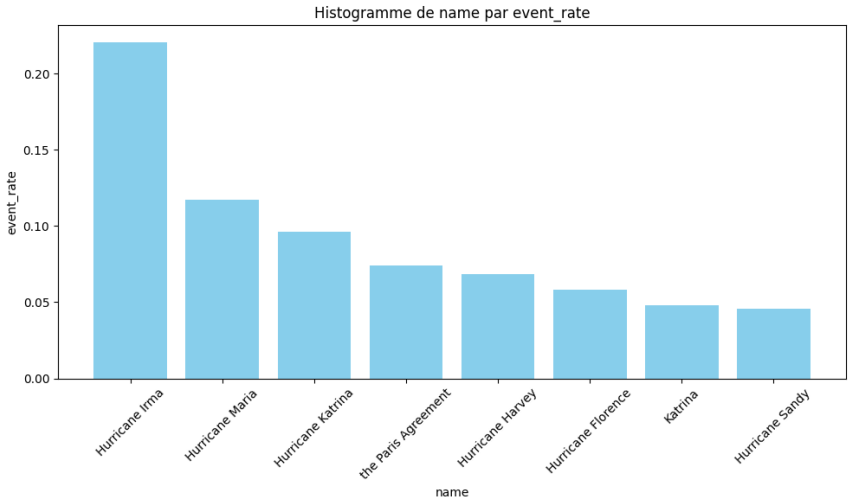

Figure 3 – Principaux événements mentionnés dans les articles en lien avec le changement climatique

Pour étudier la répartition des principaux événements, nous avons réalisé un histogramme selon le degré des noeuds de ce type. Il en ressort que les ouragans représentent la quasi totalité des événements répertoriés. En revanche, on observe une absence notable de mentions de catastrophes écologiques spécifiques autres que les ouragans. Cette focalisation sur les ouragans peut indiquer une tendance des articles à se concentrer sur les événements climatiques immédiats et spectaculaires plutôt que sur les crises écologiques plus larges et chroniques.

L'Accord de Paris (Paris Agreement) apparaît également fréquemment. Sa présence sera étudiée plus en détails avec l'aide du graphe principal.

5.3 Produits

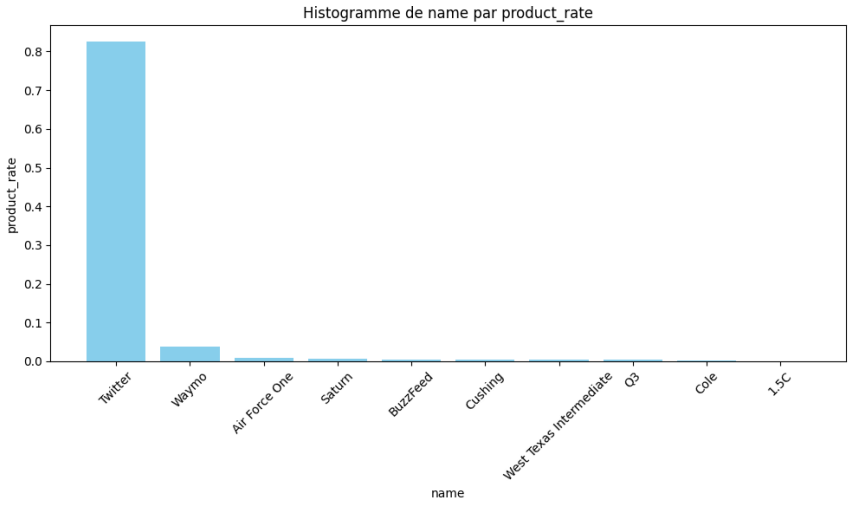

Figure 4 – Principaux produits mentionnés dans les articles en lien avec le changement climatique

Du côté des produits, il a été remarqué une très faible proportion de mentions concernant le charbon (coal) et l'objectif de limitation du réchauffement climatique à 1.5°C. Cela suggère que les articles ne contiennent pas d'avertissements notables contre les énergies fossiles, contrairement à ce que l'on aurait pu attendre.

5.4 Graphe

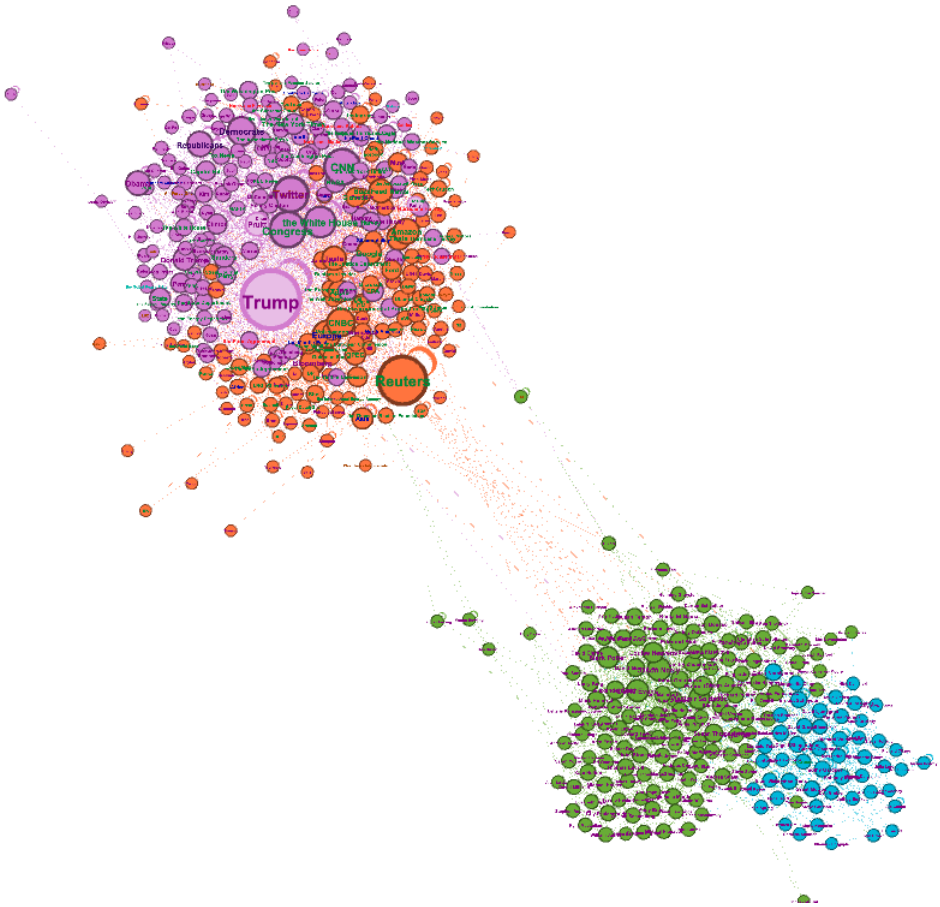

Figure 5 – Graphe principal

Voici la représentation du graphe global obtenue par Gephi. La couleur des noeuds correspond à leur communauté et celle des labels à leur type. La taille des noeuds et des labels est proportionnelle à leur degré. Nous allons maintenant analyser ce graphe en nous concentrant sur ses communautés.

5.4.1 • Communauté liée au débat politique américain

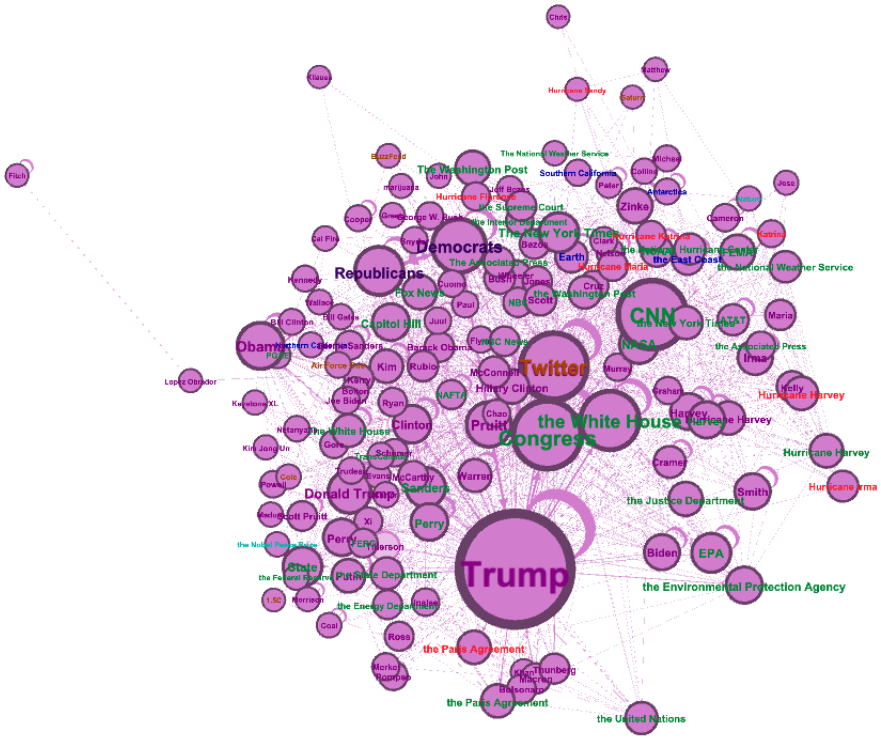

Figure 6 – Communauté liée au débat politique américain

Cette communauté gravite essentiellement autour de Donald Trump et des sujets politiques aux États-Unis. On peut donc supposer que le changement climatique y est principalement abordé comme un sujet de débat politique. L'entité Paris Agreement (Accord de Paris) y est présente. Elle est essentiellement lié à Trump. Il apparait donc cohérent de supposer que ce lien vient de la décision de Trump de retirer les États-Unis de cet accord. Les différents ouragans des principaux événements apparaissent également dans cette communauté.

Ainsi, il semble que cette communauté ne présente pas de tendance marquée vers une mise en garde contre le changement climatique ni vers une critique des modes de consommation actuels. Au lieu de cela, les discussions se concentrent davantage sur les aspects politiques et les événements climatiques spécifiques.

5.4.2 • Communauté liée aux industries

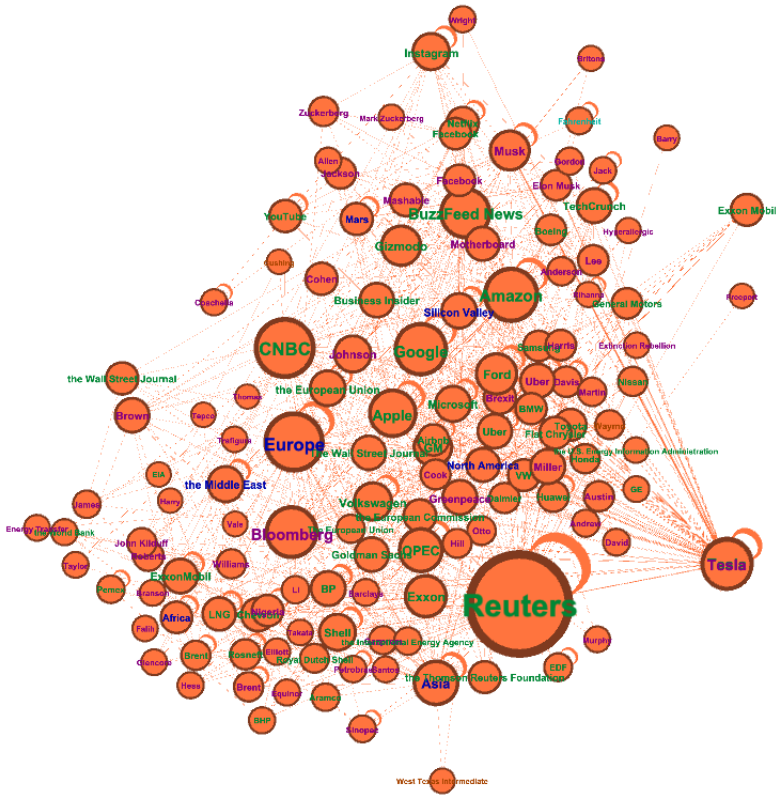

Figure 7 – Communauté liée aux industries

La deuxième communauté identifiée concerne principalement des entités de type organisation et entreprise, avec des références spécifiques à des localisations. Notamment, on y trouve des marques de voitures et des pollueurs notoires comme Exxon Mobil et Shell. Cela permet de supposer que ces articles traitent davantage des activités industrielles polluantes. Cette communauté semble donc se concentrer sur les aspects industriels et économiques du changement climatique, en particulier sur les entreprises et leurs contributions à la pollution ce qui est un petit peu plus satisfaisant.

5.4.3 • Deux dernières communautés

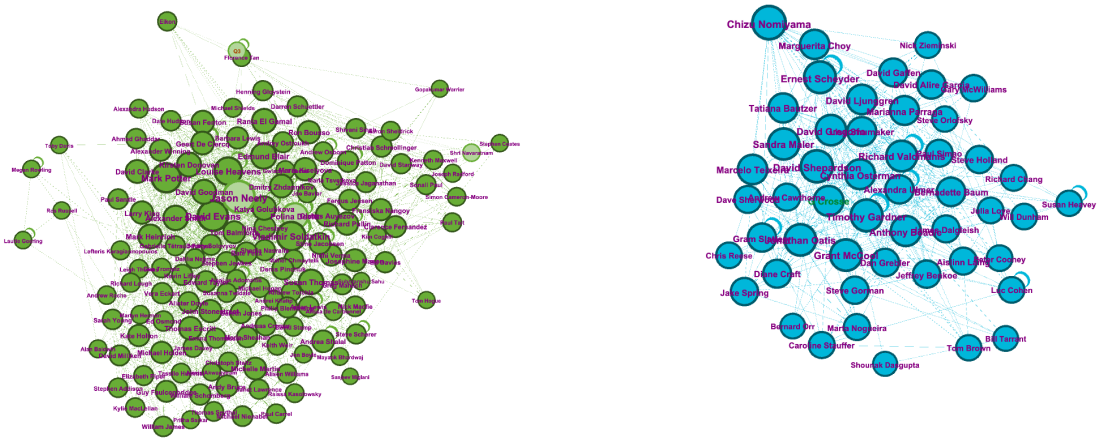

Figure 8 – Communautés liées à des personnalités

Enfin, les deux dernières communautés identifiées impliquent des personnalités que nous ne connaissons pas suffisamment pour pouvoir les analyser avec précision. Par conséquent, il est difficile de tirer des conclusions claires sur la nature des articles dans ces communautés.

En somme, l'analyse du graphe montre seulement une légère tendance vers des articles qui pourraient réellement mettre en garde contre le réchauffement climatique et nos modes de consommation. La majorité des articles semblent plutôt axés sur des débats politiques, des activités industrielles polluantes, et des aspects économiques, avec une attention limitée portée aux avertissements directs sur l'urgence climatique et la nécessité de changer nos modes de vie.

6
EXPLOITATION DE WIKIDATA POUR ÉTUDIER LES PRINCIPAUX POLLUEURS

6.1 Graphe concernant les pollueurs

En cherchant des informations complémentaires sur Wikidata sur nos pollueurs, nous avons récolté les informations suivantes : alias, secteur d'activité, pays, place de cotation, revenu annuel, appartenance à des groupes d'entreprises. Malheureusement, parmi notre liste de pollueurs, tous n'étaient pas présents sur Wikidata : nous avons pu récupérer des informations pour seulement 54 d'entre eux. L'intégration de ces données nous a permis de compléter notre analyse en apportant un contexte supplémentaire sur les entreprises et en facilitant l'identification des liens entre ces pollueurs et les mentions dans les articles de presse analysés. Après avoir récupéré les données pertinentes de Wikidata sous forme de fichier JSON, nous avons entrepris de construire un graphe représentatif des relations entre les entreprises pollueuses et les informations extraites. Chaque nœud du graphe représente une entité (par exemple une entreprise, un secteur, un pays), et chaque arête représente une relation entre deux entités (par exemple, une entreprise opérant dans un certain secteur).

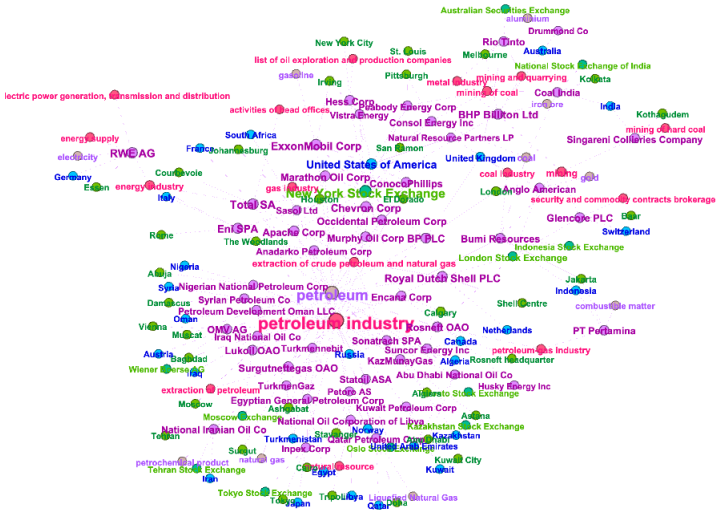

Figure 9 – Graphe des entreprises et leurs caractéristiques (Wikidata)

Nous avons ensuite appliqué un filtre pour garder les noeuds ayant un degré supérieur à 5, afin de mettre en évidence les entités les plus importantes :

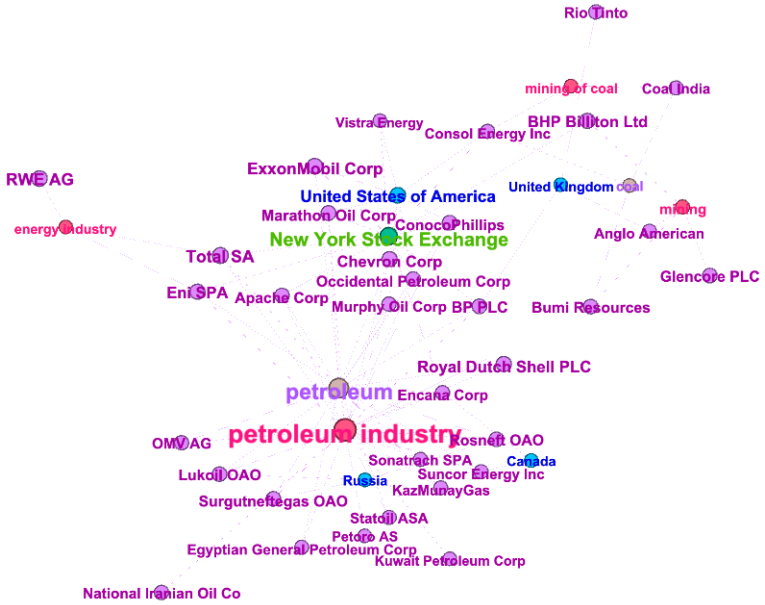

Figure 10 – Graphe des entreprises et leurs caractéristiques (Wikidata), degré supérieur à 5

Il apparait ainsi clairement que l'industrie du pétrole, suivie par les secteurs du charbon, de l'énergie et du

minage sont les industries les plus polluantes.

De plus, le graphe a mis en évidence que la plupart de ces entreprises sont cotées à New York, soulignant l'importance de cette place de cotation pour les grands pollueurs mondiaux.

En termes de localisation géographique, on remarque les principaux pollueurs sont situés dans des pays développés.

6.2 Mention des pollueurs dans les articles

Nous avons aussi utilisé les alias pour identifier et retracer le nombre de mentions de chaque pollueur dans les articles portant sur le changement climatique. Cette approche nous a permis d'obtenir des données quantitatives sur la visibilité médiatique de chaque pollueur, offrant ainsi es informations sur leur perception publique et leur implication dans les débats sur le climat :

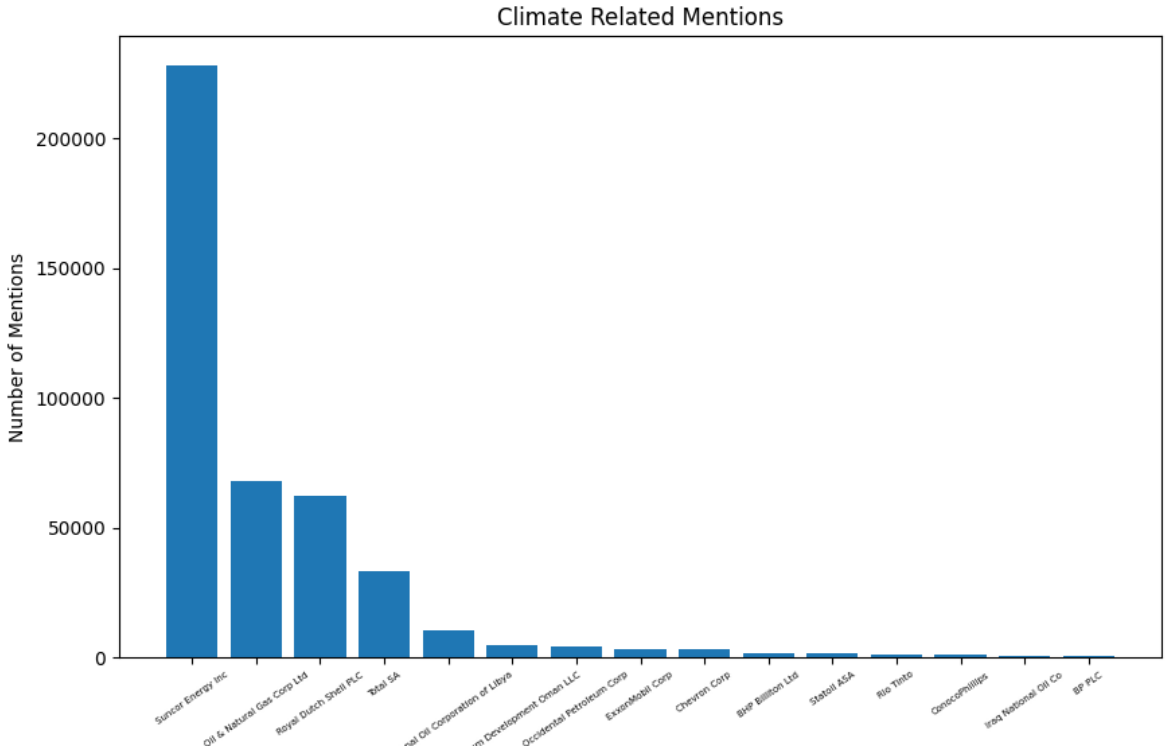

Figure 11 – Mentions des pollueurs dans les articles relatifs au climat

On peut remarquer la prédominance de Suncor Energy Inc, suivi par une entreprise d'huile et de gaz indienne, de Shell et de Total.

Dans un second temps, en alliant recherche des pollueurs et les sentiments fournis par ClimateBERT, nous avons pu obtenir le nombre de mentions par pollueurs associées aux différents labels : opportunity, neutral et risk.

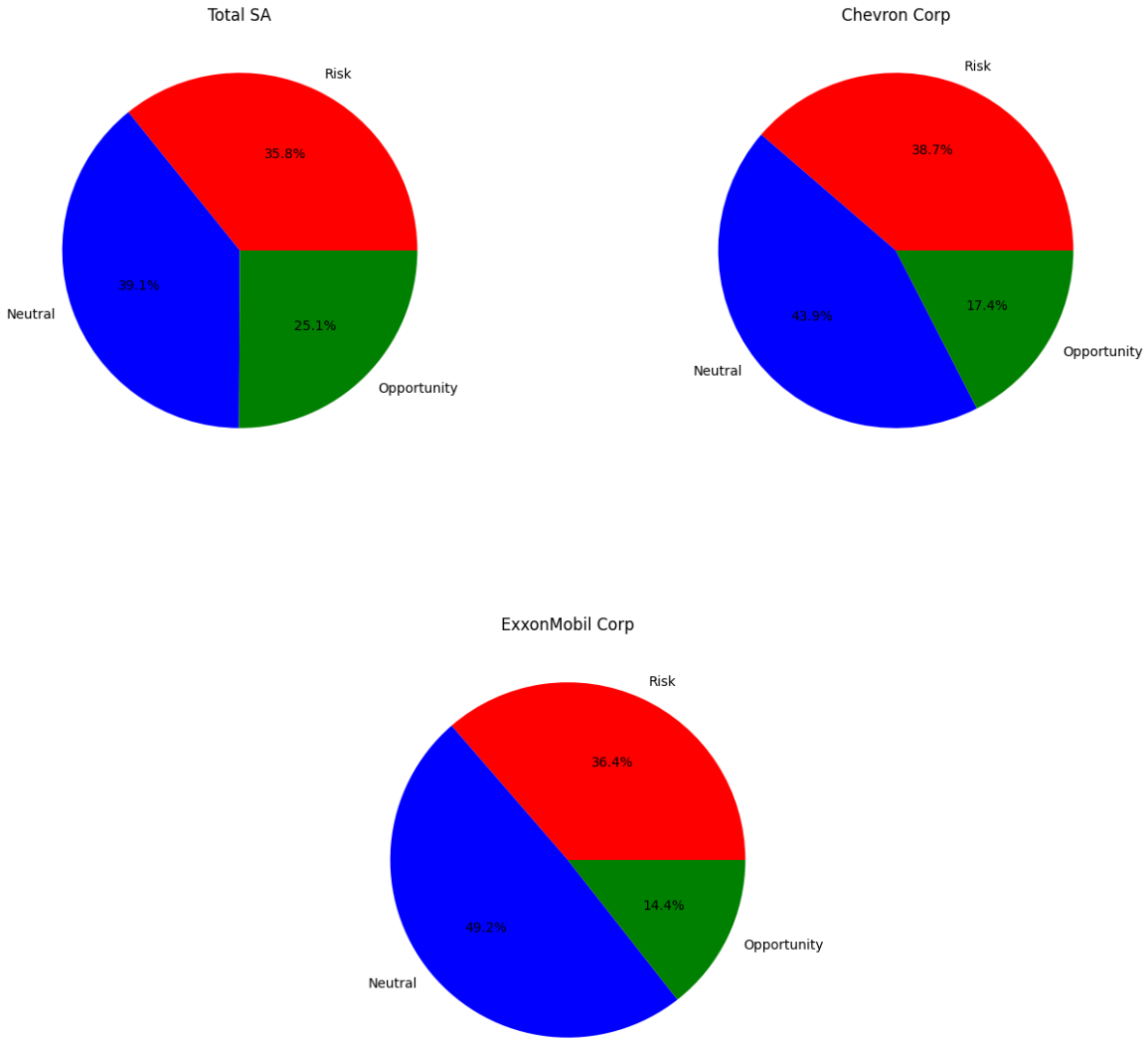

Figure 12 – Proportions risk/neutral/opportunity des pollueurs (liste non exhaustive)

On remarque alors, pour certains pollueurs, une proportion d'articles associant le changement climatique à un risque plus grande que sur l'ensemble des articles. La proportion d'articles l'associant à une opportunité est, elle, plus faible. Cela dénote tout de même finalement d'une association dans la presse de ces pollueurs aux risques qu'ils engendrent. Néanmoins, il reste une grande part d'articles qualifiés de neutres qui mériteraient d'être étudiés en détails.

7
CONCLUSION, PISTES D'AMÉLIORATION ET DE POURSUITE DU PROJET

En conclusion, l'analyse des articles de presse concernant le changement climatique révèle plusieurs tendances notables. Le sujet du changement climatique n'est présent que dans une faible partie des articles. Cette sous-représentation est d'autant plus préoccupante que le changement climatique est un enjeu majeur qui va influencer de manière significative nos sociétés. De plus, l'absence de considération de nombreuses conséquences du réchauffement climatiques telle que la question des migrations climatiques ou l'augmentation des maladies liées à la pollution est regrettable. Ces conséquences sont pourtant un aspect crucial de la crise climatique. Enfin, la remise en cause des activités des plus gros pollueurs industriels n'apparait que très peu. Ces omissions contribuent à une vision partielle et incomplète du problème. En conséquence, on peut conclure que la presse de 2016 à 2020, ou du moins cette sélection d'articles, semble montrer une forme d'inaction en matière de sensibilisation et d'incitation à un changement d'attitude global pour lutter contre le réchauffement climatique.

Cependant, ces interprétations doivent être nuancées. Elles représentent des tendances globales de la base de données et non une analyse détaillée de chaque article pris individuellement. Pour améliorer et approfondir cette analyse, plusieurs pistes peuvent être envisagées. Il serait pertinent d'affiner l'analyse en étudiant plus précisément les prédicats utilisés dans les articles, ce qui permettrait une compréhension plus fine des discours et des relations entre les entités. Par ailleurs, étendre la base de données, affiner le tri initial des articles pertinents ou reproduire l'étude sur une base d'articles plus récente pourrait enrichir les résultats et rendre notre analyse plus pertinente. Enfin, appliquer le modèle ClimateBERT sur l'ensemble des articles, et non seulement sur les titres, offrirait une analyse plus exhaustive et précise des contenus.

En résumé, cette étude met en lumière un certain nombre de lacunes dans la couverture médiatique du changement climatique entre 2016 et 2020, tout en proposant des pistes pour poursuivre et améliorer l'analyse dans le futur.