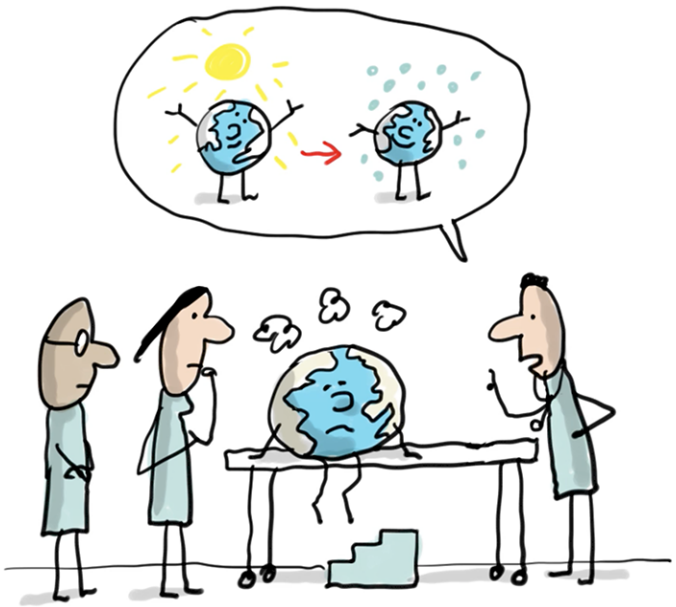
AUGUSTYN Patricia
BRISSET Lise
SAUTREAU Laurine

Sorbonne Nouvelle
université des cultures

LES BRAS PAS CASSÉS

PROJET : TEXT MINING

LZML041



SERRAR Loubna

Année universitaire : 2022-2023

PLAN DU PROJET

I. Introduction

A. Choix du thème

Le Text Mining est un ensemble de techniques qui permettent de trouver des informations dans des corpus choisis. Ces informations nous permettent d'analyser les résultats obtenus à l'aide de certaines fonctions. Par exemple, il existe l'analyse de sentiment où l'on peut identifier les émotions exprimées par l'auteur, ou encore détecter automatiquement les éléments importants du texte comme la fréquence.

Par ailleurs, cette méthode nous permet aussi d'explorer des corpus volumineux et cela est un avantage considérable pour notre projet. Nous avons décidé de travailler sur le thème de l'écologie et plus particulièrement sur un compte Twitter : ONU Climat France (https://twitter.com/CCNUCC). Nous pensons que l'écologie est un sujet d'actualité en ce qui concerne les périodes : le passé, le présent et le futur. Le thème de l'écologie est très présent chez l'ONU et pose énormément de questions sur la situation actuelle. Nous sommes impactées par de nombreux désastres écologiques dont le réchauffement climatique, et nous faisons face de plus en plus à des phénomènes écologiques violents dont la fonte de glace etc… Le fait que les États Membres de l'Organisation des Nations Unies prennent conscience de ce sujet est donc important dans la prévision des actions pour améliorer la situation. Ainsi, nous allons voir si l'utilisation de ce réseau social est une stratégie pour transmettre des informations.

Pour cela nous avons utilisé la technique du web scraping qui permet de récupérer des tweets sur internet. Il était possible de récupérer tous les tweets mais nous voulions seulement reprendre des tweets dans une période précise : depuis 2015 à fin 2022. Ainsi, cette technique nous a permis de créer notre corpus nous même et de ne pas télécharger un corpus déjà préparé.

Ce projet nous a posé quelques interrogations et nous allons tenter d' y répondre. Quelles sont les principales actions prévues pour l'écologie en France ? Quelle est l'évolution dans le temps des principaux thèmes problématiques dans la lutte climatique ?
Parmi nos hypothèses, nous en avons émis plusieurs dont trois principales :

1) Le fait de poster des tweets sur les réseaux permet au gouvernement de s'adresser à un public jeune. Par exemple, avec l' utilisation d'emojis ou de hashtags pour viser un public précis.

2) On s'attend à l'utilisation d'un vocabulaire impactant pour inciter les followers à être actifs dans la lutte contre le réchauffement climatique.

3) On s'attend à retrouver des thèmes sur divers moyens de mobilisation tels que le tri sélectif, avoir moins de déchets non recyclables, avoir moins de pollution etc…

B. Construction du corpus

Il est possible de faire du Web Scraping avec Python grâce à la librairie *snscrape* spécialisée dans les scrapings de réseaux sociaux. C'est donc avec cette méthode que nous sommes allées récupérer les tweets qui nous intéressent.

Souhaitant travailler sur la question du climat, des urgences écologiques, nous nous sommes tournées vers le compte ONU Climat, qui est le compte officiel de l'ONU climat français. Nous avons récupéré l'ensemble de tous les tweets publiés sur ce compte pour avoir une analyse complète dessus. La recherche avancée a été effectuée depuis le compte @CCNUCC, du 1er janvier 2015 au 1er janvier 2023. Ce qui nous a donné la requête suivante sur Python : "(from:CCNUCC) until:2023-01-01 since:2015-01-01".

Suite à la récupération de nos tweets avec *snscrape*, nous pouvons observer de quoi est constitué le corpus de tweets. Il y a 9759 tweets allant du 23 janvier 2015 au 21 décembre 2022.

La librairie *pandas*, nous permet d'afficher nos tweets dans un tableau. Ce tableau s'appelle DataFrame, c'est ce dernier que nous allons convertir en document csv.
Le code permettant de venir enregistrer le corpus de tweets est le suivant : *df.to_csv('tweets_ONUClimat_Lise_BRISSET.csv')*.

Pour la suite de l'étude nous n'avons gardé que les tweets français, ce qui nous fait un total de 9630 tweets.

C. Pre-processing

Avant toute manipulation et transformation de notre corpus, nous avons fait l'analyse de l'utilisation des hashtags (voir section II.B.).

C'est à la suite de cette section que nous avons choisi de faire le pre-processing de notre corpus. C'est-à-dire que nous sommes allées nettoyer ce qui n'était plus pertinent à analyser par la suite. Nous avons fait les choix suivants :

Éléments à retirer	Raisons
Les liens	Ce n'est pas du texte à proprement parler, il pourrait nous donner des informations intéressantes sur quel genre de site l'ONU Climat renvoie ses followers, mais ce ne serait que trop peu présent.
Tous les hashtags et les tags vidéos/audio	Difficile de faire de l'analyse textuelle sur ce genre de données. L'analyse des hashtags a été faite plus tôt.
Les retweets et les références à d'autres comptes	Les retweets ne viennent pas directement du compte ONU Climat et les références à d'autres comptes n'apportent pas beaucoup d'éléments de réponse.
Diverses ponctuations, espaces en trop et les caractères de retour à la ligne	Les ponctuations sont des éléments qui sont trop fréquents et viendrait nuire à la qualité des listes des tokens les plus fréquents du corpus.

Cependant certains éléments restent intéressants à garder dans le cas de notre recherche. En effet, les chiffres en font partie car le climat étant un sujet en perpétuelle évolution et sans cesse mesuré ; les chiffres sont des indicateurs alarmants ainsi que des dates clefs qui pourraient régulièrement revenir.

Il y a aussi les emojis, l'une de nos hypothèses portant sur les jeunes et la manière dont ONU Climat cherche à sensibiliser les jeunes et s'adresser à eux, il est donc utile de garder les emojis dans nos futures analyses.

Nous allons aussi venir retirer les stopwords, c'est-à-dire les mots les plus fréquents en français n'apportant pas d'informations, ex : "le, se, à, etc".
Nous finirons par la tokenisation et lemmatisation du corpus afin de venir segmenter le corpus et analyser chaque mot de manière séparée lorsqu'il est utile de le faire.

Ce nettoyage, pre-processing et segmentation ont été rendus possibles grâce à l'utilisation des librairies *nltk* et *spacy*. Voici pour exemple le premier tweet de notre corpus qui n'a pas été nettoyé mais seulement segmenté :

```
print(lemmatizing(tokenize(Text)))
```

```
['#', 'LeSaviezVous', '❓', 'à', 'le', '#', 'COP15', ',', 'le', 'gouverneme
nt', 'se', 'être', 'engager', 'à', 'protéger', '30', 'pourcent', 'de', 'ter
re', 'et', 'un', 'eau', 'considérer', 'comme', 'important', 'pour', 'le',
'#', 'biodiversité', 'de', 'ici', 'à', '2030', '.', 'que', 'être', '-ce',
'que', 'cela', 'vouloir', 'dire', '?', 'voir', 'par', 'vous', 'même', ':',
'https://t.co/k57sprz8hy', 'https://t.co/agrdh9ct9n']
```

Et le résultat après nettoyage des cents premiers tokens de notre corpus :

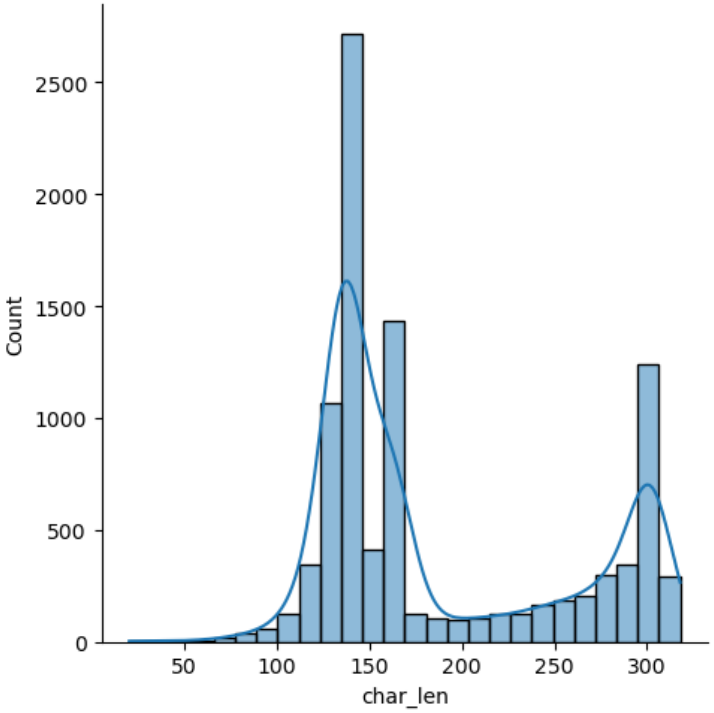
```
list_tweets_clean[0][0:100]
```

```
58]: ['❓',
      'gouvernement',
      'engager',
      'protéger',
      'terre',
      'eau',
      'considérer',
      'comme',
      'important',
      'é',
      'de',
      'ici',
      'que',
      'estce',
      'cela',
      'vouloir',
      'dire',
      'voir']
```

II. Analyse

A. Provenance et répartition des tweets

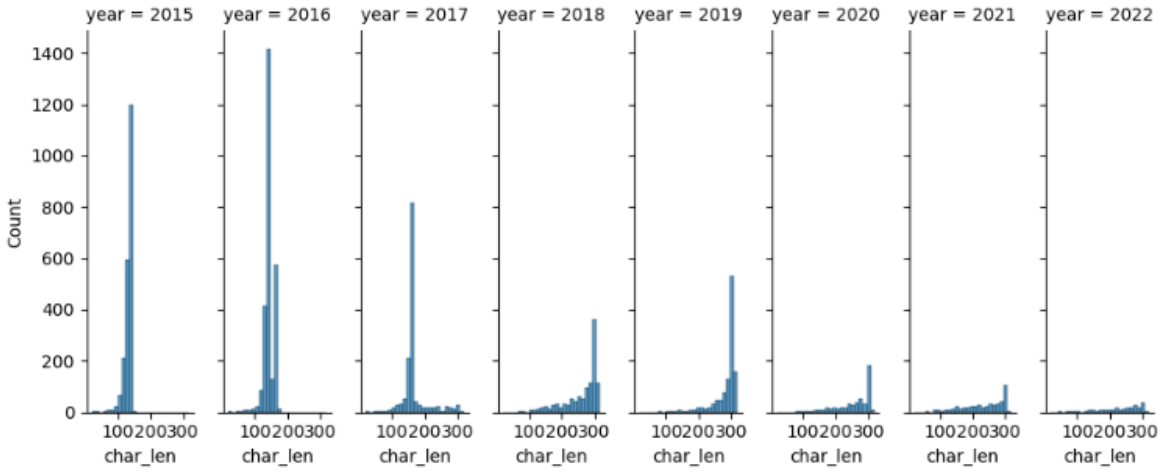
Pour faire l'analyse des tweets, nous avons découpé le corpus en plusieurs corpus de tweets répartis en fonction des mois. Ceci nous a donné un total de 93 périodes, donc 93 mois entre 2015 et 2022.



Le graphique ci-contre représente le nombre de caractères par tweet :

Nous observons un pic considérable entre 100 et 150 caractères pour plus de 2500 tweets postés. Un deuxième pic moins fort que le premier à environ 160 caractères pour à peu près 1500 tweets postés et un troisième pic à 300 caractères pour environ 1400 tweets postés. Cela signifie que le compte twitter de ONUClimat France poste généralement des tweets d'une longueur moyenne voire courte.

La poursuite de notre analyse, nous a conduit à obtenir le graphique suivant. Il représente le nombre de caractères par tweets sur la période de 2015 à 2022.

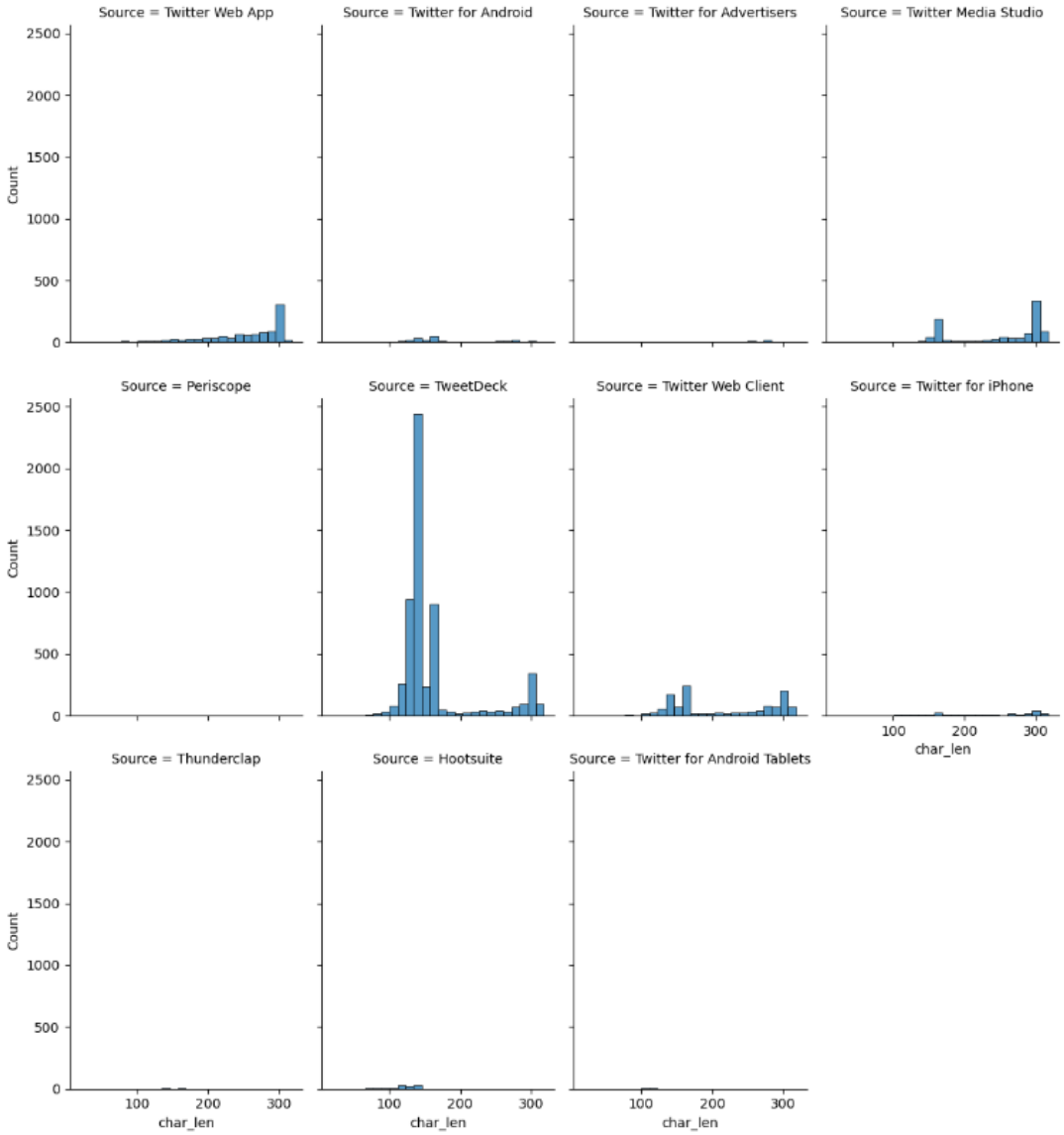


Nous constatons que les premières années, c'est-à-dire 2015 à 2017, sont les années où le compte ONU Climat France a le plus tweeter et que les dernières années 2020 à 2022, il y a peu de tweets. L'année 2016 est l'année où il y a eu le plus de tweets postés environ 1400 et que la longueur en nombre de caractères était d'environ 150. Nous nous

sommes posées la question : pourquoi il y a un pic de tweets cette année-là. Après des recherches, nous avons découvert que l'année 2016 est l'année des records au niveau des températures climatiques, voir le lien ci-dessous du site "Organisation Météorologique Mondiale" (OMM).

Disponible en ligne :
https://public.wmo.int/fr/medias/communiqu%C3%A9s-de-presse/nombreux-records-pulv%C3%A9ris%C3%A9s-en-2016-dans-le-domaine-du-climat-avec-des#:~:text=Selon%20l'Organisation%20m%C3%A9t%C3%A9orologique%20mondiale,et%20du%20r%C3%A9chauffement%20des%20oc%C3%A9ans. (Consulté le 3 mai 2023)

Le graphique suivant représente les différentes sources d'où les tweets ont été envoyés.



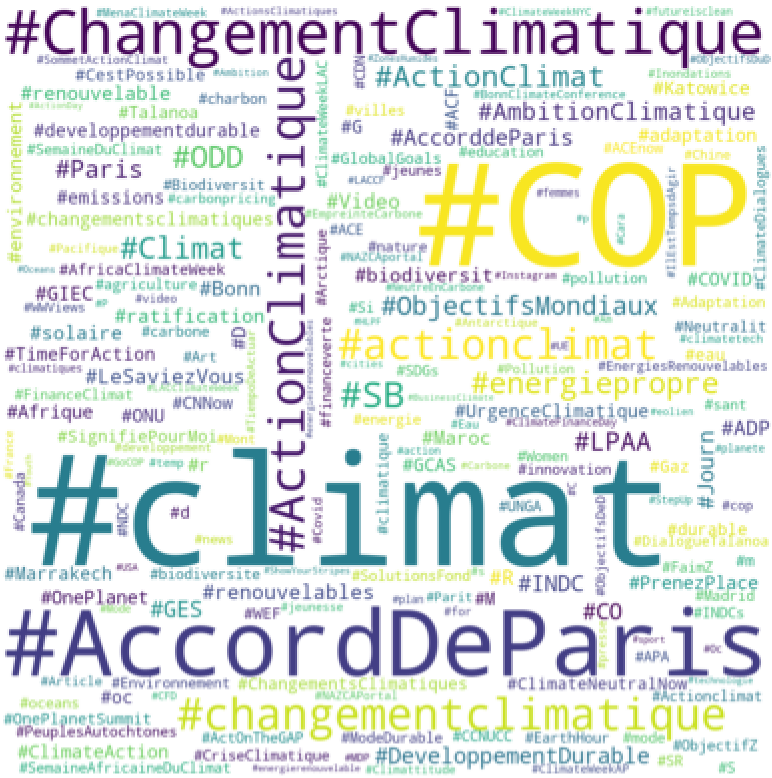
La source TweetDeck est celle qui a envoyé le plus de tweets entre 2015 à 2022. Nous ne connaissons pas TweetDeck, alors nous avons fait des recherches et nous avons trouvé que c'est un logiciel permettant de gérer plusieurs comptes à la fois, nous pensons donc que c'est quelqu'un ou une équipe de communication qui gère le compte principalement.

Voir le lien ci-dessous, disponible en ligne :
https://help.twitter.com/fr/using-twitter/how-to-use-tweetdeck (Consulté le 3 mai 2023).

B. Utilisation des hashtags

Les hashtags sont des éléments très présents au sein de Twitter. Il est donc normal de faire une partie de l'analyse à ce sujet.

Le graphique représenté ci-dessous est un nuage de mots. Plus le mot est grand, plus il est fréquent et plus le mot est petit, moins il le sera.



Les principaux hashtags de ce nuage sont : "#climat", "#COP", "#AccordDeParis", "#ChangementClimatique". Ce nuage soulève les actions envisagées et les thèmes qui reviennent le plus souvent.

Nous pouvons noter des éléments intéressants qui soulignent le souhait de se mobiliser et trouver des solutions au désastre climatique, tels que "#renouvelable", "#DeveloppementDurable", "#PrenezPlace" "#AmbitionClimatique". Ces derniers montrent un souhait d'arriver à des objectifs et d'inciter les français à se mobiliser écologiquement.

Par ailleurs, nous trouvons des éléments soulignant l'urgence écologique tels que "#CriseClimatique", "#UrgenceClimatique". Mais ces derniers ont l'air moins présents dans les hashtags. Ce qui montre que ONU Climat cherche principalement à motiver ses followers avec des hashtags plutôt encourageants.

C. Fréquence et apparition des termes

Dans cette partie, nous allons nous pencher sur la présence plus ou moins constante des mots les plus fréquents de l'ensemble de notre corpus durant la période 2015-2022.

Cette capture d'écran, nous montre les tokens les plus fréquents avant le pre-processing. Cette liste de mots appartient à celle de la tokenisation.

```
[('#', 24799), (':', 18799), ('https', 13021), ('de', 10256), ('@', 8517), ('la', 6039), ('à', 500
4), ('.', 4547), (',', 4466), ('le', 4372), ('climat', 4280), ('les', 3493), ('des', 3204), ('pour',
2855), ('en', 2781), ('l', 2777), ("'", 2771), ('du', 2666), ('et', 2468), ('http', 2367), ('sur', 2
203), ('!', 2082), (';', 1996), ('&', 1979), ('COP21', 1707), ('amp', 1669), ('AccordDeParis', 163
8), ('au', 1467), ('ChangementClimatique', 1347), ('’', 1342), ('pr', 1246), ('un', 1134), ('Le', 11
06), ('est', 1090), ('La', 1027), ('COP22', 967), ('Les', 909), ('dans', 896), ('que', 855), ('une',
850), ('?', 816), ('ActionClimatique', 794), ('a', 739), ('d', 730), ('ONU_fr', 711), ('CCNUCC', 67
2), ('nous', 665), ('par', 642), ('changementclimatique', 640), ("'", 633)]
```

Les résultats désignent l'utilisation de liens "https" à hauteur de 13021 références, de d'autres comptes comme celui des Nations Unies en France : "ONU_fr" soit 711 références, de beaucoup de hashtags soit 24799 références.

Cette capture d'écran a été prise après le pre-processing.

```
Top 100 des tokens les plus fréquents du compte ONUClimat France : --- [('le', 2958), ("'", 2666), ('de', 1955), ('amp', 162
4), ("'", 1231), ('pour', 1118), ('👉', 949), ('avoir', 930), ('#', 829), ('climatique', 760), ('mondial', 721), ('ce', 65
1), ('être', 641), ('plus', 635), ('', 601), ('faire', 598), ('action', 592), ('tout', 581), ('pays', 551), ('pouvoir', 54
1), ('climat', 480), ('conférence', 447), ('contre', 442), ('se', 440), ('devoir', 424), ('ici', 408), ('émission', 396),
('«', 390), ('»', 388), ('pourcent', 381), ('nouveau', 379), ('plan', 370), ('via', 365), ('rapport', 359), ('jour', 351),
('objectif', 345), ('degré', 336), ('changement', 326), ('mettre', 324), ('semaine', 321), ('vidéo', 316), ('atteindre', 30
4), ("aujourd'hui", 297), ('1', 294), ('suivre', 283), ('un', 281), ('🌎', 266), ('direct', 266), ('année', 265), ('monde',
264), ('➕', 261), ('lors', 244), ('présenter', 240), ('énergie', 238), ('pari', 237), ('comment', 236), ('voir', 231), ('d
s', 220), ('face', 216), ('participer', 216), ('ville', 213), ('selon', 210), ('site', 210), ('aussi', 208), ('ne', 205),
('premier', 204), ('prix', 203), ('agir', 202), ('é', 201), ('prendre', 201), ('dernier', 200), ('découvrir', 199), ('jeun
e', 199), ('si', 198), ('2', 198), ('leur', 198), ('solution', 192), ('effet', 191), ('🌍', 189), ('entrer', 188), ('accor
d', 186), ('projet', 184), ('entreprise', 184), ('grand', 182), ('15', 181), ('an', 181), ('que', 179), ('niveau', 179), ('d
éveloppement', 178), ('programme', 177), ('planète', 176), ('lutter', 176), ('avant', 174), ('savoir', 169), ('conf', 168),
('pr', 168), ('carbone', 166), ('lutte', 165), ('engagement', 165), ('presse', 165)]
```

Ceci complémente le nuage de mot, vu précédemment. Cette capture d'écran représente les 100 tokens les plus fréquents du compte ONU Climat France. Nous avons beaucoup de mots qui nous impactent comme "atteindre", "énergie", "participer", "découvrir", "mondial", "jeune", "pari", "entreprise", "engagement", "vert", "réduire", "action", "planète", "lutte".

Ici, les stopwords ont été enlevés et nous retrouvons les 10 premières listes de la collection cleaning, c'est-à-dire qu'il n'y a plus de mots vides, de caractères indésirables… Cependant nous remarquons quelques anomalies.

```
['❓', 'gouvernement', 'engager', 'protéger', '30', 'pourcent', 'terre', 'eau', 'considérer', 'comme', 'important', 'é', 'd
e', 'ici', '2030', 'que', 'estce', 'cela', 'vouloir', 'dire', 'voir']
['❌', '22', 'million', 'personne', 'déplacer', 'faire', 'de', 'évènement', 'météorologique', 'extrême', 'chaque', 'année',
'❌', 'sans', 'action', 'urgent', 'nombre', 'personne', 'pouvoir', 'atteindre', '216', 'million', 'de', 'ici', '2050', 'èco
utez', 'entretien', 'exclusif', 'savoir', '➕', '', '👇']
['découvrir', '🔟', 'programme', 'mond', 'entier', '🌍', 'récompenser', "'", 'rôle', 'restauration', 'monde', 'naturel']
['❓', 'perte', 'gaspillag', 'alimentaire', 'représenter', 'entrer', '8⃣', 'pourcent', '🔟', 'pourcent', '#', 'émission',
'gaz', 'effet', 'serre', 'chaque', 'année', 'comment', '❓', 'dire', '➕', '👇']
['«', 'ne', 'y', 'avoir', 'planète', 'b', 'ce', 'être', 'réparer', 'monde', '»']
['se', 'agir', 'transformer', 'relation', 'société', 'é', 'permettre', 'vie', 'harmonie', 'nature', 'panèt', '🌍']
['triple', 'crise', 'planétaire', '⚠', 'changement', 'climatique', '⚠', 'pollution', '⚠', 'perte', 'biodiversité', 'alor
s', 'avoir', 'démarrer', 'rappelonsnous', 'risque', 'de', 'extinction', 'espèce', 'augmenter', 'chaque', 'fraction', 'degr
é', 'réchauffement']
['changement', 'climatique', 'venir', 'terminer', 'charm', 'elcheikh', 'égypt', 'ɛ', 'ɢ', 'é', 'se', 'ouvrir', 'montréal',
'canada', 'c', 'ᴀ', 'pourquoi', 'deux', 'cop', 'pourquoi', '15', '27']
['secteur', 'agricole', '➕', '➕', 'vulnérable', 'changement', 'climatique', 'avoir', 'developper', '"', 'fast', '"', 'pro
gramme', 'multipartite', 'viser', '💰', 'renforcer', 'financement', 'transformation', 'le', 'agriculture', '🦾', 'contribue
r', 'effort', "'", '👉']
```

Nous voyons de nombreux emojis ce qui rejoint notre hypothèse qu'ils peuvent être destinés à un public visé jeune. Nous observons qu'il y a beaucoup de mots qui font sens avec le

climat, nous avons “climatique”, “crise”, “météorologique”, “gaz”, “effet”, “serre”, ou d'éventuelles actions à mener comme “pollution”, “gaspillage”, “programme” et “protéger”.

Dans cette partie, TF-IDF est un indicateur qui suppose que les caractéristiques les plus discriminantes sont celles qui apparaissent fréquemment dans le document actuel et rarement dans d'autres documents, d’après le cours *Text Mining* de Serrar Loubna . Cet indicateur se calcule ainsi :

- TF : Nombre d’occurrence du terme analysé / Nombre de termes total.
- IDF : Log(Nombre total de documents / Nombre de documents contenant le terme analysé
- TF-IDF = TF * IDF

Ci-dessous une capture d’écran du terme le plus significatif dans chacun des documents par rapport aux autres documents par mois.

```
Le terme le plus signifiant de  Dec-2022  est  : --- chaque
Le terme le plus signifiant de  Nov-2022  est  : --- charm
Le terme le plus signifiant de  Oct-2022  est  : --- climatique
Le terme le plus signifiant de  Sep-2022  est  : --- climatique
Le terme le plus signifiant de  Aug-2022  est  : --- gabon
Le terme le plus signifiant de  Jul-2022  est  : --- hybride
Le terme le plus signifiant de  Jun-2022  est  : --- 2022
Le terme le plus signifiant de  May-2022  est  : --- forêt
Le terme le plus signifiant de  Apr-2022  est  : --- climatique
Le terme le plus signifiant de  Mar-2022  est  : --- climatique
Le terme le plus signifiant de  Feb-2022  est  : --- climatique
Le terme le plus signifiant de  Jan-2022  est  : --- 2022
Le terme le plus signifiant de  Dec-2021  est  : --- côté
Le terme le plus signifiant de  Nov-2021  est  : --- glasgow
Le terme le plus signifiant de  Oct-2021  est  : --- climatique
Le terme le plus signifiant de  Sep-2021  est  : --- climatique
Le terme le plus signifiant de  Aug-2021  est  : --- éemondialehumanitair
Le terme le plus signifiant de  Jul-2021  est  : --- épassement
Le terme le plus signifiant de  Jun-2021  est  : --- climatique
Le terme le plus signifiant de  May-2021  est  : --- mai
Le terme le plus signifiant de  Apr-2021  est  : --- climatique
Le terme le plus signifiant de  Mar-2021  est  : --- dire
Le terme le plus signifiant de  Feb-2021  est  : --- pourcent
Le terme le plus signifiant de  Jan-2021  est  : --- 2021
Le terme le plus signifiant de  Dec-2020  est  : --- tout
Le terme le plus signifiant de  Nov-2020  est  : --- éro
Le terme le plus signifiant de  Oct-2020  est  : --- wébinaire
Le terme le plus signifiant de  Sep-2020  est  : --- idée
Le terme le plus signifiant de  Aug-2020  est  : --- refroidissement
Le terme le plus signifiant de  Jul-2020  est  : --- reprise
```

```
Le terme le plus signifiant de  Jul-2018  est  : --- émission
Le terme le plus signifiant de  Jun-2018  est  : --- 0emissions
Le terme le plus signifiant de  Feb-2018  est  : --- récif
Le terme le plus signifiant de  Jan-2018  est  : --- 2017
Le terme le plus signifiant de  Dec-2017  est  : --- summit
Le terme le plus signifiant de  Nov-2017  est  : --- climatique
Le terme le plus signifiant de  Oct-2017  est  : --- érature
Le terme le plus signifiant de  Sep-2017  est  : --- ratification
Le terme le plus signifiant de  Aug-2017  est  : --- participer
Le terme le plus signifiant de  Jul-2017  est  : --- 2100
Le terme le plus signifiant de  Jun-2017  est  : --- participer
```

Nous observons que le terme “climatique” est prédominant depuis 2017.
Nous n’avons pas pris les périodes 2015-2016 et 2019 car les tokens “pr” et “amp” revenaient souvent et ce n’était pas pertinent pour notre étude.

Dans le tableau ci-contre, la colonne TF Keywords indique les occurrences du terme “action” et la colonne TF-IDF Keywords représente les occurrences du terme “action” sur l’ensemble des tweets.

	period	TF Keywords	TFIDF Keywords
0	Dec-2022	[chaque, climatique, avoir, changement, pource...	[chaque, harmonie, 216, évènement, panèt, gasp...
1	Nov-2022	[charm, elcheikh, direct, plus, climatique, ch...	[charm, elcheikh, direct, préjudice, membre, p...
2	Oct-2022	[climatique, rapport, climat, propre, nouveau,...	[climatique, envers, perte, propre, fréquent, ...
3	Sep-2022	[climatique, être, climat, pays, mondial, avoi...	[climatique, éedelairpur, contenu, être, essay...
4	Aug-2022	[climat, climatique, africain, face, semaine, ...	[gabon, librevill, climat, continent, africain...
...	...	...	...
88	May-2015	[climatique, climat, changement, amp, nouveau,...	[summit, 20minute, climatique, livre, amendeme...
89	Apr-2015	[climatique, changement, mesure, africain, éne...	[kilani, africain, 2015, climatique, mesure, f...
90	Mar-2015	[climatique, amp, réchauffement, aujourd, hui,...	[climatique, météo, énergiepropr, mars, réchau...
91	Feb-2015	[climatique, genève, changement, conférence, p...	[genève, mt, climatique, 2015, hier, accord, t...
92	Jan-2015	[mt, 2015, négociation, disponible, suivre, re...	[mt, négociation, 2015, réseal, disponible, re...

Nous pouvons souligner que le terme “action” est toujours lié au terme “climatique”, “climat”, “changement” entre le début de la période 2015 et la fin de la période 2022.

La capture d'écran suivante génère une liste des tokens les plus pertinents avec le terme recherché "action" en utilisant l'indicateur des scores TF-IDF.

```
# Définir le mot recherché
query = "action"
# trié la colonne correspondant à ce terme dans la matrice des scores TF-IDF
index = tf_idf_vector.sort_values(by=[query], ascending=False).index.to_list()
#Utiliser les index pour trouver le corpus le plus pertinent
for i in range(10):
    print("mois #", index[i], ":",
          df_results['period'][index[i]],"____ :",
          df_results['TFIDF Keywords'][index[i]])
```

```
mois # 84 : Sep-2015 ____ : ['plan', 'action', 'présenter', 'choisir', 'pays', 'reporter', 'aideznous', 'présent', 'accord',
'voter']
mois # 83 : Oct-2015 ____ : ['action', 'plan', 'présenter', 'amp', 'site', 'initiative', 'pays', 'accord', 'amont', 'pari']
mois # 82 : Nov-2015 ____ : ['pari', '2015', 'accord', 'site', 'action', 'plan', 'présente', 'initiative', 'présent', 'amp']
mois # 85 : Aug-2015 ____ : ['envoyer', 'participer', 'plan', 'rejoindre', 'amp', 'jour', 'aideznous', 'choisir', 'détail',
'action']
mois # 86 : Jul-2015 ____ : ['amp', 'chanson', 'aideznous', 'scientifique', 'améliorer', 'site', 'action', 'présenter', 'sem
aine', 'sommet']
mois # 20 : Apr-2021 ____ : ['climatique', 'éedelaterre', 'climat', 'changement', 'éedelaterr', 'planète', 'québec', 'actio
n', 'prix', 'mondial']
mois # 87 : Jun-2015 ____ : ['amp', 'dialogue', 'vidéo', 'citoyen', 'rome', 'encycliqu', 'accord', 'action', 'mondial', 'mor
al']
mois # 11 : Jan-2022 ____ : ['2022', 'climat', 'tout', 'charbon', 'baser', 'climatique', 'ski', 'dumoyenorient', 'athlèt',
'éeinternationaledeléducation']
mois # 29 : Jul-2020 ____ : ['reprise', 'contre', 'juillet', 'énergie', 'éro', 'action', 'certain', 'relance', 'émission',
'événement']
mois # 24 : Dec-2020 ____ : ['tout', 'dvpt', 'être', 'humain', 'éedesmigrant', 'sommet', 'climatique', 'vaccin', 'décembre',
'anniversaire']
```

Nous ne retrouvons pas les mots attendus comme, "trier"ou "recycler". Mais nous retrouvons des verbes commes "participer", "rejoindre" et "choisir".
Il y aussi des synonymes intéressants comme "plan", "améliorer", "changement", "citoyen" (droits et devoirs).
Dans la dernière liste, nous distinguons "dvpt" qui signifie développement. C'est une abréviation utilisée dans le langage des jeunes.

La partie N-gram est une approche contextuelle qui va nous permettre de nous pencher sur les mots qui suivent.
Le calcul de n-grams constitue la méthode la plus simple pour tenir compte du contexte. Dans cette partie, ce sont les bigrams qui vont le plus nous intéresser pour les cooccurrences. Une cooccurrence est une combinaison de mots qui apparaissent fréquemment ensemble dans un corpus. Les collocations sont eux une forme privilégiée de cooccurrence : ce sont tout simplement des mots qui tendent à apparaître ensemble, d'après le cours *Text Mining* de Serrar Loubna.

Dans notre notebook, nous avons noté le vocabulaire impactant avec comme mot choisi "action". Lorsque l'on s'intéresse aux concordances, le vocabulaire qui en ressort est "urgent", "crucial", "nouveau", "décisif", "nouveau monde", "transformer", "promesse", "ambitieux" et "nécessité". Pour les collocations (voir ci-dessous), nous constatons que le vocabulaire est marquant de part notre sujet.

```
changement climatique; lutter contre; lutte contre; atteindre
objectif; ban kimoon; gaz effet; mettre œuvre; présenter plan; déposer
instrument; entrée vigueur; haut niveau; ici 2050; faire face; new
york; secrétaire général; ici 2030; action climatique; effet serr;
jamais enregistrer; amendement doha
```

Nous pouvons faire une hypothèse concernant les nombres 2030/2050 qui consistent à prévenir des projets d'actions futures, ceci afin d'atteindre les objectifs fixés au niveau du climat.

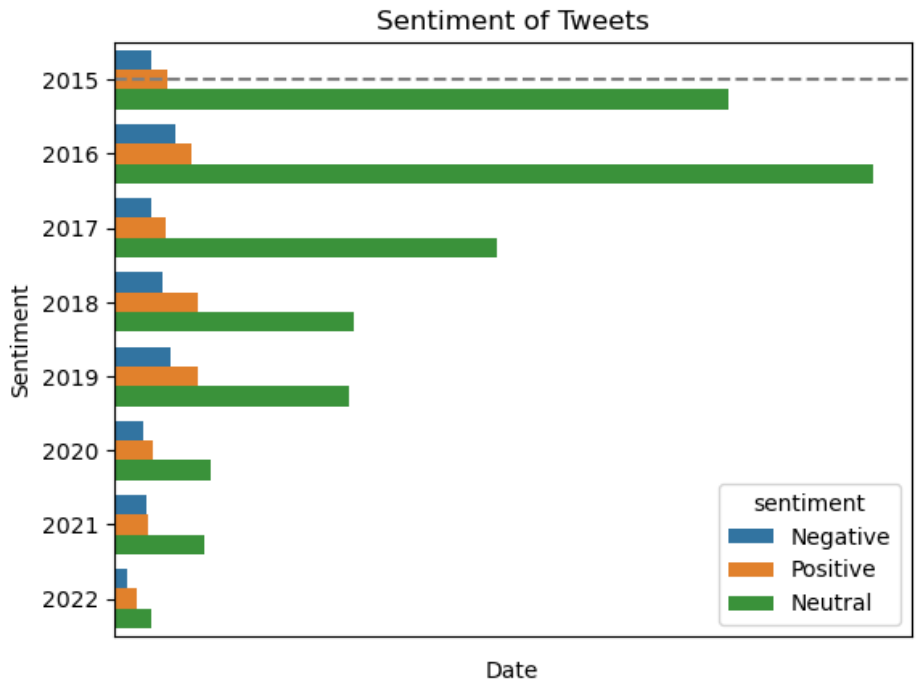
D. Sentiment analysis

Dans cette partie, nous allons tenter d'identifier qui se cache derrière le compte d'ONU Climat. En règle générale, un humain exprime des opinions ou des émotions qu'elles soient positives ou négatives.

sentiment	Date	period	year	Source	Tweet_clean	scores
Negative	946	90	8	8	861	1
Neutral	7353	93	8	11	6720	1
Positive	1328	89	8	9	1228	1

Lorsque nous avons commencé cette partie, nous avons été surpris par les résultats. Dans ce tableau, nous pouvons voir que notre corpus contient une majorité de tweets qui sont neutres. En effet, 6720 d'entre eux sont neutres, alors que 1228 sont positifs et 861 sont négatifs.

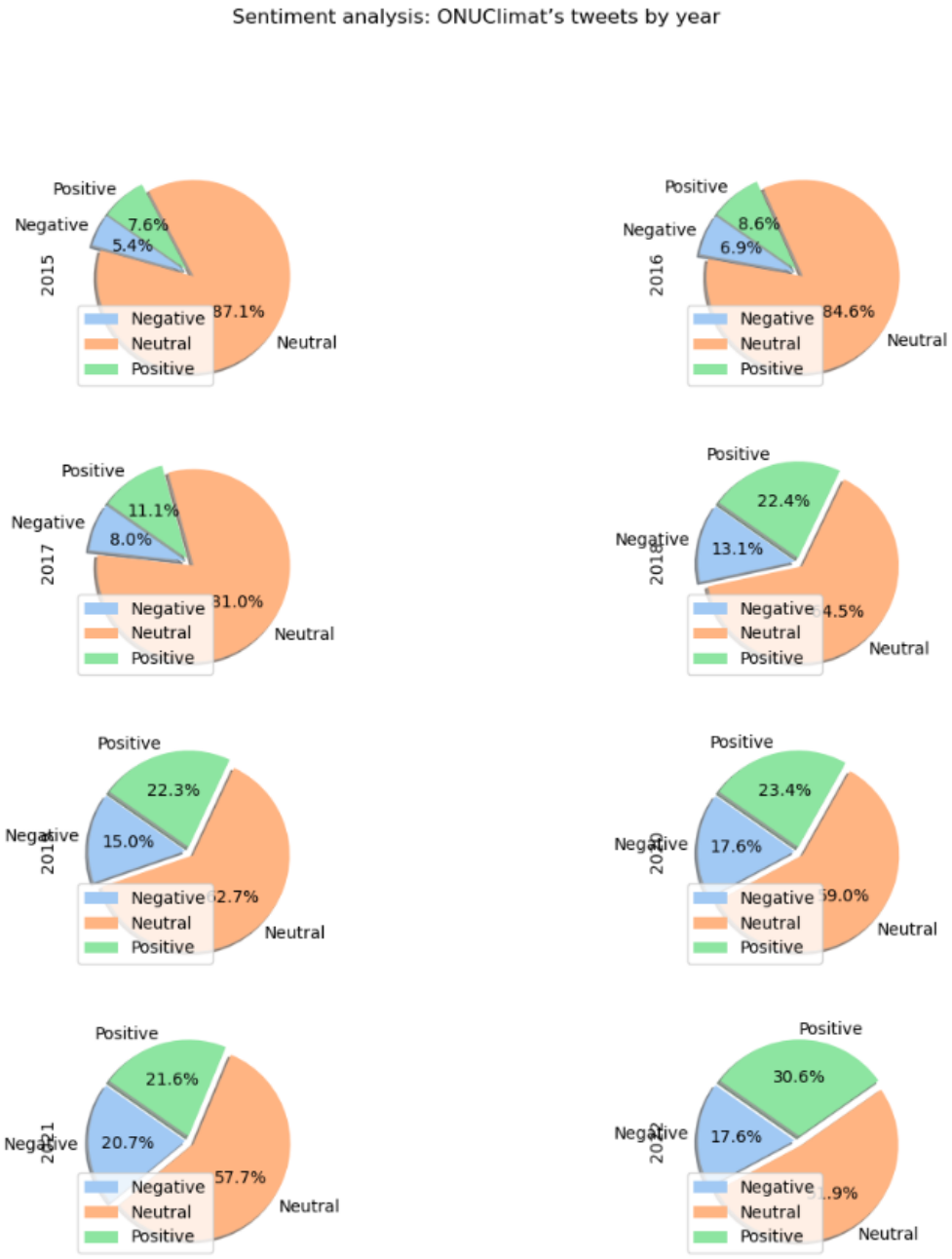
En continuant notre recherche, nous avons trouvé un graphique qui contient le nombre de tweets neutres, positifs et négatifs selon les années. Voici, la capture d'écran ci-dessous.



Nous retrouvons encore cette période de 2015 et 2016 où les tweets étaient nombreux, mais ici on observe que cette période a été exponentielle sur les tweets neutres. Ainsi, nous nous sommes rapidement rendu compte qu'il pourrait y avoir un lien entre ce résultat mais aussi de la provenance des tweets. En amont, nous avons expliqué que la majorité des tweets provenait de TweetDeck. En faisant des recherches, nous avons compris que c'est un outil de gestion de Twitter. Il permet à un utilisateur de gérer plusieurs comptes Twitter, mais il a aussi de nombreuses fonctionnalités qui sont intéressantes pour certains comptes. Par exemple, suivre les hashtags qui sont en TT, programmer des tweets pour une publication future ou encore améliorer l'interaction avec sa communauté. Cet outil est souvent utilisé par des influenceurs ou encore des community managers.

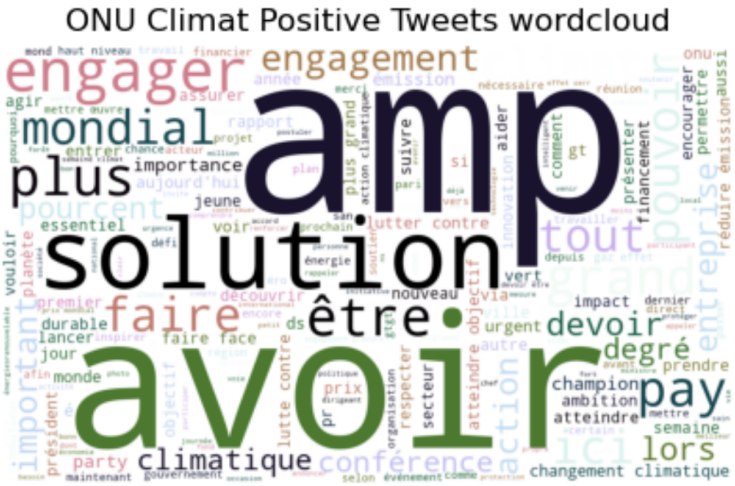
Ainsi, nous pouvons voir que le compte d'ONU Climat est géré par une équipe de communication comme des community managers.

De plus, on observe encore une baisse des tweets à partir de 2017, et plus particulièrement des tweets neutres. Cette baisse significative est accentuée avec la capture ci-dessous.



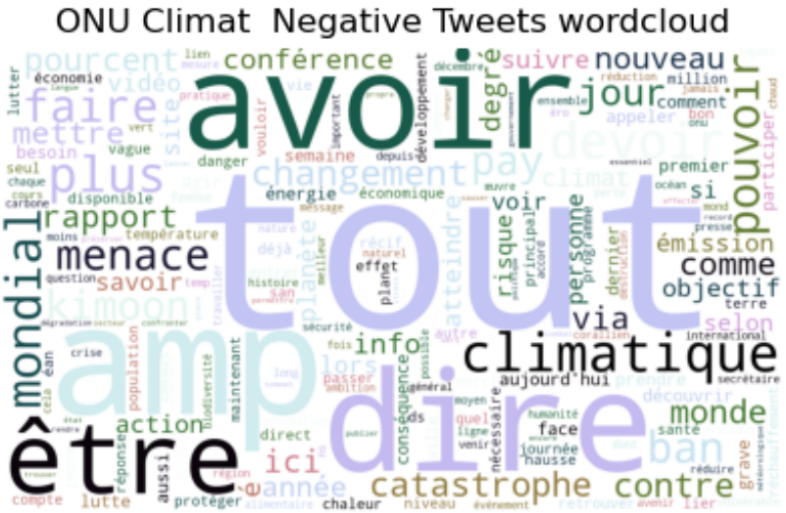
Nous pouvons voir que nous passons d'environ 80% de tweets neutres à 50%, alors que les tweets positifs et négatifs augmentent légèrement de 5 à 7 % jusqu'à environ 17 et 30 %.

De plus, nous avons cherché quels sont les mots les plus utilisés pour les tweets positifs et négatifs.



Nous allons commencer par les mots positifs. Nous pouvons voir que le champ lexical est positif et qui est relié à aux actions de l'ONU. Par exemple, nous observons des verbes tels que "faire", "agir", "atteindre", "encourager" ou "engager". Mais aussi des mots comme "solution", "plan", "défi", "ambition", "jeune" ou "engagement". Nous pouvons voir que ce vocabulaire encourage le public à une action collective pour lutter contre ce phénomène mondial.

Nous allons regarder les mots négatifs. Contrairement aux mots positifs, nous pouvons voir que le champ lexical utilisé est plus frappant. Par exemple, nous pouvons voir "catastrophe", "degré", "émission", "risque", "danger", "menace", "chaleur" ou encore "destruction". Ce sont des mots très percutants pour les followers qui lisent ses tweets. Ainsi, nous pouvons voir que tous les tweets ne sont pas écrits au hasard car les mots utilisés sont intentionnels et réfléchis.



III. Conclusion

Pour conclure, nous pouvons voir que le compte de l'ONU Climat France publie du contenu axé sur le changement climatique et la lutte contre le réchauffement climatique. Pour cela, une équipe de communication publie des tweets sur les décisions prises lors des COP, ou encore des actions et des initiatives prises par les organisations. Créer un compte Twitter est une stratégie pour sensibiliser et viser un public ciblé, les jeunes. Pour cela, nous avons vu que le vocabulaire utilisé est impactant mais aussi compréhensible avec l'utilisation massive de hashtags et d'émojis.

L'objectif étant de venir trouver des éléments pouvant affirmer ou non nos hypothèses, voici les éléments de réponses que nous avons pu regrouper suite à notre recherche et nos analyses.

1) *Le fait de poster des tweets sur les réseaux permet au gouvernement de s'adresser à un public jeune. Par exemple, avec l' utilisation d'emojis ou de hashtags pour viser un public précis*.

En effet, nous avons observé une utilisation importante d'emojis qui pourrait être un bon moyen de s'adresser aux jeunes. Ceci avec quelques abréviations tel que le terme "dvlp".
Concernant les hashtags, nous n'avons pas trouvé des hashtags directement adressés aux jeunes mais plutôt à un public de tout âge en général avec des termes significativement impactant comme "#UrgenceClimatique", "#CriseClimatique" et sollicitant le monde à se mobiliser comme "#PrenezPlace". Mais aussi, ces hashtags permettent d'interagir directement avec la communauté du compte, tels que "#LeSaviezVous?".

Cette première hypothèse n'est donc qu'affirmer en partie, au niveau de l'utilisation des émojis pour s'adresser aux jeunes. ⚠️

2) *On s'attend à l'utilisation d'un vocabulaire impactant pour inciter les followers à être actifs dans la lutte contre le réchauffement climatique.*

Comme sur le dernier point, nous avons trouvé des termes et hashtags impactants. Les termes qui reviennent le plus fréquemment sont par exemple "atteindre", "participer", "engagement", "réduire", "action", "lutte" et "crise". L'analyse des collocations a fait ressortir les nombres 2030 et 2050, nous pensons que cela peut avoir un lien avec de futurs actions mises en place ou à mettre en place contre le réchauffement climatique.

Cette seconde hypothèse est donc affirmée. ✅

3) *On s'attend à retrouver des thèmes sur divers moyens de mobilisation tels que le tri sélectif, avoir moins de déchets non recyclables, avoir moins de pollution etc…*

Concernant les moyens de mobilisation, nous avons trouvé quelques éléments de réponse tels que l'apparition des termes faisant référence à l'utilisation de transports plus écologiques et moins polluants comme le terme "hybride". Il y a aussi des termes faisant référence au tri sélectif comme "trier" et "recycler" venant s'opposer aux termes "pollution" et "gaspillage". Ainsi que beaucoup de mots amenant les followers à être mobilisés dans la lutte contre le réchauffement climatique avec les termes "engager", "participer", "rejoindre" et "protéger".

Cette dernière hypothèse est donc affirmée. ✅