



UFR 6

Université Paul Valéry, Montpellier III

Rapport TER

## Visualisation de tweets controversés

Alya Zouzou, Audric Girondin, Houria Sayah, Jonathan Duckes, Maéva Maïo

Encadrants pédagogiques : Sandra Bringay, Maximilien Servajean

janssen  
Horizon



## Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation :

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature : 

Date : 04/05/2024

Signature : 

Date : 04/05/2024

Signature : 

Date : 04/05/2024

Signature : 

Date : 04/05/2024

Signature : 

Date : 04/05/2024

Nous tenons à vous exprimer notre profonde gratitude pour l'ensemble des enseignements que vous nous avez prodigués tout au long de notre parcours au sein du Master MIASHS. Votre dévouement, votre expertise et votre engagement ont été des piliers essentiels dans notre formation et notre développement académique.

Nous sommes également reconnaissants pour la précieuse opportunité qui nous a été offerte de participer à un projet de recherche tout au long de l'année. Cette expérience enrichissante a grandement contribué à approfondir notre compréhension des enjeux actuels dans notre domaine d'études et à enrichir nos compétences.

Nous tenons également à adresser nos remerciements les plus chaleureux à Sandra Bringay et Maximilien Servajean pour leur accompagnement attentif et leur disponibilité tout au long de ce projet. Leur expertise, leur soutien et leur volonté de répondre à nos questions ont été d'une aide inestimable et ont grandement facilité notre progression.

## Résumé

En tant qu'étudiants en première année de Master Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales (MIASHS), nous sommes chargés d'entreprendre un projet d'étude d'envergure qui s'étendra sur toute la durée de l'année universitaire. Ce projet nous offre l'opportunité d'appliquer nos compétences universitaires et de renforcer notre esprit d'analyse ainsi que nos compétences en collaboration. De plus, il enrichit notre compréhension du domaine de la data science en nous permettant de mettre en pratique des méthodes et des techniques concrètes dans des contextes réels.

## Abstract

As first-year students in the Master's program in Applied Mathematics and Computer Science for the Humanities and Social Sciences, we have been tasked with undertaking a large-scale research project that will span the entire academic year. This project provides us with the opportunity to apply our academic skills and enhance our analytical mindset and collaborative abilities. Furthermore, it deepens our understanding of the field of data science by allowing us to implement concrete methods and techniques in real-world contexts.

# Table des matières

<b>Déclaration de non plagiat</b>	ii
<b>Remerciements</b>	iii
<b>Résumé</b>	iv
<b>Liste des figures</b>	vi
<b>Liste des tables</b>	vii
<b>Introduction</b>	1
<b>Question de recherche</b>	1
<b>1 Contexte de recherche</b>	2
1.1 Question de recherche . . . . .	3
1.2 Contexte . . . . .	3
<b>2 Présentation du projet</b>	5
2.1 Planification et gestion de projet . . . . .	6
2.2 Présentation des données . . . . .	6
2.2.1 Traitement des données . . . . .	6
2.3 Défis . . . . .	8
<b>3 Analyses Descriptives</b>	9
3.1 Visualisation des données . . . . .	10
3.2 Poursuite des analyses . . . . .	11
<b>4 Détection de communautés</b>	12
4.1 Objectif de l'analyse . . . . .	13
4.2 Méthodologie . . . . .	13
4.2.1 Présentation des résultats . . . . .	13
<b>5 Analyse des sentiments et des émotions</b>	15
5.1 Objectifs de l'analyse . . . . .	16

5.2 Méthodologie	16
5.3 Présentation des résultats	17
<b>6 Analyse des opinions</b>	<b>19</b>
6.1 Objectifs de l'analyse	20
6.2 Méthodologie	20
6.3 Présentation des Résultats	21
6.4 Défis et perspectives	22
<b>7 Topic Modelling et Latent Dirichlet Allocation</b>	<b>24</b>
7.1 Objectif de l'analyse	25
7.2 Méthodologie	25
7.3 Présentation des résultats	26
<b>8 Indice de controverse</b>	<b>28</b>
8.1 Objectif de l'analyse	29
8.2 Méthodologie	29
8.3 Présentation des résultats	30
<b>Conclusion</b>	<b>32</b>
<b>Bibliographie</b>	<b>33</b>

# Table des figures

2.1 Tableau Trello.	6
2.2 Vue d'un tweet.	7
2.3 Jeu de données nettoyé et traité.	7
3.1 Nuage de mots des hashtags.	10
3.2 Dashboard descriptif.	11
4.1 Communauté d'Hashtags.	14
5.1 Scores moyens des sentiments.	17
5.2 Scores moyens des émotions.	17
5.3 t-SNE Visualization of Feature Spaces for Sentiment Analysis	18
7.1 Sujets et mots clefs.	26
7.2 Visualisation du Topic Modelling avec le LDA.	27
8.1 Tweet impacté faiblement par les caractéristiques	30
8.2 Tweet impacté fortement par les caractéristiques	30
8.3 Tweet analysé	31
8.4 Les seuils font diminuer l'indice	31
8.5 Les seuils font augmenter l'indice	32
8.6 Un des tweets comportant un indice de controverse maximal	32

# Liste des tableaux

6.1	Détails de la phase d'entraînement . . . . .	21
6.2	Détails de l'entraînement du modèle . . . . .	21
6.3	Accuracy du modèle par classe et globale . . . . .	21

# Introduction

Les opinions divergent sur les réseaux sociaux, en particulier sur Twitter. Les utilisateurs exposent ouvertement leurs points de vue, débattent et partagent activement des informations, des analyses et des réflexions sur tous les sujets.

Ces débats argumentés autour d'opinions divergentes sont caractérisés par la controverse. De nombreux sujets de santé génèrent également des discussions et des confrontations d'idées sur la plateforme.

Dans ce contexte, notre projet d'étude, financé par la fondation Janssen en lien avec le projet Controverse et mené au LIRMM, a pour objectif l'analyse de tweets controversés autour de la santé. Cette initiative s'inscrit dans une démarche visant à comprendre et quantifier les niveaux de controverse autour de sujets sensibles tels que la Covid-19 et les interventions non médicamenteuses, sur Twitter.

De manière générale, notre étude s'articule autour de cinq missions principales : la collecte et l'exploration des données, la création d'un tableau de bord interactif, l'analyse et la visualisation des thèmes, la classification des tweets "chauds" à l'aide de méthodes avancées d'apprentissage automatique et la définition d'un score de controverse. Nous avons choisis de restreindre notre analyse autour de la question des Interventions non Médicamenteuses (INM) pour la guérison du cancer.

Ce rapport retrace notre parcours en tant qu'équipe d'étudiants engagés dans un projet de recherche, offrant une présentation approfondie de notre sujet d'étude et des activités que nous avons menées. Il offre une plongée complète dans notre exploration des tweets controversés, mettant en avant les missions réalisées, les résultats obtenus, et les défis surmontés au cours du processus.

# Chapitre 1

## Contexte de recherche

### Sommaire

---

<b>1.1 Question de recherche . . . . .</b>	<b>3</b>
<b>1.2 Contexte . . . . .</b>	<b>3</b>

---

## 1.1 Question de recherche

En quoi les visualisations de données peuvent-elles contribuer à une meilleure compréhension des facteurs de controverses sur Twitter, notamment en ce qui concerne les débats et les partages d'opinions divergentes ? Et, comment peuvent-elles être utilisées pour améliorer la classification des tweets grâce à l'analyse du texte et des thèmes ?

Dans la poursuite du projet, nous nous sommes particulièrement concentrés sur l'analyse des interventions non médicamenteuses (INM), notamment dans le contexte de la lutte contre le cancer, en utilisant les insights tirés des visualisations de données pour affiner nos méthodes d'analyse et pour approfondir notre compréhension des perceptions et des attitudes des utilisateurs sur ce sujet spécifique.

## 1.2 Contexte

Le cancer est une maladie très complexe qui a suscité énormément de débats au sein de la communauté médicale, et aucun traitement efficace à 100 % n'a encore été découvert. À l'heure actuelle, le traitement principal pour guérir le cancer est la chimiothérapie, qui est un traitement très invasif et peu apprécié en raison de ses nombreux effets secondaires négatifs, comme la perte de cheveux, la baisse de moral, etc. Les principales méthodes non médicamenteuses présentes dans nos données sont l'utilisation du cannabis, l'activité sportive ou encore le jeûne. Les bienfaits d'une activité sportive quotidienne et de la pratique du jeûne occasionnel sont très connus. Pour autant, les bienfaits de certaines substances addictives comme la marijuana ou le cannabis restent encore très controversés. En effet, ces substances sont illégales dans la plupart des pays du monde entier à cause de leur conséquences sur la santé individuelle et la prise de décision. La consommation de cannabis peut entraîner une variété d'effets psychologiques, tels que l'euphorie, la relaxation et des modifications de la perception sensorielle. Certains bienfaits de ces substances ont, en effet, été mis en avant ces dernières années. Par exemple, il y a eu la mise en place de cannabis thérapeutique, afin d'aider les patients en fin de vie à apaiser leur douleur. [1] L'expérimentation de l'usage du cannabis à des fins thérapeutiques a d'ailleurs débuté en France, le 26 Mars 2021 et s'est achevé le 26 mars 2024 [2]. Toutefois, ces substances peuvent également provoquer des symptômes moins désirables comme la paranoïa, l'anxiété, ainsi que des difficultés de concentration et de mémorisation à court terme. Sur le plan physique, les consommateurs peuvent expérimenter des yeux rouges, une bouche sèche, une augmentation de l'appétit et une diminution de la coordination motrice. À long terme, l'usage régulier de cannabis est associé à des risques pour la santé mentale, notamment un risque accru de développer des troubles comme la schizophrénie ou la dépression, particulièrement chez les jeunes. De plus, il peut conduire à une dépendance et à un syndrome amotivationnel, se manifestant par une réduction de la motivation et une altération du fonctionnement quotidien. Ces effets combinés soulignent la complexité des

impacts du cannabis sur la santé individuelle qui explique les divergences au sein de la communauté médicale.

D'autres recherches ont mis en évidence que l'utilisation de ces substances pourrait contribuer à la lutte contre le cancer, notamment grâce à des études réalisées sur des cellules et des animaux. Le National Institutes of Health (NIH) affirme que les cannabinoïdes, présents dans le cannabis, ont démontré des effets anti-tumoraux en laboratoire [3]. Ces études, menées sur des cellules de souris et de rats, montrent que les cannabinoïdes peuvent inhiber la croissance des tumeurs en provoquant la mort des cellules cancéreuses, en bloquant leur division et en empêchant l'angiogenèse, c'est-à-dire la formation de nouveaux vaisseaux sanguins qui alimentent la tumeur, tout en préservant les cellules saines. Cependant, ce sujet reste largement débattu et ne fait pas consensus. Nous allons donc, dans cette étude, tenter de saisir comment la controverse entourant ces méthodes non médicamenteuses pour le traitement du cancer.

# Chapitre 2

## Présentation du projet

### Sommaire

---

<b>2.1 Planification et gestion de projet . . . . .</b>	<b>6</b>
<b>2.2 Présentation des données . . . . .</b>	<b>6</b>
<b>2.2.1 Traitement des données . . . . .</b>	<b>6</b>
<b>2.3 Défis . . . . .</b>	<b>8</b>

---

## 2.1 Planification et gestion de projet

Dans le cadre de notre projet, Alya Zouzou a été élue Team Leader. Son rôle a été de coordonner les efforts de toute l'équipe, en veillant à ce que chacun trouve sa place selon ses compétences, ses forces et ses intérêts. Nous avons adopté une approche Agile, travaillant par sprint d'une semaine, avec des réunions régulières avec nos commanditaires, Sandra Bringay et Maximilien Servajean, deux fois par mois pour présenter nos avancées et ajuster notre direction si nécessaire.

Pour faciliter notre gestion de projet, nous avons choisi Trello [4] comme outil principal. Trello nous a permis de visualiser et de suivre nos tâches de manière intuitive. Nous avons également opté pour GitHub [5] comme plateforme de travail collaboratif, permettant ainsi à tous les membres de l'équipe de contribuer efficacement au projet.

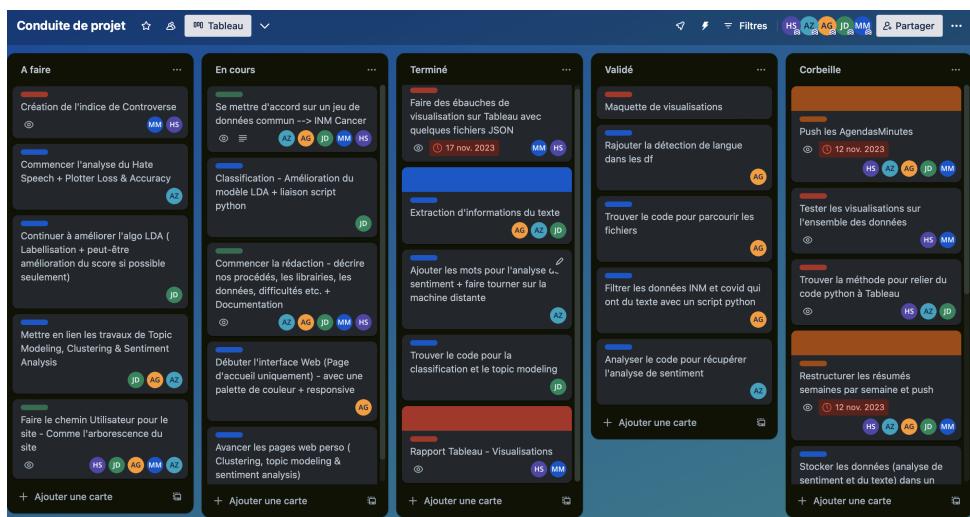


FIGURE 2.1 – Tableau Trello.

En début de chaque sprint, nous avons révisé notre planification en fonction des résultats du sprint précédent. Nous avons systématiquement archivé les récapitulatifs de chaque réunion sur GitHub, assurant ainsi un suivi rigoureux des échanges et des décisions prises.

En parallèle, nous avons participé en groupe au Travail d'Étude et de Recherche (TER) pendant les horaires de cours dédiés. De plus, chacun a consacré du temps personnel en dehors des cours pour s'aligner sur les attentes du sprint et garantir la progression du projet.

## 2.2 Présentation des données

### 2.2.1 Traitement des données

Nous avons reçu un ensemble de données, recoltées pendant la thèse de S.Benslimane, sous la forme de fichiers JSON, chaque fichier représentant un tweet.

```

{
  "id": "1283417024593625088",
  "creation": 1594825438.0,
  "user_id": "1283057398706573314",
  "social_network": "twitter",
  "nsfw": false,
  "request": ["a8e1fbe8-2136-43a8-9eba-42bfeb74bc00"],
  "metrics": {"a8e1fbe8-2136-43a8-9eba-42bfeb74bc00": {"retweet_count": 0, "reply_count": 0, "like_count": 0, "quote_count": 0}},
  "text": "Multiple eye bracelet\\n 3 different stones\\nTiger eye / hawk eye / bull eye Force / returns negative energies / soothes stress \\n#landofspirit #lithotherapy #pierresnaturelles #england #jewels #nature #sport #mode #sant\u00e9 #force #mensjewels https://t.co/gTpJc00joD",
  "first_save": 1649146953.786382,
  "hashtags": ["#LANDOFSPIRIT", "#LITHOTHERAPY", "#PIERRESNATURELLES", "#ENGLAND", "#JEWELS", "#NATURE", "#SPORT", "#MODE", "#SANT\u00e9", "#FORCE", "#MENSJEWELS"]
}

```

FIGURE 2.2 – Vue d'un tweet.

Pour traiter ces données, nous avons suivi plusieurs étapes. Tout d'abord, nous avons récupéré les données en dézippant les fichiers via les commandes du terminal d'une machine distante. Ensuite, nous avons trié ces fichiers en utilisant la structure des données. Certains fichiers contenaient des requêtes utilisateur qui répertoriaient les identifiants de tweets associés à un hashtag spécifique. Nous avons donc développé un script Python pour filtrer les tweets pertinents pour chacun des huit hashtags qui nous intéressaient. Cela nous a permis de réduire la taille des fichiers à traiter, passant de plus de 5 millions de tweets à 252 400 tweets. Voici un récapitulatif du nombre de tweets pour chaque

hashtag :

- cancer sport : 592
- cancer cannabis : 51 255
- cancer fasting : 1 036
- children\_covid\_vaccine : 1 393
- chloroquine : 17 669
- covid vaccine : 177 309
- carecall : 2 995
- lithotherapy : 151

Une fois les données filtrées, nous avons utilisé un autre script Python pour parcourir ces fichiers, extraire le texte de chaque tweet et effectuer un nettoyage à l'aide de TextBlob [6]. Cette analyse comprenait la tokenisation et la détection de la langue. Les résultats ont été stockés dans un dataframe et transformés en fichier CSV (249 456 lignes et 17 colonnes).

ID	Nb retweet	Nb like	Nb réponses	Nb citations	Hashtags	Texte	Mots
1147982997838340099	5	17	2	0.0	['#FASTING', '#PANCREATIC', '#CANCER']	#Fasting reduces intestinal toxicity from radi...	['Fasting', 'reduces', 'intestinal', 'toxicity...']
10725224640425985	0	0	0	0.0	['#ALKALINE', '#FOOD', '#HEALER', '#NATURALREM...']	https://t.co/TanAjkG1L\\nTips for the best tim...	['https', 't.co/TanAjkG1L', 'Tips', 'for', 't...']
1405348861078478852	0	0	0	0.0	['#COFFEE', '#CANCER', '#FASTING', '#INTERMITT...']	It's been testified that #coffee intake is b... b...	['It', 's', 'been', 'testified', 'that...']
1407724898550755329	0	0	0	0.0	['#REMINDERPOST', '#HINDU', '#FESTIVALS', '#PU...']	Tomorrow is Purnima\\n24 June 2021\\n#reminderp...	['Tomorrow', 'is', 'Purnima', '24', 'June', '#reminderp...']
627028927458988032	0	0	0	0.0	['#WANNAKNOW', '#FASTING', '#HEALTH', '#CANCER']	#WannaKnow Does #fasting help shrink a tumor o...	['WannaKnow', 'Does', 'fasting', 'help', 'shri...']

FIGURE 2.3 – Jeu de données nettoyé et traité.

## 2.3 Défis

Nous avons rencontré des défis lors du traitement de ces données, notamment en ce qui concerne la gestion du temps d'exécution et de la mémoire de nos ordinateurs. Pour améliorer l'efficacité de nos scripts, nous avons optimisé le traitement en parallèle sur plusieurs cœurs de nos machines.

De plus, l'analyse des tweets sur le vaccin COVID a été particulièrement longue, nécessitant plus de 31 heures. Nous avons également dû résoudre des problèmes de détection de langue, ce qui a exigé l'utilisation d'une solution alternative. Surmonter ces défis a demandé un effort supplémentaire pour résoudre les obstacles techniques rencontrés.

# Chapitre 3

## Analyses Descriptives

### Sommaire

---

<b>3.1 Visualisation des données . . . . .</b>	<b>10</b>
<b>3.2 Poursuite des analyses . . . . .</b>	<b>11</b>

---

### 3.1 Visualisation des données

Avant d'entamer notre étude approfondie sur les tweets controversés, nous avons jugé essentiel de procéder à des analyses descriptives préliminaires. Ces analyses nous ont permis d'examiner les métriques générales de nos tweets afin d'obtenir une vue d'ensemble de notre jeu de données et d'identifier des pistes potentielles pour des études plus approfondies.

Dans cette optique, nous avons effectué deux types d'analyses : une analyse des hashtags les plus fréquents à l'aide d'un nuage de mots, et la création d'un tableau de bord interactif sur Tableau.

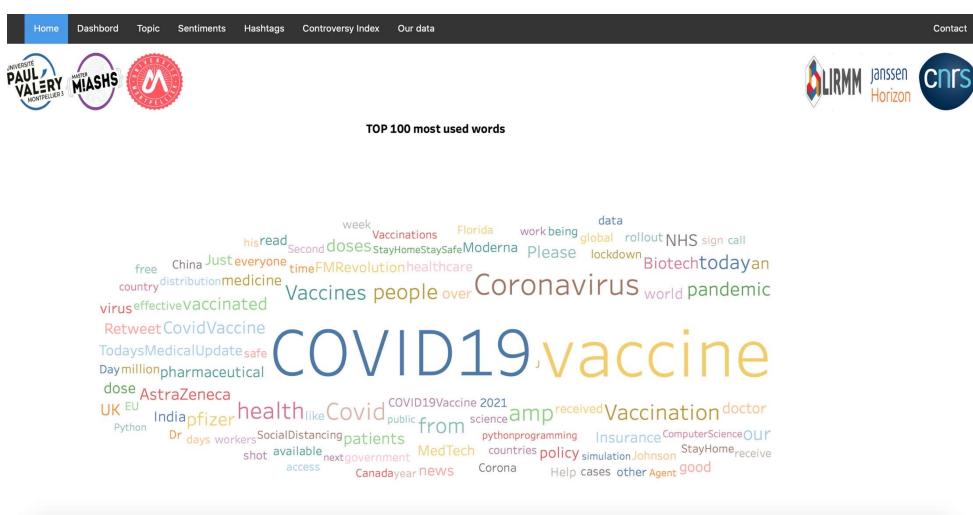


FIGURE 3.1 – Nuage de mots des hashtags.

Le nuage de mots des 100 hashtags les plus fréquents nous permet d'identifier les sujets les plus discutés dans notre ensemble de données. Parmi les hashtags les plus présents, nous remarquons des termes tels que *Covid19*, *vaccine*, *Coronavirus*, *Vaccines*, *pandemic*, etc. Cette observation suggère une forte préoccupation et une discussion intense autour de la pandémie de COVID-19 et des vaccins associés.

Par ailleurs, bien que notre jeu de données couvre diverses sujets, allant des interventions non médicamenteuses (INM) à la pandémie de COVID-19, nous avons constaté que les discussions liées au COVID-19 étaient prédominantes par rapport aux INM. En conséquence, nous avons pris la décision stratégique de recentrer notre analyse sur les INM. Cette décision s'avère être une utilisation plus efficace de nos ressources, en réduisant la charge de traitement des données et en nous permettant de nous concentrer sur un domaine spécifique pour des analyses plus approfondies.

Ainsi, le tableau de bord interactif se concentre uniquement sur les résultats relatifs aux interventions non médicamenteuses (INM). Il présente :

- le nombre de likes, de retweets, de commentaires et de citations ;
- le pourcentage de likes par hashtag ;
- le nombre de retweets par hashtag ;

- l'effectif des hashtags ;
- les combinaisons de hashtags les plus fréquentes.

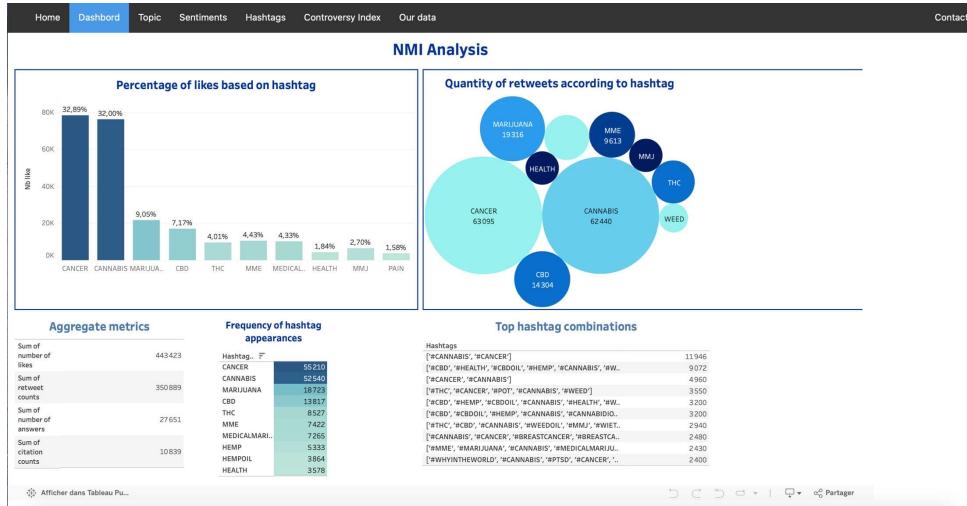


FIGURE 3.2 – Dashboard descriptif.

On observe que les hashtags `#cancer` et `#cannabis` sont les deux plus utilisés et les plus appréciés. De plus, ces deux hashtags apparaissent souvent ensemble, renforçant ainsi l'idée d'une association entre le cancer et le cannabis. On peut penser que le cannabis pourrait être largement discuté comme une possible intervention non médicamenteuse dans le traitement du cancer.

## 3.2 Poursuite des analyses

Après avoir réalisé ces analyses préliminaires, nous avons poursuivi nos investigations en analysant divers aspects des tweets controversés. L'objectif était d'approfondir notre compréhension de la controverse et d'identifier les différents éléments qui contribuent à sa dynamique. De plus, afin de présenter de manière claire et interactive les résultats de notre étude, nous avons entrepris la création d'un site web dédié. Chaque page de ce site web met en avant une analyse spécifique de la controverse, offrant ainsi une vue détaillée sur les différents aspects de notre étude.

# Chapitre 4

## Détection de communautés

### Sommaire

---

<b>4.1 Objectif de l'analyse</b> . . . . .	<b>13</b>
<b>4.2 Méthodologie</b> . . . . .	<b>13</b>
<b>4.2.1 Présentation des résultats</b> . . . . .	<b>13</b>

---

## 4.1 Objectif de l'analyse

Dans le cadre de cette étude, nous avons analysé un ensemble de tweets traitant du cancer en relation avec divers autres sujets comme le sport, le jeûne, et le cannabis. L'objectif principal était de découvrir des motifs et des relations entre les hashtags utilisés dans ces tweets mais aussi de pouvoir les quantifier, permettant ainsi de comprendre les discussions et les perceptions autour de ces thèmes.

## 4.2 Méthodologie

Nous avons débuté notre analyse en chargeant les données issues de trois sources distinctes, toutes liées au cancer. Pour équilibrer les ensembles de données nous avons procédé à un échantillonnage aléatoire de 1,5% sur le jeu Cancer et Cannabis qui est le plus volumineux. Cette méthode assure une représentation plus équilibrée des données dans notre analyse globale contenant un total de 2 391 tweets.

Pour analyser les données des tweets, nous avons employé l'algorithme de clustering K-Means [7], un choix populaire pour le partitionnement de données en groupes (ou clusters) basés sur la similarité. L'objectif est de regrouper les tweets de manière à ce que ceux qui sont dans le même cluster soient plus similaires entre eux qu'avec ceux d'autres clusters.

Les tweets ont été d'abord préparés par un nettoyage où nous avons converti les hashtags en chaînes de caractères. Cela simplifie l'analyse en éliminant les variations inutiles dans les formats de données.

Nous avons ensuite appliqué une vectorisation TF-IDF [8] (Term Frequency-Inverse Document Frequency) sur les hashtags. Cette méthode évalue l'importance d'un hashtag dans l'ensemble des tweets tout en réduisant l'impact des termes qui apparaissent très fréquemment, ce qui pourrait biaiser l'analyse.

Le choix du nombre de clusters est crucial. Nous avons utilisé la méthode du coude pour estimer le nombre optimal de clusters. Cette méthode implique le calcul de la somme des carrés des distances (WSS) entre les points et les centres de leurs clusters respectifs pour différents nombres de clusters et la recherche du point où l'ajout de plus de clusters n'apporte pas une réduction significative de la WSS.

Avec le nombre de clusters déterminé, nous avons appliqué l'algorithme K-Means. Le paramètre random state a été fixé pour assurer la reproductibilité des résultats. L'algorithme attribue chaque tweet à un cluster de sorte à ce que la variance intra-cluster soit minimisée.

### 4.2.1 Présentation des résultats

Après le clustering, chaque tweet a été associé à un cluster. Nous avons examiné les hashtags prédominants dans chaque cluster pour comprendre les thèmes spécifiques

discutés. Cela a été réalisé par l'agrégation des hashtags de tous les tweets d'un même cluster et en comptant leur fréquence d'apparition.

Chaque cluster a été caractérisé par les trois hashtags les plus fréquents, ce qui aide à identifier rapidement le focus de discussion dans chaque groupe. Par exemple, un cluster pourrait être dominé par des hashtags liés au cannabis médicinal dans le contexte du cancer, tandis qu'un autre pourrait concerner le cancer et le jeûne.

Le processus de clustering a permis de regrouper les tweets en catégories significatives basées sur les hashtags. Par exemple, certains clusters étaient fortement associés à des hashtags comme Cannabis et Cancer, indiquant une discussion fréquente sur l'usage médical du cannabis dans le contexte du cancer.

Nous avons élaboré un graphe de co-occurrence pour mettre en évidence les relations entre différents clusters basés sur leurs hashtags communs à l'aide de Pyvis [9]. Ce graphe illustre non seulement l'interconnexion entre divers sujets, mais aussi quantifie le nombre de co-occurrences, offrant ainsi une mesure claire des liens entre les clusters. Les interactions se distinguent par l'épaisseur des liens, ce qui souligne visuellement l'intensité de leur relation. Pour l'analyse de ce graphe, nous avons utilisé l'algorithme de Louvain [10] pour la détection de communautés, révélant des groupes de discussion plus vastes et interconnectés. De plus, nous avons employé l'algorithme de Force Atlas [11] pour la visualisation, améliorant la lisibilité et l'interprétation des structures complexes du graphe.

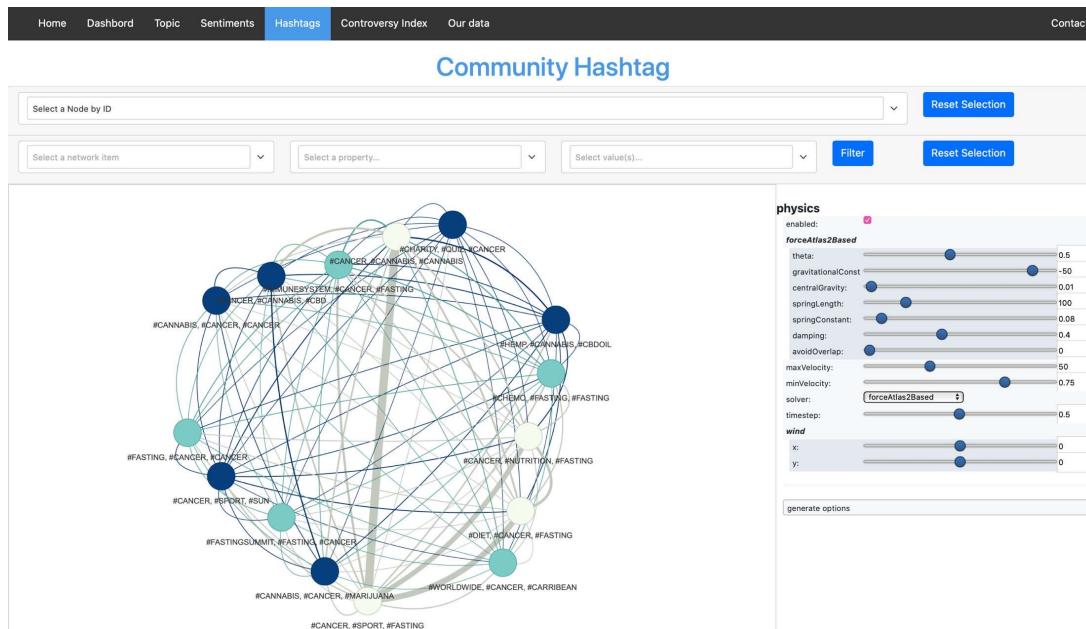


FIGURE 4.1 – Communauté d’Hashtags.

La visualisation obtenue avec ces algorithmes met en lumière les connexions entre différents thèmes et facilitent l'identification des principaux sujets de discussion ainsi que leurs interrelations. Ce graphique est indispensable pour saisir la structure et la dynamique des échanges sur Twitter concernant le cancer, aidant les chercheurs et les praticiens à mieux comprendre les patterns de communication publique autour de cette maladie.

# Chapitre 5

## Analyse des sentiments et des émotions

### Sommaire

---

<b>5.1 Objectifs de l'analyse</b> . . . . .	<b>16</b>
<b>5.2 Méthodologie</b> . . . . .	<b>16</b>
<b>5.3 Présentation des résultats</b> . . . . .	<b>17</b>

---

## 5.1 Objectifs de l'analyse

Notre objectif principal consiste à examiner la diversité des sentiments et des émotions et des utilisateurs concernant l'utilisation des méthodes non médicamenteuses (INM) pour traiter le cancer. Nous cherchons à comprendre la manière dont ces méthodes sont perçues et discutées par le public, en mettant particulièrement l'accent sur les aspects émotionnels et les nuances de la controverse qui les entourent. En explorant les discours des utilisateurs, nous espérons capturer une image holistique des attitudes envers les INM dans le contexte du traitement du cancer.

## 5.2 Méthodologie

Dans le cadre de notre projet, nous avons utilisé des modèles de traitement du langage naturel (NLP) pré-entraînés, fournis par Hugging Face, pour analyser les sentiments et les émotions exprimés dans les tweets concernant le Covid et les interventions non médicamenteuses (INM).

Notre objectif principal était de comprendre comment les utilisateurs expriment leurs réactions et leurs perceptions à travers différents vecteurs émotionnels et sentimentaux sur des plateformes de communication comme Twitter. Nous cherchions à déterminer si les utilisateurs manifestaient des sentiments positifs à l'égard de l'utilisation d'approches moins conventionnelles pour traiter le cancer.

Pour mener à bien cette analyse, nous avons choisi le modèle Twitter-RoBERTa-Base-Sentiment-Latest [12]. Ce modèle, spécifiquement entraîné sur 128 millions de tweets, est idéal pour comprendre les nuances du langage informel et elliptique utilisé sur Twitter. Nous pouvons ainsi capturer avec précision les sentiments des utilisateurs, qui peuvent varier subtilement en fonction du contexte de chaque tweet. Cette analyse nous permet d'identifier les tendances des sentiments dans les discussions sur Twitter concernant le Covid et les INM, offrant des indicateurs sur les opinions publiques et les controverses autour de ces sujets.

En complément de cette analyse, nous avons utilisé le modèle Twitter-RoBERTa-Base-Emotion-Multilabel-Latest [13] pour observer les réactions émotionnelles. Basé sur une large collection de 58 millions de tweets, ce modèle distingue les nuances émotionnelles qui vont au-delà des simples expressions de sentiments. Il explore la gamme complète des émotions exprimées par les utilisateurs, permettant une compréhension plus approfondie du paysage émotionnel sur Twitter. Cette approche enrichit notre analyse en révélant comment les émotions influencent la réception et la propagation des informations. Ces deux modèles Pré-entraînés sont basés sur le modèle RoBERTa [14] (Robustly optimized BERT approach) développé par Facebook AI Research.

### 5.3 Présentation des résultats

Tout d'abord, nous avons explorer la répartition des scores moyens des sentiments et des émotions à travers notre ensemble de données. Les visualisations ci-dessous présentent ces scores moyens respectifs, offrant ainsi un aperçu de la distribution des sentiments et des émotions sur notre jeu de données.

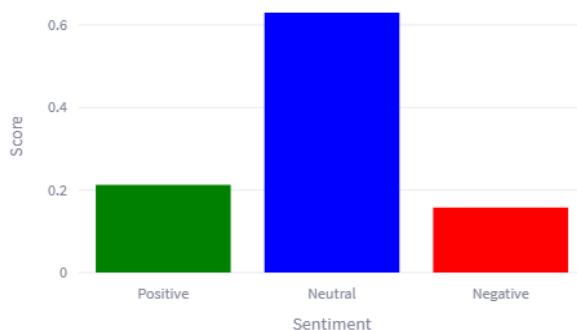


FIGURE 5.1 – Scores moyens des sentiments.

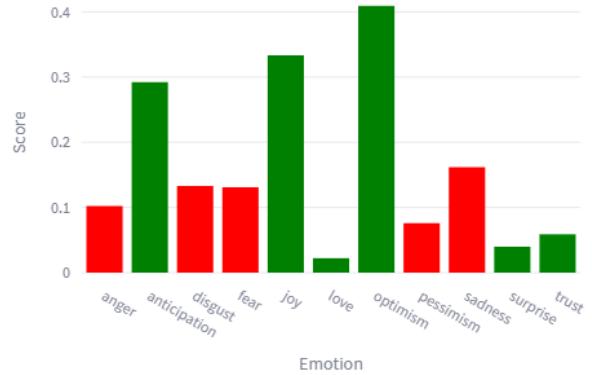


FIGURE 5.2 – Scores moyens des émotions.

La figure 4.2 montre que le sentiment neutre domine nettement avec un score approximativement de 0.6, indiquant une tendance générale des tweets à exprimer une absence de sentiment positif ou négatif fort. Le sentiment positif et négatif montrent des scores plus bas, avec le positif légèrement au-dessus de 0.2 et le négatif juste en dessous, reflétant une présence modérée de ces sentiments dans les tweets analysés.

La figure 4.3 indique une prédominance de l'émotion 'joie', avec un score autour de 0.3, suivi de près par 'optimisme'. Ces deux émotions positives sont les plus marquées parmi les utilisateurs, suggérant une réaction globalement positive aux sujets discutés. D'autre part, les émotions telles que la 'peur', 'la tristesse', et 'la surprise' montrent des scores significativement plus bas, avec la 'confiance' et 'l'amour' exprimées à des niveaux encore inférieurs, ce qui peut indiquer une réticence à partager des émotions profondément personnelles ou à afficher une confiance explicite dans le contexte des sujets abordés.

Ces analyses suggèrent que, bien que les discussions puissent être globalement neutres en termes de sentiment, il y a un penchant vers des réactions émotionnelles positives lorsque des émotions spécifiques sont exprimées. Cette dualité entre neutralité sentimentale et expressivité émotionnelle positive peut fournir des indications sur l'état d'esprit des utilisateurs et leur réaction aux thèmes discutés.

Par ailleurs, nous avons réalisé une t-SNE [15] (t-distributed stochastic neighbor embedding) sur un échantillon représentatif de 3000 tweets, sélectionné parmi un total de 55 000 tweets. Cet échantillon a été choisi de manière à refléter la diversité et la distribution des sentiments présents dans l'ensemble des données, à partir d'une sélection aléatoire de tweets.

À partir de ce graphique, nous pouvons identifier trois groupes distincts correspondant aux sentiments neutres en bleu, positifs en vert et négatifs en rouge. Les tweets neutres et positifs se chevauchent quelque peu, suggérant que les utilisateurs partagent souvent des opinions qui ne sont ni clairement positives ni clairement neutres, mais quelque part entre les deux. Cela peut également refléter une certaine hésitation ou un questionnement de la part des utilisateurs sur le sujet. En revanche, les tweets négatifs forment un groupe plus compact, ce qui suggère que les utilisateurs sont peut-être plus unanimes ou plus directs lorsqu'ils expriment des sentiments négatifs.

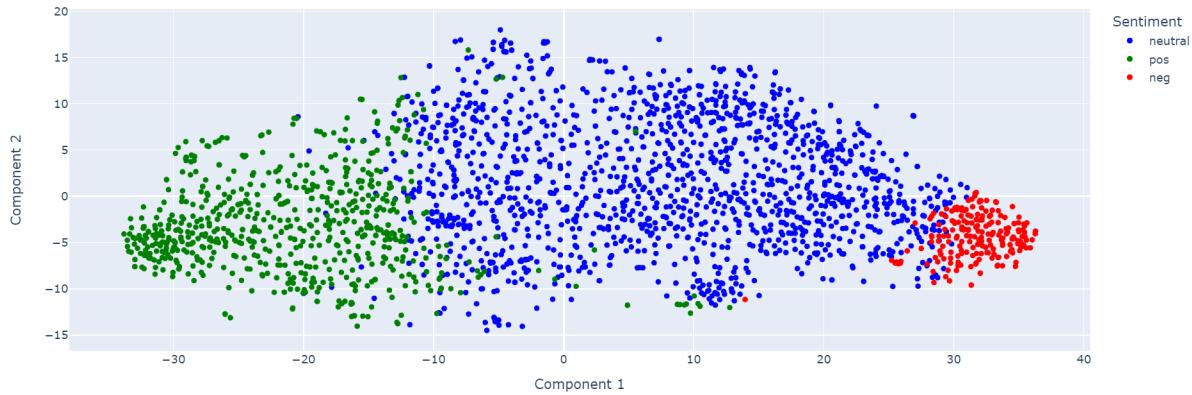


FIGURE 5.3 – t-SNE Visualization of Feature Spaces for Sentiment Analysis

Cette répartition nous donne un aperçu utile des différents points de vue exprimés sur Twitter en ce qui concerne les INM. En analysant plus en détail les tweets de chaque groupe, nous pourrions mieux comprendre les opinions spécifiques et les discussions qui animent le débat sur ce sujet controversé.

En complément de ces visualisations, nous avons développé une application Streamlit déployée sur Hugging Face Spaces [16], permettant une visualisation interactive des sentiments et des émotions exprimés dans des tweets sélectionnés aléatoirement.

# Chapitre 6

## Analyse des opinions

### Sommaire

---

<b>6.1 Objectifs de l'analyse</b>	20
<b>6.2 Méthodologie</b>	20
<b>6.3 Présentation des Résultats</b>	21
<b>6.4 Défis et perspectives</b>	22

---

## 6.1 Objectifs de l'analyse

Notre objectif est de répondre à la question suivante : comment les modèles d'apprentissage automatique peuvent-ils être efficacement utilisés pour classifier les tweets exprimant des opinions favorables, neutres ou opposées concernant l'utilisation des interventions non médicamenteuses (INM) dans le traitement du cancer ?

L'intérêt principal de cette problématique réside dans son exploration des attitudes publiques envers les méthodes de traitement alternatives du cancer, telles que le jeûne, la consommation de cannabis médicinal ou l'exercice physique.

## 6.2 Méthodologie

Dans un premier temps, nous avons développé et utilisé un script de labellisation spécialement conçu pour classer les tweets en trois catégories : FOR (pour), AGAINST (contre), et NEUTRAL (neutre). Au total, 500 tweets ont été labelisés, fournissant une base solide pour tester et entraîner nos modèles de classification des opinions. Nous avons adopté une répartition de 80% des données pour l'entraînement et 20% pour les tests, afin de maximiser l'apprentissage tout en validant efficacement la performance du modèle. Cette base, relativement petite, a été choisie en raison de contraintes de temps qui ne permettaient pas un étiquetage plus étendu.

En ce qui concerne l'analyse des opinions exprimées dans les tweets, nous avons opté pour un fine-tuning d'un modèle pré-entraîné BERT [17] (Bidirectional Encoder Representations from Transformers). L'architecture de BERT est particulièrement adaptée à notre problématique, car elle permet de comprendre les nuances du langage naturel grâce à sa capacité à traiter les mots dans leur contexte bidirectionnel. De plus, BERT est largement documenté, ce qui a facilité le processus de fine-tuning.

Le fine-tuning consiste à ajuster les paramètres d'un modèle pré-entraîné sur un ensemble de données spécifique (450 tweets). Dans notre étude, nous avons mappé les catégories des opinions comme suit afin de permettre la tokenisation : 0 pour 'FOR', 1 pour 'AGAINST', et 2 pour 'NEUTRAL'. Nous avons ensuite adapté les couches supérieures du modèle BERT pour qu'elles reflètent mieux les spécificités de notre jeu de données de tweets. Cette adaptation a impliqué le réentraînement du modèle sur nos données labellisées, utilisant l'instance Trainer de Hugging Face [18] pour faciliter l'apprentissage. L'utilisation de cette instance permet d'optimiser le processus d'apprentissage, garantissant ainsi que le modèle adapte au mieux ses prédictions aux nuances des opinions exprimées dans les tweets concernant les interventions non médicamenteuses (INM).

TABLE 6.1 – Détails de la phase d’entraînement

Parameters	Valeur
Number of Examples	450
Number of Epochs	3
Instantaneous Batch Size Per Device	8
Total Train Batch Size (w. parallel, distributed & accumulation)	8
Gradient Accumulation Steps	1
Total Optimization Steps	171

Le nombre total d’exemples traités était de 450, répartis sur 3 époques. Nous avons fait le choix de 3 époques seulement afin de limiter l’apparition de l’over-fitting trop tôt dans la phase d’entraînement et aussi par contrainte de ressources computationnelles. Chaque lot instantané traité contenait 8 tweets, et compte tenu que notre configuration n’incluait pas de parallélisation, de distribution sur plusieurs appareils, ni d’accumulation de gradients, la taille totale des lots d’entraînement était également de 8, ce qui simplifiait le processus de mise à jour des gradients puisque chaque mise à jour était effectuée séquentiellement après le traitement de chaque lot. Le pas d’accumulation des gradients était fixé à 1, simplifiant le processus de mise à jour des gradients sans accumulation supplémentaire. Enfin, le nombre total d’étapes d’optimisation s’élevait à 171, indiquant le nombre de fois que les poids du modèle ont été ajustés au cours de l’entraînement.

### 6.3 Présentation des Résultats

Les résultats de l’entraînement du modèle sont présentés ci-dessous. Ces données illustrent les variations de perte et de précision au fil des époques.

TABLE 6.2 – Détails de l’entraînement du modèle

Epoch	Learning Rate	Loss
1	$3.725 \times 10^{-5}$	0.3329
2	$1.863 \times 10^{-5}$	0.3242
3	0.0	0.195

TABLE 6.3 – Accuracy du modèle par classe et globale

Accuracies	Values
Global Accuracy	0.86
Class 0 FOR Accuracy	0.971
Class 1 AGAINST Accuracy	0.500
Class 2 NEUTRAL Accuracy	0.643

Les résultats montrent une amélioration significative de la perte au fil des époques, indiquant que le modèle s’adapte bien aux données. La précision globale du modèle atteint

86%, ce qui est prometteur. Cependant, la précision par classe révèle un déséquilibre, avec une performance particulièrement basse pour la classe 1 qui sont les opinions "against" (50%). Cette situation est causée par le déséquilibre dans les classes de nos données, où justement la classe 'against' est minoritaire. Ce déséquilibre peut conduire à une moins bonne généralisation du modèle sur les classes moins représentées, ce qui suggère la nécessité d'adopter des stratégies afin de contrer ce biais.

## 6.4 Défis et perspectives

Les performances de notre modèle ont été significativement impactées par un déséquilibre des classes dans le jeu de données, avec la classe "against" minoritaire. Afin de remédier à ce problème, nous envisageons deux principales stratégies d'amélioration :

1. **Augmentation de données** : Cette technique vise à générer artificiellement de nouveaux exemples de formation en modifiant légèrement les tweets existants, ce qui peut inclure le changement de certains mots, la reformulation ou l'utilisation de synonymes. L'objectif serait de créer un ensemble de données plus équilibré où les trois classes (FOR, AGAINST, NEUTRAL) seraient mieux représentées. Cela devrait aider à réduire le biais du modèle envers la classe majoritaire et améliorer sa capacité à généraliser à partir de données plus diverses.
2. **Fonctions de Perte pour Gérer le Déséquilibre des Classes** : Adapter la fonction de perte pour tenir compte du déséquilibre des classes peut également être une approche efficace. En utilisant une fonction de perte qui pénalise davantage les erreurs sur les classes minoritaires, ou en attribuant des poids plus élevés aux classes moins représentées, nous pouvons inciter le modèle à prêter plus d'attention à ces classes. Dans notre contexte des déséquilibres significatifs des classes dans les jeux de données de classification, deux fonctions de perte se distinguent par leur efficacité à aborder ce problème :
  - (a) **Perte Entropique Croisée Pondérée (Weighted Cross-Entropy Loss [19])** : Cette fonction de perte est une variation de la perte entropique croisée standard, souvent utilisée dans les tâches de classification multiclass. La spécificité ici est l'introduction de poids attribués à chaque classe, généralement inversément proportionnels à la fréquence des classes dans le jeu de données. Cette pondération permet d'augmenter l'importance des erreurs sur les classes moins représentées, contribuant ainsi à un apprentissage plus équilibré. La formule ajustée de cette perte peut être exprimée comme suit :

$$L = - \sum_{i=1}^C w_i y_i \log(p_i)$$

où  $w_i$  représente le poids de la classe  $i$ ,  $y_i$  est la vérité terrain et  $p_i$  est la probabilité prédictive pour la classe  $i$ .

- (b) **Focal Loss [20]** : Développée initialement pour aborder les difficultés dans la détection d'objets où le déséquilibre entre les classes de fond et d'objet est prononcé, la Focal Loss modifie la perte entropique croisée en y ajoutant un facteur de mise au point. Ce facteur réduit l'impact des exemples facilement classifiés, permettant au modèle de se concentrer sur les cas difficiles, souvent ceux des classes minoritaires. Cette fonction est particulièrement utile pour prévenir la domination de la classe majoritaire pendant l'entraînement. Sa formule est la suivante :

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

où  $\alpha_t$  et  $\gamma$  sont des hyperparamètres qui contrôlent respectivement l'équilibrage des classes et le taux auquel les exemples faciles sont pénalisés.

Ces stratégies devraient contribuer à une meilleure équité dans la prédiction des classes et à une augmentation de la précision globale du modèle.

# Chapitre 7

## Topic Modelling et Latent Dirichlet Allocation

### Sommaire

---

7.1 Objectif de l'analyse . . . . .	25
7.2 Méthodologie . . . . .	25
7.3 Présentation des résultats . . . . .	26

---

## 7.1 Objectif de l'analyse

L'objectif principal de cette analyse est d'utiliser le Topic Modelling pour explorer et comprendre les sujets controversés présents dans un ensemble de tweets. Cette approche permet d'identifier des clusters de mots significatifs, révélant ainsi les thèmes clés abordés dans ces tweets.

Nous nous appuyons sur le modèle LDA (Latent Dirichlet Allocation) [21], développé par David Blei, pour atteindre cet objectif. Il permet d'extraire des clusters de mots qui peuvent être interprétés comme des thèmes principaux dans les documents textuels. En l'appliquant à notre corpus de tweets, nous visons à déduire et à explorer les thèmes clés abordés dans ces discussions en ligne.

## 7.2 Méthodologie

Tout d'abord, nous avons combiné trois ensembles de données de tweets pour construire un modèle LDA sur les interventions non médicales pour le traitement du cancer, la pratique du sport, la consommation de cannabis et le jeûne intermittent.

Après l'extraction du texte, les étapes suivantes ont été réalisées :

- Suppression des mots vides et de ceux de moins de trois lettres, tels que *the* et *a*.
- Lemmatisation et stemming des mots pour ramener les mots à leur forme canonique et réduire les mots à leur racine, par exemple *happiness* devient *happy*.
- Conversion du texte en un dictionnaire et en une représentation de fréquence de sacs de mots.

Nous avons initié les premiers modèles LDA en utilisant la bibliothèque Gensim [22], en ajustant tous les paramètres qui impactent la répartition des sujets et des mots, incluant le nombre d'itérations et de sujets, ainsi que les valeurs d'alpha et de beta. L'alpha, représentant la concentration de Dirichlet pour la distribution des sujets par document, et le beta, pour la distribution des mots par sujet, influent sur la dispersion des sujets et des mots. Des valeurs plus élevées induisent une uniformité accrue, affectant la cohérence et la diversité des sujets à travers l'ensemble des documents.

Finalement, nous avons choisi d'utiliser la bibliothèque scikit-learn [23], en ajustant le nombre de sujets et la méthode d'apprentissage. Cette décision a abouti à un score de cohérence de 0,53, ce qui est considéré comme un bon résultat. En effet, un modèle est jugé approprié lorsque son score se situe entre 0,4 et 0,7. Grâce à cela, nous avons pu définir des thèmes clairs et distincts, que nous développerons dans la section suivante de cette analyse.

## 7.3 Présentation des résultats

Après l'application du modèle LDA, nous avons obtenu les résultats ci-dessous.

Label	Top Words
Treatment with Cannabis	[cancer, cannabi, connai, cell, marijuana, cure, studi, treatment, cannabinoid, kill]
Diet and Intermittent Fasting	[cancer, cannabi, breast, connai, breastcanc, analysi, extens, v, sport, wast]
Health Benefits of Cannabis	[cancer, health, daili, connai, cannabi, latest, benefit, cbd, hemp, ill]
CBD and Pain Management	[cancer, connai, cbd, hemp, boil, health, anxiety, cannabi, cod, epilepsi]
Health and CBD Oil Usage	[cancer, connai, cb, cod, hemp, list, mmm, e, wietoli, weedoil]
THC and Tumor Inhibition	[cancer, connai, mme, tumor, cod, growth, cannabinoid, cannabi, the, marijuana]
Research on Cancer Prevention	[cancer, connai, cannabi, amp, pain, fight, effect, growth, patient, cell]
Epilepsy and Medical Marijuana	[cancer, marijuana, connai, cannabi, cod, medic, mme, medicalmarijuana, despoil, help]
Medical Marijuana Usage	[cancer, connai, cannabi, marijuana, leukaemia, medicalmarijuana, mmm, health, medicin, cod]
Health Risks and Marijuana	[cancer, connai, marijuana, medicalmarijuana, pushkin, hemp, donald, cod, weed, amp]
Cannabis as a Cure	[cancer, cannabi, connai, marijuana, stop, eat, cure, plant, food, amp]
Lifestyle and Cancer Support	[cancer, connai, cannabi, health, fast, diet, sport, cure, port, amp]
THC and Cervical Cancer Treatment	[cancer, cannabi, studi, would, new, connai, great, suggest, cervic, disappear]

FIGURE 7.1 – Sujets et mots clefs.

Notre modèle LDA révèle des sujets distincts, que nous avons étiquetés à l'aide de ChatGPT [24], avec des mots-clés pertinents pour chaque sujet. Chaque tweet a été associé à l'un de ces sujets spécifiques. Il est important de noter que les discussions sur l'utilisation du cannabis comme traitement non médical du cancer ont prédominé. Cette tendance peut s'expliquer par le grand nombre de tweets dans notre corpus portant sur ce sujet.

L'examen des mots les plus fréquemment utilisés a confirmé que certains mots, en particulier le mot *cancer*, étaient particulièrement récurrents, ce qui reflète l'importance accordée par les utilisateurs à ce sujet.

Par ailleurs, notre analyse met en évidence une concentration significative du cannabis médical dans les discussions sur le cancer et la gestion de la douleur. D'autres thèmes, tels que le jeûne intermittent et le sport, sont également présents, mais dans une moindre mesure. En scrutant de plus près les groupes de mots, nous observons des similitudes entre certains d'entre eux, comme l'association fréquente des termes "cannabis", "cancer", "cbd" et "medical". Cela suggère un fort intérêt pour l'utilisation du cannabis à des fins médicales, en particulier dans le contexte du traitement du cancer et de la gestion de la douleur. Ces résultats mettent en lumière l'intérêt croissant pour des approches thérapeutiques alternatives dans la lutte contre le cancer.

De surcroît, dans l'optique d'améliorer la représentativité des sujets, une approche aurait pu consister à effectuer un échantillonnage aléatoire du jeu de données consacré aux tweets sur la consommation du cannabis. Cette stratégie aurait favorisé une répartition plus équilibrée des sujets et renforcé l'homogénéité des thématiques au sein de notre ensemble de données.

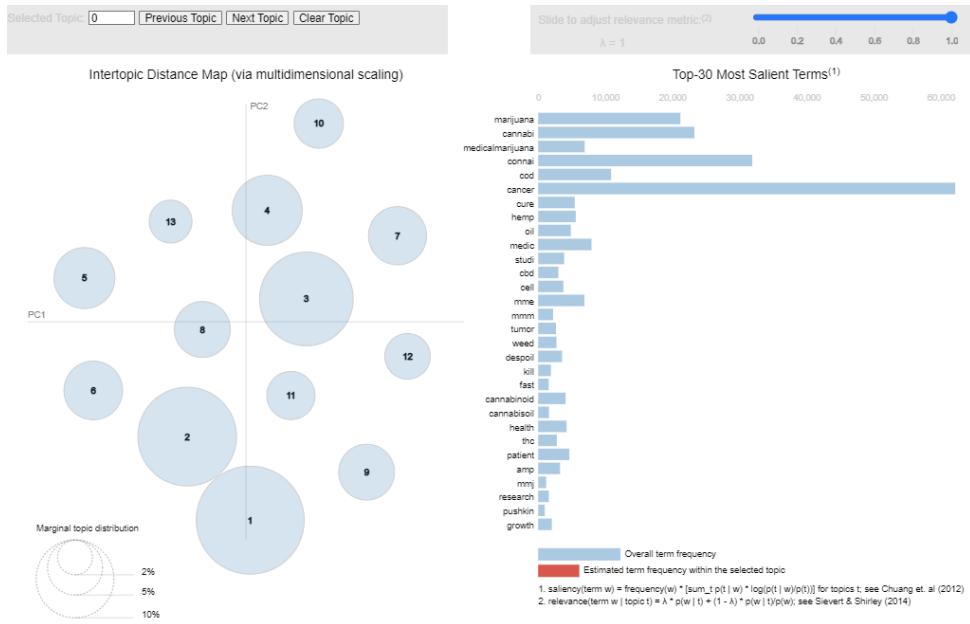


FIGURE 7.2 – Visualisation du Topic Modelling avec le LDA.

La visualisation de la carte de distance interthématische du modèle LDA révèle que les thèmes sont représentés par des cercles, dont la taille correspond à leur importance. Les cercles plus grands indiquent ainsi des thèmes plus saillants. Un cercle plus petit à l'intérieur d'un autre cercle indique que le thème est lié à son thème parent. De plus, une barre de pertinence sur la droite affiche les termes les plus caractéristiques pour un sujet sélectionné. Cette barre peut être ajustée pour équilibrer la fréquence et l'exclusivité des termes.

# Chapitre 8

## Indice de controverse

### Sommaire

---

<b>8.1 Objectif de l'analyse</b> . . . . .	<b>29</b>
<b>8.2 Méthodologie</b> . . . . .	<b>29</b>
<b>8.3 Présentation des résultats</b> . . . . .	<b>30</b>

---

## 8.1 Objectif de l'analyse

L'indice de controverse est une métrique essentielle dans l'analyse des tweets controversés, car il offre une mesure quantitative de la polémique présente dans ces discussions. En fournissant une valeur numérique de la controverse, cet indice permet de différencier les tweets qui suscitent des débats intenses des autres. Cette distinction est cruciale pour comprendre l'ampleur et l'impact des discussions sur Twitter, en particulier dans des domaines sensibles comme la santé.

## 8.2 Méthodologie

Pour établir l'indice de controverse, nous avons pris en compte plusieurs facteurs :

- Le nombre de likes
- Le nombre de retweets
- Le nombre de commentaires
- Le nombre de citations
- Le score positif des sentiments exprimés
- Le score négatif des sentiments exprimés

A partir de ces facteurs que nous avons normalisés afin de les rendre comparables entre 0 et 1, l'utilisateur attribue des poids dans le but de pondérer leur importance dans le calcul de l'indice de controverse. Celui-ci est obtenu par la somme pondérée de ces facteurs et de leurs poids respectifs [25] :

$$\begin{aligned}\text{Indice de controverse} = & (\text{Poids\_likes} \times \text{Likes}) \\ & + (\text{Poids\_retweets} \times \text{Retweets}) \\ & + (\text{Poids\_commentaires} \times \text{Commentaires}) \\ & + (\text{Poids\_citations} \times \text{Citations}) \\ & + (\text{Poids\_positif} \times \text{Score\_positif}) \\ & + (\text{Poids\_négatif} \times \text{Score\_négatif})\end{aligned}$$

Pour mieux comprendre le tweet généré aléatoirement, nous avons décidé de l'afficher avec ses caractéristiques afin de permettre une analyse approfondie, notamment par l'utilisateur attribuant les poids des facteurs.

Par la suite, nous avons constaté que, lorsque nous avons des tweets faiblement controversés, l'intérêt persistait à découvrir celui qui suscite un indice de controverse maximal selon les poids choisis.

C'est pourquoi nous avons également affiché le tweet comportant l'indice de controverse maximal pour les poids choisis par l'utilisateur. De ce fait, cela offre une analyse complète et concrète, quant à l'indice de controverse et ses facteurs.

## 8.3 Présentation des résultats

Étant donné que nous avons décidé de baser l'indice de controverse sur une attribution de poids dépendante de l'avis de l'utilisateur, plusieurs résultats ont été observés :

- La quantité des caractéristiques impacte l'indice de controverse :

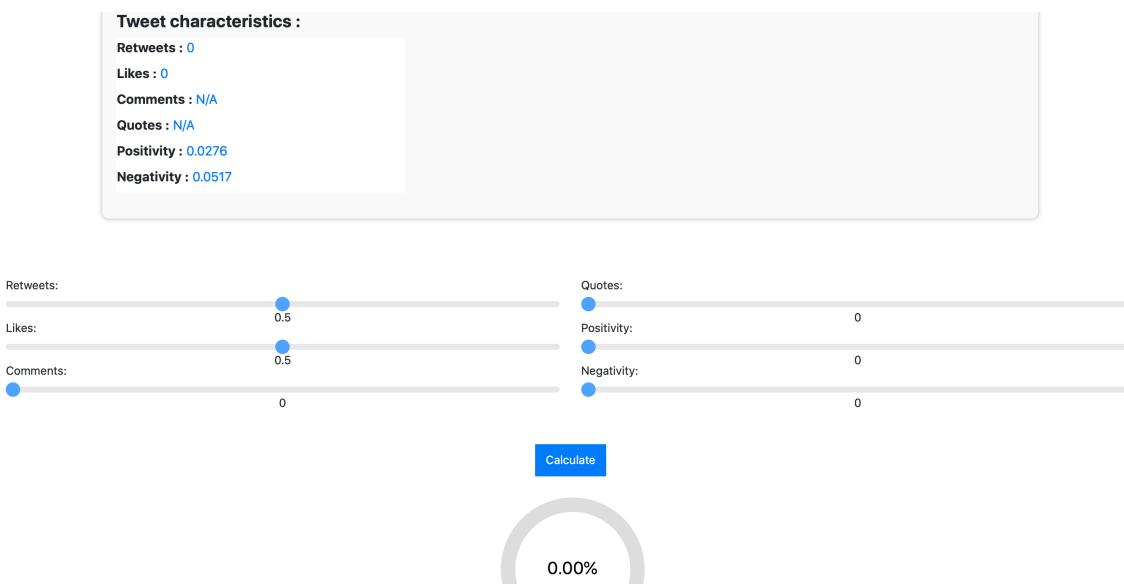


FIGURE 8.1 – Tweet impacté faiblement par les caractéristiques

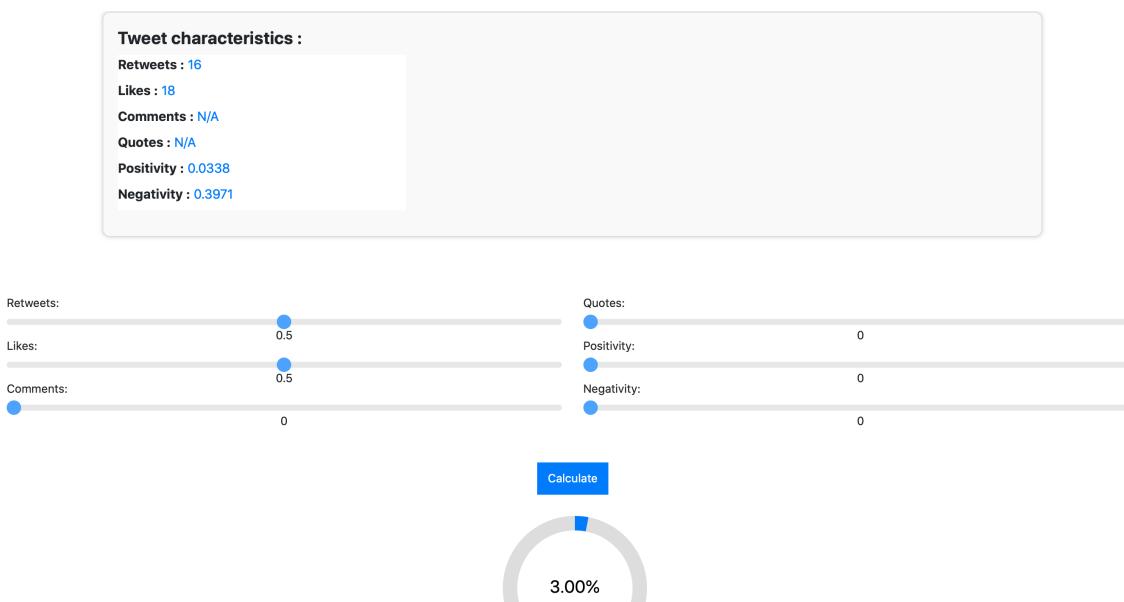


FIGURE 8.2 – Tweet impacté fortement par les caractéristiques

En effet, nous pouvons constater que pour les mêmes seuils, l'indice de controverse augmente lorsque les caractéristiques du tweet augmentent.

Cependant, les seuils jouent un rôle significatif dans le calcul de l'indice de controverse :

— Le choix des seuils impactent l'indice de controverse :

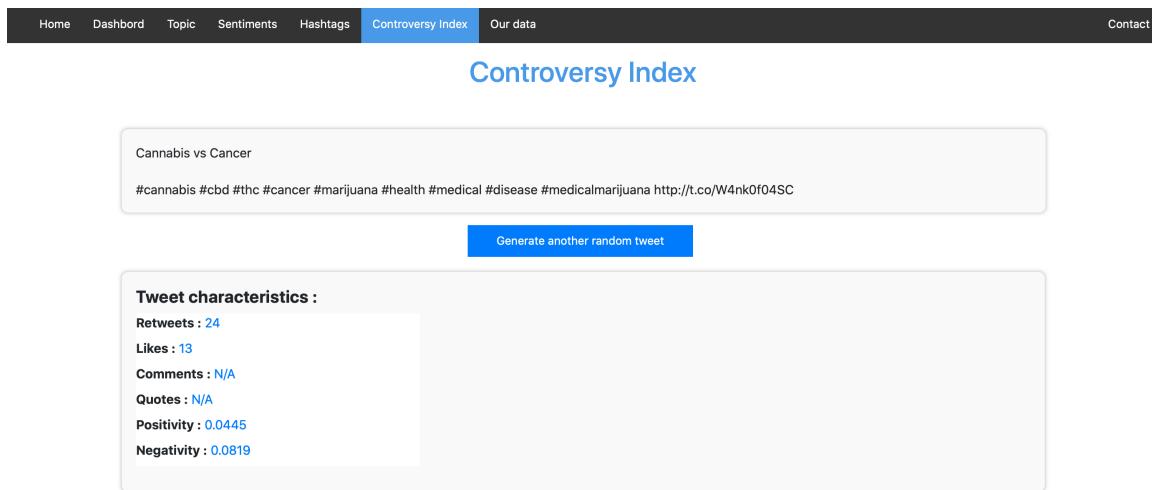


FIGURE 8.3 – Tweet analysé

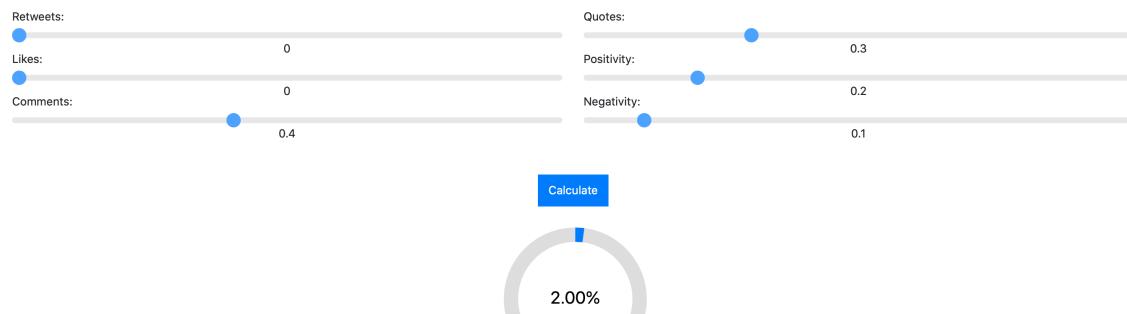
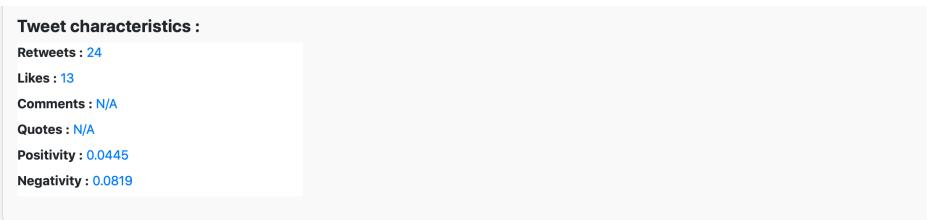


FIGURE 8.4 – Les seuils font diminuer l'indice

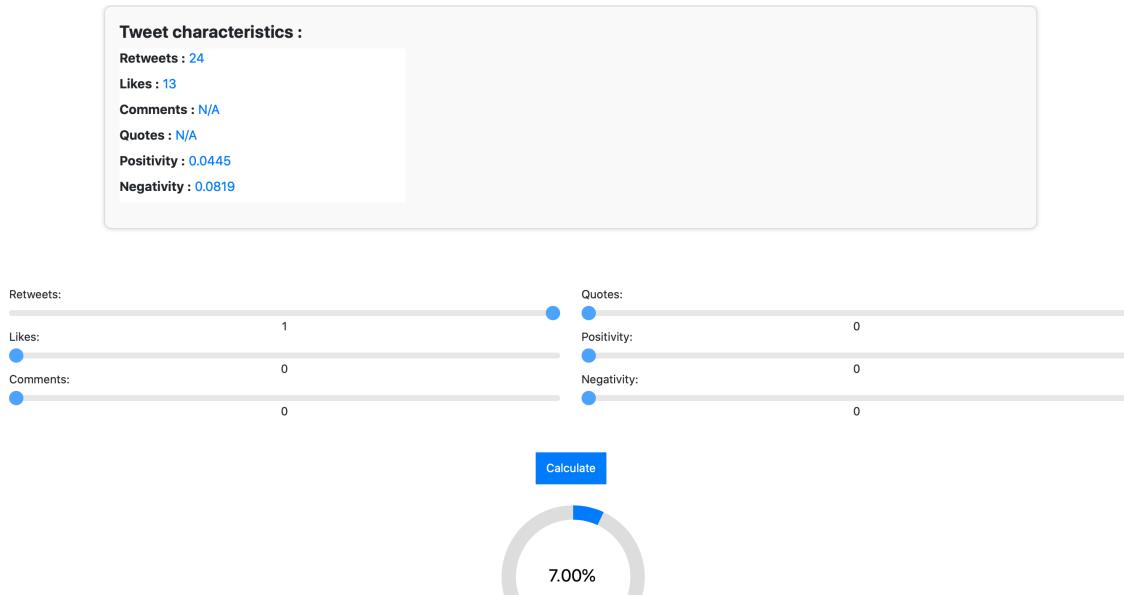


FIGURE 8.5 – Les seuils font augmenter l’indice

De ce fait, il est important de constater que les seuils peuvent avoir davantage d’impact sur l’indice de controverse que la quantité de caractéristiques.

Nous pouvons par conséquent conclure que l’assemblage des deux, des seuils choisis de manière adaptée et des forts caractéristiques contribuent de manière exponentielle l’indice de controverse.

- Le choix des seuils en plus de caractéristiques élevés ont un important impact :

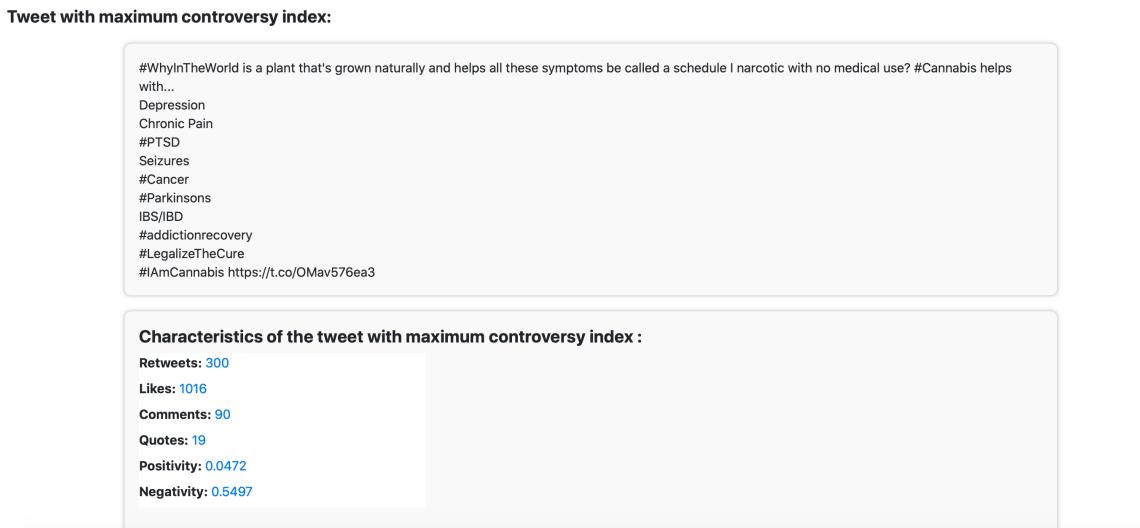


FIGURE 8.6 – Un des tweets comportant un indice de controverse maximal

Grâce à l’affichage du tweet comportant un indice maximal de controverse, nous pouvons affirmer que l’union de seuils bien choisis et la possession de forts caractéristiques impliquent un indice de controverse maximal.

# Conclusion

Cette expérience de projet TER a été profondément enrichissante pour nous tous. Elle nous a permis de mettre en pratique les compétences théoriques acquises au cours de notre formation universitaire dans un contexte réel de recherche. De plus, travailler en équipe sur un projet aussi ambitieux nous a permis d'affiner nos compétences en communication, en collaboration et en gestion de projet.

Nous avons conçu un site web regroupant l'intégralité de nos analyses. Chacune de ces analyses est présentée dans une section dédiée du site, permettant aux visiteurs de découvrir nos recherches, nos résultats ainsi que notre jeu de données. Cette plateforme constitue une ressource complète et intuitive, accessible à tous ceux qui souhaitent s'informer sur nos travaux et les approfondir.

Aussi, la rédaction de l'article "INM-Explain" pour la PFIA 2024 [26], Plate-forme Intelligence Artificielle la Plate-Forme Intelligence Artificielle (IA) réunissant des chercheurs, industriels et étudiants autour de conférences et d'ateliers consacrés à l'IA, a été une opportunité grattifiante de partager notre étude. Cet article synthétise nos analyses, met en lumière nos résultats et présente un cas d'étude qui les examine en détail.

En clôturant ce chapitre, nous sommes impatients de voir comment nos travaux pourront inspirer de nouvelles recherches et de nouvelles avancées dans le domaine de l'intelligence artificielle et de la santé.

# Bibliographie

- [1] France TV Info. Cannabis thérapeutique et réduction des douleurs liées au cancer, 2023. Consulté le : 2024-05-04.
- [2] Service Public. Le cannabis à usage médical, 2023. Consulté le : 2024-05-04.
- [3] Chandni Hindocha, Tom P Freeman, Meryem Grabski, and Jack B Stroud. An overview of products and bias in research. *PubMed Central*, 2019, 2023. Consulté le : 2024-05-04.
- [4] Trello.
- [5] Github.
- [6] Textblob.
- [7] Kmeans.
- [8] Tf-idf.
- [9] Pyvis.
- [10] Algorithme de louvain, que et al. scalable community detection with the louvain algorithm, in proceedings of the2015 ieee international parallel and distributed processing symposium, hyderabad, india, 2015, 28-37.
- [11] Algorithme de force atlas, jacomy, m., venturini, t., heymann, s., bastian, m. (2014). forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *plos one*, 9(6), e98679.
- [12] Modèle roberta pour l'analyse du sentiment sur twitter.
- [13] Modèle roberta pour l'analyse de l'émotion sur twitter.
- [14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized bert pretraining approach. *arXiv*, 2019.
- [15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 :2579–2605, 2008.
- [16] Hugging face spaces.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

- [18] Hugging Face. Trainer : Main classes in transformers, 2023. Accessed : 2024-05-04.
- [19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. The real-world-weight cross-entropy loss function : Modeling the costs of mislabeling. *arXiv preprint arXiv :2001.00570*, 2020.
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv preprint arXiv :1708.02002v2*, 2017.
- [21] Blei David. Latent dirichlet allocation.
- [22] Prabhakaran Selva. Topic modeling visualization – how to present the results of lda models ?
- [23] Abell Nicholas. Visualizing unsupervised text classification output with pyldavis.
- [24] OpenAI. Chatgpt.
- [25] Méthode mathématique d'analyse multicritère — wikipédia, l'encyclopédie libre.
- [26] PFIA. Pfia.