



# Projet AI Emotion

**Membres du groupe :**

BESSE Tony - (*N° 22400107*)  
LÉ Alain - (*N° 22415897*)  
VERBOVY Anton - (*N° 22405645*)

**Encadré par :**

TODOROV Konstantin  
LECOURT Florian

25 mai 2025

## Résumé

Ce projet vise à analyser la manière dont les émotions exprimées dans les tweets, en particulier ceux liés à la science, influencent la réception et la diffusion de l'information sur les réseaux sociaux. Un corpus de 1140 tweets a été annoté manuellement selon sept catégories émotionnelles, puis exploité pour entraîner et comparer plusieurs modèles de classification (SVM, Naive Bayes, BERT). Ce travail met en lumière l'importance des émotions dans le discours scientifique en ligne et ouvre des perspectives pour l'étude des dynamiques émotionnelles et informationnelles sur les réseaux sociaux.

## **Remerciements**

Au nom de l'équipe d'Optimus Gang, nous tenions sincèrement à remercier nos encadrants, Mr. Todorov K. et Mr. Lecourt F., pour leurs conseils, leur soutien et leur disponibilité tout au long de ce long processus pour la finalisation du projet.

Nous nous en ressortons moins bêtes.

# Table des matières

<b>Introduction</b>	<b>6</b>
<b>1 Cadre théorique et méthodologie</b>	<b>8</b>
<b>1.1 Revue de la littérature . . . . .</b>	<b>8</b>
<b>1.2 Positionnement du projet par rapport à la littérature</b>	<b>11</b>
<b>2 Annotation du corpus de labels d'émotions</b>	<b>14</b>
<b>2.1 Définition des émotions et élaboration du protocole d'annotation . . . . .</b>	<b>14</b>
2.1.1 Définition des émotions et cadre théorique . . . . .	14
2.1.2 Élaboration du protocole d'annotation . . . . .	17
<b>2.2 Procédure d'annotation manuelle . . . . .</b>	<b>19</b>
2.2.1 Méthodologie détaillée . . . . .	19
2.2.2 Mesure de la fiabilité inter-annotateurs (Kappa de Fleiss)	19
2.2.3 Création du fichier annoté final . . . . .	22
<b>2.3 Analyse exploratoire des données . . . . .</b>	<b>24</b>
2.3.1 Répartition émotionnelle selon les contextes . . . . .	24
2.3.2 Typologie des tweets par émotion . . . . .	28
2.3.3 Visualisations : nuages de mots, matrice de confusion, statistiques . . . . .	29
<b>3 Classification automatique des émotions</b>	<b>39</b>
<b>3.1 Préparation des données et modèles . . . . .</b>	<b>39</b>
3.1.1 Démarche méthodologique et traitement du corpus . .	39
3.1.2 Transformation des tweets en données exploitables .	40
3.1.3 Sélection et configuration des modèles . . . . .	42
<b>3.2 Résultats et analyse comparative des modèles . . . . .</b>	<b>43</b>
3.2.1 Méthodologie d'évaluation et critères de performance .	43
3.2.2 Analyse comparative des modèles . . . . .	44

3.2.3	Validation des modèles sur un corpus équilibré généré	51
<b>4</b>	<b>Discussion et perspectives</b>	<b>54</b>
4.1	Limites du projet	54
4.2	Améliorations possibles et perspectives futures	55
	<b>Conclusion</b>	<b>57</b>

# Table des figures

2.1	Exemple d'annotation . . . . .	18
2.2	Répartition des cas résolus lors de l'annotation des tweets . . . . .	23
2.3	Distribution globale des émotions . . . . .	25
2.4	Comparaisons des émotions dans les labels . . . . .	27
2.5	Proportion des tweets scientifiques vs non scientifiques par émotions . . . . .	27
2.6	Wordcloud global sur texte brut . . . . .	29
2.7	Wordcloud global sur texte prétraité . . . . .	30
2.8	Wordcloud sur l'émotion peur . . . . .	30
2.9	Wordcloud sur l'émotion colère . . . . .	31
2.10	Wordcloud sur l'émotion joie . . . . .	31
2.11	Wordcloud sur l'émotion surprise . . . . .	31
2.12	Wordcloud sur l'émotion tristesse . . . . .	32
2.13	Wordcloud sur l'émotion dégout . . . . .	32
2.14	Wordcloud sur l'émotion neutre . . . . .	32
2.15	Matrice de confusion des annotations : Alain vs. Anton . . . . .	33
2.16	Matrice de confusion des annotations : Alain vs. Tony . . . . .	34
2.17	Matrice de confusion des annotations : Anton vs. Tony . . . . .	35
2.18	Distribution de la longueur des tweets selon l'émotion annotée	36
3.1	Treemap du dataset SciTweets . . . . .	41

# Liste des tableaux

1.1	Synthèse des modèles émotionnels et techniques NLP utilisés dans les études analysées . . . . .	11
3.1	Performances des modèles avec et sans rééquilibrage (validation croisée à 5 folds) . . . . .	45
3.2	Meilleurs F1-macro obtenus pour les modèles classiques . . . . .	46
3.3	Meilleurs hyperparamètres identifiés pour chaque modèle . . . . .	46
3.4	Comparaison des performances globales des modèles de classification d'émotions . . . . .	47
3.5	Résultats par émotion pour chaque modèle (précision, rappel, F1-score) . . . . .	48
3.6	Comparaison des performances <i>finetuné</i> entre EmoBERTa et DistilRoBERTa. . . . .	49
3.7	Résumé des paramètres et stratégies de fine-tuning du modèle DistilRoBERTa. . . . .	51
3.8	Accuracy moyenne des modèles sur corpus équilibré et réel . . . . .	52
3.9	Performances des modèles sur corpus réel et équilibré avec interprétation . . . . .	52

# Introduction

L’analyse des émotions dans les échanges en ligne, notamment sur les réseaux sociaux, est devenue un enjeu majeur pour mieux comprendre la circulation de l’information, les raisons pour lesquelles les débats peuvent s’intensifier, et l’impact du numérique sur la perception publique de la science<sup>1</sup>. Jusqu’à présent, les outils informatiques développés pour étudier ces phénomènes se sont souvent focalisés soit sur la détection des émotions, soit sur l’identification des contenus scientifiques, mais rarement sur la combinaison des deux.

Notre projet se distingue par cette approche innovante. Il ne s’agit pas de créer un nouvel outil informatique ni d’améliorer les modèles existants pour détecter les émotions, mais plutôt de proposer un nouveau terrain d’étude scientifique. Nous cherchons à comprendre comment les émotions influencent la réception, le partage et la contestation du discours scientifique sur Internet, en particulier dans un contexte où la circulation des « *fake news* » déstabilise la confiance du public envers la science<sup>2</sup>.

Pour répondre à cette question, nous avons adopté une démarche rigoureuse, en mobilisant des outils informatiques comme instruments d’analyse d’un phénomène social complexe. Plusieurs jeux de données ont été explorés, notamment *SemEval-2018*<sup>3</sup> et *CrowdFlower*<sup>4</sup>, afin de mieux comprendre les approches existantes en matière de détection des émotions. Toutefois, notre projet s’est principalement appuyé sur le jeu de données *SciTweets*<sup>5</sup>, qui nous a semblé particulièrement pertinent à étudier.

---

1. Réseaux sociaux et propagation des émotions : *qu'est-ce que la “conscience collective numérique”*, Science et Vie, consulté le 03 mars 2025.

2. Jukka Jouhki. “*The Future of Journalism : Fake News, Misinformation and Fact-Checking*”. In : *Journalism Practice* 10.7 (2016). Consulté le 04 mars 2025.

3. Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, Svetlana Kiritchenko. *SemEval-2018 Task 1 : Affect in Tweets*. Consulté le 10 février 2025.

4. CrowdFlower. *Emotion in Text Dataset*. Consulté le 10 février 2025.

5. Schellhammer, Sebastian ; Pfeiffer, Jonas ; Gurevych, Iryna. *SciTweets – A Dataset and Annotation Framework for Detecting Scientific Online Discourse*. Consulté le 04 mars 2025.

Notre méthodologie s'est articulée autour de plusieurs étapes clés : une revue de la littérature sur la détection des émotions dans les tweets ; une appropriation du corpus *SciTweets*, suivie d'une annotation manuelle des émotions selon une classification reconnue ; une distinction entre tweets à contenu scientifique et non scientifique ; une analyse statistique des données ; et enfin, l'expérimentation de plusieurs modèles de classification automatique des émotions.

Cette démarche s'inscrit dans une approche de science des données, qui consiste à formuler une question originale, organiser méthodiquement l'étude, puis analyser rigoureusement les résultats. Nous montrons ainsi que les émotions ne sont pas de simples « interférences » à éliminer, mais qu'elles fournissent au contraire des clés précieuses pour mieux comprendre les liens entre science et société.

Partant du postulat que les émotions jouent un rôle déterminant dans la manière dont la science est abordée et discutée sur X (*anciennement Twitter*), nous proposons une première cartographie des émotions associées aux discours à caractère scientifique et non scientifique. Les outils informatiques mobilisés ne constituent que des moyens au service de l'analyse ; notre contribution principale réside dans l'élaboration d'un protocole méthodologique, l'interprétation des premiers résultats, ainsi que la construction d'un corpus annoté de manière croisée, articulant science et émotions, qui servira de base aux analyses futures.

Ce rapport s'articule en quatre grandes parties. La première partie pose le cadre théorique et méthodologique, en présentant d'abord une revue de la littérature existante sur la détection des émotions, puis en positionnant notre projet par rapport à ces travaux. La deuxième partie est consacrée à l'annotation du corpus de labels d'émotions : elle détaille la définition des émotions retenues, l'élaboration du protocole d'annotation, la procédure d'annotation manuelle, ainsi qu'une analyse exploratoire des données annotées. La troisième partie s'intéresse à la modélisation automatique : elle décrit la préparation des données et des modèles, et présente une analyse comparative des performances obtenues par différents algorithmes. Enfin, la quatrième partie propose une discussion critique du projet, en identifiant ses limites, et en ouvrant sur des pistes d'amélioration et des perspectives pour des recherches futures. Le rapport se conclut par une synthèse des apports et des résultats obtenus.

# Chapitre 1

## Cadre théorique et méthodologie

### 1.1 Revue de la littérature

L’analyse des émotions dans les contenus numériques est un champ de recherche en plein essor depuis une dizaine d’années. Les premiers travaux se sont principalement concentrés sur la détection de la polarité (positive, neutre, négative), avant d’évoluer vers des classifications plus fines des émotions, en s’appuyant sur des modèles issus de la psychologie, tels que ceux de **Paul Ekman**<sup>1</sup> ou de **Robert Plutchik**<sup>2</sup>. Si ce champ est aujourd’hui bien établi dans de nombreux contextes notamment commerciaux ou médiatiques, l’étude des émotions dans le cadre des discours scientifiques, et plus particulièrement sur les réseaux sociaux, reste encore largement inexplorée. C’est précisément dans cette perspective que s’inscrit notre projet : analyser les émotions exprimées dans des tweets relatifs à la science publiés sur le réseau social X, en mobilisant des techniques de traitement automatique du langage naturel (NLP, pour *Natural Language Processing*).

Avant de débuter notre projet, nous avons souhaité réaliser une analyse de l’état de l’art. Pour cela, nous avons étudié trois travaux scientifiques qui nous ont permis d’identifier les méthodes d’analyse et les modèles de classification les plus pertinents pour la réussite de notre projet. Cette revue nous a également permis d’examiner les jeux de données utilisés dans ces articles, afin d’en déceler les spécificités et les éventuelles limites.

---

1. Ekman, Paul. *An argument for basic emotions*. Consulté le 05 mars 2025.

2. Plutchik, Robert. *A general psychoevolutionary theory of emotion*. Consulté le 05 mars 2025.

Le premier article, *Sentiment Analysis and Emotion Analysis : A Survey*<sup>3</sup>, distingue clairement deux axes majeurs dans l'analyse du langage : l'analyse des sentiments, qui vise à déterminer la polarité globale d'un texte (positive, négative ou neutre), et l'analyse des émotions, qui cherche à identifier des états émotionnels spécifiques tels que la joie, la colère ou la peur.

L'article met en évidence plusieurs défis méthodologiques importants : la subjectivité de l'annotation émotionnelle, l'ambiguïté inhérente au langage naturel, et la difficulté d'adapter les méthodes à des contextes multilingues. Ces enjeux résonnent particulièrement avec notre projet, qui se concentre sur des tweets rédigés en anglais, mais avec une perspective d'élargissement à d'autres langues dans le futur.

Sur le plan théorique, la revue présente deux grandes familles de modèles émotionnels : les modèles dits *discrets*, comme ceux de Ekman ou Plutchik, qui proposent un nombre limité d'émotions fondamentales ; et les modèles *dimensionnels*, qui positionnent les émotions selon des axes continus tels que la valence (positivité/négativité), l'activation (niveau d'excitation) ou la dominance.

Sur le plan méthodologique, l'article décrit l'usage de lexiques émotionnels, de méthodes d'apprentissage automatique classiques (SVM, Naive Bayes) ainsi que de réseaux de neurones profonds, notamment des modèles de type BERT ou Transformers. Ces apports ont directement influencé notre approche : nous avons choisi de comparer des modèles classiques et avancés afin d'évaluer leur capacité à classifier des tweets courts, informels et souvent bruités. Nous avons également adopté une annotation monolabel, c'est-à-dire qu'un seul type d'émotion est attribué à chaque tweet, afin de garantir la cohérence et la robustesse des résultats.

Enfin, l'article souligne la prépondérance des corpus en anglais et l'absence de prise en compte des éléments visuels (comme les images ou les emojis), des limites que nous avons également rencontrées dans la constitution de notre propre corpus. C'est également dans cet article que sont présentés deux jeux de données de référence très utilisés dans le domaine : **SemEval-2018** et **CrowdFlower**, qui ont servi de base pour de nombreux travaux sur l'analyse des émotions dans les textes en ligne.

Le deuxième article, *The Good, The Bad, and Why : Unveiling Emotions in Generative AI*<sup>4</sup>, propose une perspective innovante sur l'impact des émotions dans les réponses générées par des modèles d'intelligence artificielle comme ChatGPT ou GPT-4. Les auteurs introduisent trois mécanismes originaux : *EmotionPrompt*, qui consiste à insérer des stimuli émotionnels positifs dans les prompts ; *EmotionAttack*, qui simule des attaques

---

3. Evgeny Kim, Roman Klinger. *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies*. Consulté le 05 mars 2025.

4. Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, Xing Xie. *The Good, The Bad, and Why : Unveiling Emotions in Generative AI*. Consulté le 05 mars 2025.

émotionnelles négatives ; et *EmotionDecode*, qui permet d’analyser les représentations internes activées par ces stimuli. Les résultats montrent que ces manipulations ont un impact significatif sur les performances des modèles : les prompts positifs améliorent la qualité des réponses de plus de 13 %, tandis que les prompts négatifs peuvent entraîner une dégradation allant jusqu’à 50 %. L’étude repose sur des tâches variées et une évaluation combinant des métriques automatiques et des validations humaines. Elle met en évidence la forte sensibilité des modèles multimodaux, qui combinent texte et image. Cette approche ouvre des perspectives intéressantes pour notre projet : nous pourrions tester la robustesse de nos modèles face à des tweets ironiques, ambigus ou chargés émotionnellement, mais aussi explorer l’intégration d’éléments visuels tels que les emojis, omniprésents dans les interactions sur les réseaux sociaux.

Le troisième article, *Emotion Analysis in NLP : Trends, Gaps and Roadmap for Future Directions*<sup>5</sup>, dresse un panorama critique des avancées du domaine à partir de l’analyse de 150 publications récentes. Les auteurs mettent en lumière plusieurs limites structurelles : un manque de diversité culturelle et linguistique dans les jeux de données, des catégories émotionnelles parfois inadaptées aux contextes réels, une terminologie hétérogène, et un manque d’interdisciplinarité dans les travaux. Ils soulignent également la domination persistante des modèles discrets (Ekman, Plutchik), mais notent la montée en puissance des modèles de type Transformer (BERT, RoBERTa) ainsi que des approches *zero-shot*, qui permettent de généraliser à de nouveaux domaines sans données annotées. Cette synthèse a nourri notre réflexion sur la nécessité d’enrichir notre corpus dans le futur, en intégrant davantage de langues et de contextes, mais aussi sur l’importance d’une annotation plus fine et mieux encadrée, notamment pour mieux gérer la subjectivité des annotateurs.

---

5. Flor Miriam Plaza-del-Arco, Alba Curry, Amanda Cercas Curry, Dirk Hovy. *Emotion Analysis in NLP : Trends, Gaps and Roadmap for Future Directions*. Consulté le 05 mars 2025.

Étude	Modèles émotionnels	Méthodes et techniques NLP
<i>Sentiment Analysis and Emotion Analysis : A Survey</i>	Modèles discrets (Ekman, Plutchik), Modèles dimensionnels (VAD)	Lexiques émotionnels (WordNet-Affect), Modèles classiques (SVM, Naive Bayes), Deep learning (RNN, LSTM, BERT), Prétraitement (tokenisation, stop-words), Évaluation (précision, rappel, F1-score)
<i>The Good, The Bad, and Why : Unveiling Emotions in Generative AI</i>	Émotions générées par IA (via GPT-4), Pas de modèle psychologique explicite	Mécanismes : EmotionPrompt, EmotionAttack, EmotionDecode, IA génératives (GPT-4, LLaMA2, ChatGPT, GPT-4V, BLIP-2, LLaVa, CogVLM), Analyse multimodale (texte + image), Visualisation de l'attention, Évaluation humaine et automatisée
<i>Emotion Analysis in NLP : Trends, Gaps and Roadmap for Future Directions</i>	Modèles discrets (Ekman, Plutchik), Modèles évaluatifs (Rosman), Dimensionnels (VAD)	Transformers (BERT, RoBERTa), Approches zero-shot, Classification supervisée, Annotation manuelle, Jeux de données variés (EmoEvent, MELD, IEMOCAP), Métriques classiques (F1-score, etc.)

TABLE 1.1 – Synthèse des modèles émotionnels et techniques NLP utilisés dans les études analysées

## 1.2 Positionnement du projet par rapport à la littérature

Notre projet s’inscrit dans le prolongement direct des avancées récentes sur l’analyse des émotions dans les discours scientifiques en ligne, tout en se distinguant par l’exploitation du dataset *SciTweets*, proposé par Hafid et al. Ce choix méthodologique et empirique nous permet de répondre à plusieurs limites identifiées dans la littérature, tout en apportant une contribution originale à ce champ de recherche.

## **Adossement à un dataset de référence : *SciTweets***

Contrairement à de nombreux travaux qui collectent leurs propres corpus ou se fondent sur des jeux de données généralistes, notre projet repose sur le dataset *SciTweets*, publié par Hafid et al. Ce corpus, composé de 1 261 tweets annotés par des experts, offre un cadre rigoureux pour l'étude des discours scientifiques sur Twitter. Il propose une annotation fine de la « scientificité » des tweets, avec un accord inter-annotateurs satisfaisant (coefficient de Fleiss  $\kappa = 0,63$ ), et une structuration multi-label permettant d'identifier différentes formes de contenus scientifiques (affirmations, références, vulgarisation, etc.).

Cette approche répond à un besoin largement souligné dans la littérature récente sur l'analyse des émotions et des discours sociaux : disposer de jeux de données de qualité, annotés selon des critères transparents et reproductibles. L'article de Binali et al. (2023) souligne en effet que le manque de diversité et de rigueur dans les jeux de données constitue un obstacle majeur à la généralisation des modèles d'analyse émotionnelle.

## **Spécificité du contexte : émotion et scientificité**

La littérature met également en lumière le manque de travaux articulant explicitement analyse émotionnelle et analyse de la scientificité dans les discours sociaux. Le dataset *SciTweets* permet précisément cette mise en relation, en autorisant un croisement entre les émotions exprimées (à travers une annotation complémentaire que nous avons réalisée, inspirée du modèle d'Ekman) et le degré d'ancrage scientifique des tweets. Là où la majorité des études abordent séparément la détection des émotions (comme le fait la revue de Yadollahi et al., 2021) et l'identification des contenus scientifiques, notre projet propose une approche conjointe, ouvrant la voie à une meilleure compréhension de l'impact des émotions sur la diffusion et la réception de l'information scientifique en ligne.

## **Méthodologie rigoureuse et reproductible**

En nous appuyant sur *SciTweets*, nous bénéficions d'un protocole d'annotation éprouvé, accompagné d'une documentation détaillée, ce qui renforce la robustesse de notre démarche. Nous avons enrichi ce corpus par une annotation manuelle des émotions selon sept catégories : six émotions fondamentales issues du modèle de Paul Ekman (joie, colère, tristesse, peur, surprise, dégoût) et une classe *Neutre*. Cette annotation a été réalisée selon une procédure itérative et collaborative, visant à maximiser la cohérence entre annotateurs, conformément aux recommandations de la littérature sur les bonnes pratiques en annotation émotionnelle (notamment dans Binali et al., 2023).

La combinaison des annotations de scientifcité et d’émotions constitue une avancée méthodologique rare, dans un champ où la reproductibilité et la transparence restent souvent limitées, comme le soulignent les auteurs de la revue « Emotion Analysis in NLP ».

### **Comparaison de modèles classiques et avancés**

La littérature récente, notamment Yadollahi et al. (2021), insiste sur l’essor des modèles de type *Transformer* (BERT, RoBERTa, EmoBERTa) pour la classification des émotions, tout en rappelant que les modèles plus simples (comme les SVM ou Naive Bayes) peuvent rester pertinents dans le cas de corpus de taille modeste ou spécialisés.

Dans cette continuité, notre projet propose une évaluation comparative rigoureuse de ces deux familles de modèles. Nous portons une attention particulière à l’adaptation des modèles pré-entraînés à un corpus spécialisé comme *SciTweets*, annoté finement. Nous avons également intégré des techniques de rééquilibrage des classes et adopté des métriques appropriées à la distribution déséquilibrée des émotions, en lien avec les recommandations formulées dans l’article « The Good, the Bad, and Why », qui met en évidence la sensibilité des modèles aux variations émotionnelles dans les inputs.

## Chapitre 2

# Annotation du corpus de labels d'émotions

## 2.1 Définition des émotions et élaboration du protocole d'annotation

### 2.1.1 Définition des émotions et cadre théorique

Cette partie se concentre sur la phase cruciale d'annotation manuelle de notre corpus de tweets. Nous détaillerons le choix des émotions, l'élaboration du protocole d'annotation, le processus d'annotation lui-même, l'évaluation de sa cohérence, la création du fichier final et enfin, une première analyse des annotations.

Pour définir les émotions à identifier, notre travail s'appuie directement sur les recherches du psychologue américain **Paul Ekman**. Nous avons repris, les six émotions de base qu'il a identifiées dans ses travaux :

- Peur
- Colère
- Joie
- Surprise
- Tristesse
- Dégout

Ces émotions sont reconnues pour avoir des manifestations distinctes et pour préparer l'individu à réagir face à des situations significatives. À ces six émotions, nous avons ajouté une catégorie "*Neutre*" pour classer les textes sans charge émotionnelle claire, ce qui est fréquent dans la communication. Ce cadre *ekmanien* nous offre une base solide et reconnue pour notre tâche d'annotation.

Afin de permettre un jugement plus uniforme entre les annotateurs, les définitions des émotions ont été rappelées et précisées. Ces définitions ont été formulées et inspirée à partir du dictionnaire *Larousse* afin de garantir une compréhension accessible et partagée. Ce protocole nous permettra d'unifier les définitions de ces émotions, nous permettant ainsi de partir sur une base commune.

**Peur (1)** *Définition* : Trouble émotionnel ressenti en présence ou à la pensée d'un danger, réel ou supposé, qui se manifeste par un sentiment d'insécurité ou d'inquiétude.

*Exemples de notre corpus :*

- Ligne 33 : "Caution! 3D Printers Could Cause Health Problems. Find out more here : <http://t.co/BRAwvRgq28>" (Le mot "Caution!" et "Health Problems" induisent la peur).
- Ligne 84 : ".@GreenJournal study show this infection can increase risk of strokes and heart attacks <http://t.co/OnPKAwpmoh>" (L'annonce d'un risque accru de problèmes de santé graves).

*Contre-exemples (génériques) :*

- Ressentir de l'inquiétude diffuse sans raison précise (plutôt de l'anxiété).
- Être en colère contre un collègue (émotion différente).

**Colère (2)** *Définition* : Émotion vive résultant d'un sentiment de mécontentement, d'une frustration ou d'une injustice, accompagnée souvent d'une réaction violente ou agressive.

*Exemples de notre corpus :*

- Ligne 13 : "@bpkelly12 if that damn book lady decides to stop being a fuckin retard ill be out of here in 2 days" (Usage de termes injurieux et expression d'une forte frustration).
- Ligne 24 : "God dammit ... Poor decisions lead to negative consequences. Won't happen again" (Interjection exprimant la colère et le regret).

*Contre-exemples (génériques) :*

- Être triste après une mauvaise nouvelle (pas de colère).
- Ressentir de la peur face à une menace (autre émotion).

**Joie (3)** *Définition* : Émotion agréable et profonde, sentiment de bonheur, de plaisir, de satisfaction intense, généralement causé par un événement positif ou la réalisation d'un désir.

*Exemples de notre corpus :*

- Ligne 0 : "Knees are a bit sore. i guess that's a sign that my recent treadmilling is working" (Satisfaction d'un résultat positif).

- Ligne 1 : "McDonald's breakfast stop then the gym " (Expression d'un moment agréable et d'enthousiasme).

*Contre-exemples (génériques) :*

- Être soulagé après un danger (plutôt du soulagement).
- Être envieux du succès d'autrui (pas de joie).

**Surprise (4)** *Définition :* Émotion provoquée par quelque chose d'inattendu, de soudain, qui étonne ou déconcerte momentanément.

*Exemples de notre corpus :*

- Ligne 36 : "RT @rgj : Barrick Gold reports \$8.56 billion Q2 loss http://t.co/fuaTuAH54i" (Annonce d'une perte financière massive et inattendue).
- Ligne 78 : "A lucky find of a toe fossil unlocks a genetic encyclopedia for neanderthals http://t.co/gfjo79RuSv" (Découverte inattendue et significative).

*Contre-exemples (génériques) :*

- S'attendre à un résultat et ne pas être étonné (pas de surprise).
- Être triste à l'annonce d'une mauvaise nouvelle attendue (pas de surprise).

**Tristesse (5)** *Définition :* État affectif pénible, marqué par le chagrin, la peine, la mélancolie ou la douleur morale, souvent en réaction à une perte ou une déception.

*Exemples de notre corpus :*

- Ligne 11 : "@HankAzaria @TheSimpsons oh loneliness and cheeseburgers are a dangerous mix" (Expression de la solitude).
- Ligne 19 : "About to just go to bed maybe cry a little and maybe stop breathing okay? Okay." (Pensées sombres et évocation de pleurs).

*Contre-exemples (génériques) :*

- Être en colère après une injustice (autre émotion).
- Ressentir de la joie lors d'un succès (opposé).

**Dégoût (6)** *Définition* : Réaction de répulsion ou d'aversion, souvent accompagnée d'une sensation physique désagréable, provoquée par quelque chose de répugnant, d'inacceptable ou de moralement choquant.

*Exemples de notre corpus :*

- Ligne 3 : "Couch-lock highs lead to sleeping in the couch. Gotta stop doing this shit." (Le terme "shit" exprime un rejet et un dégoût de la situation).
- Ligne 14 : "Science sucks." (Expression directe d'aversion).

*Contre-exemples (génériques) :*

- Être en colère face à une injustice (pas du dégoût).
- Ressentir de la peur devant un animal dangereux (autre émotion).

**Neutre (7)** *Définition* : Utilisé pour les cas où aucune des six émotions principales n'est clairement identifiable ou prédominante dans le texte. Le message est factuel, informatif sans charge émotionnelle marquée.

*Exemples de notre corpus :*

- Ligne 2 : "Can any Gynecologist with Cancer Experience explain the dangers of Transvaginal Douching with Fluoride or other toxins such as Dioxin ? PDX" (Question factuelle cherchant une information).
- Ligne 4 : "Does daily routine help prevent problems with bipolar disorder http://t.co/XGUfUDoLJB" (Question informative sur un sujet médical).

### 2.1.2 Élaboration du protocole d'annotation

Dès les premières phases d'annotation, nous avons constaté des divergences d'interprétation entre annotateurs, notamment pour certaines émotions perçues différemment selon les individus. Ces écarts posaient un risque pour la cohérence du corpus final et ont rapidement mis en évidence le besoin d'un cadre plus structuré.

Bien que quelques règles aient été évoquées oralement en début de projet, aucune consigne n'avait été formalisée. C'est à la suite des premières évaluations de cohérence (*cf* 2.2.2) que la nécessité de rédiger un protocole d'annotation s'est imposée. Ce document visait à harmoniser nos pratiques et à garantir un processus reproductible.

Le protocole avait plusieurs fonctions : rappeler les émotions à identifier, préciser les critères permettant de les reconnaître dans les tweets, et proposer une méthode simple pour les annoter.

L'annotation était réalisée dans un tableur partagé, où chaque ligne représentait un tweet. Pour chacun, nous indiquions l'émotion perçue (*ou deux selon le tweet*) ainsi qu'un niveau de confiance (*faible, moyen ou fort*). Le tableau comprenait également des données issues du jeu de données *SciTweets*, comme l'identifiant du tweet et son contenu textuel.

Ce travail de formalisation a permis de mieux structurer notre protocole et d'améliorer progressivement l'homogénéité des annotations au sein du groupe.

<b>text</b>	<b>annotateur_1 émotions</b>	<b>confiance</b>	<b>annotateur_2 émotions</b>	<b>confiance</b>
McDonald's breakfast stop then the gym 🍔💪	3	low	4	medium

FIGURE 2.1 – Exemple d'annotation

Certains tweets étaient difficiles à interpréter, notamment ceux qui se contentaient de partager le titre d'un article scientifique. Après discussion, nous avons décidé que les tweets ayant comme corps de message, le titre de l'article (*Exemple : "Versatile microRNAs Choke Off Cancer Blood Supply, Suppress > Metastasis - Science Daily (press release) http://t.co/.. "*) pouvait contenir une charge émotionnelle implicite, même si le tweet ne contenait pas d'opinion ou de commentaire. Par exemple, un titre alarmant ou sensationnaliste pouvait être associé à la peur ou à la surprise.

Nous avons aussi envisagé d'annoter deux émotions par tweet, mais cette option a été écartée. Dans la majorité des cas, une seule émotion dominait, et ajouter une seconde compliquait inutilement l'analyse. Nous avons donc choisi une annotation dite « mono-label » (*une seule émotion par tweet*), ce qui a simplifié notre travail et renforcé la cohérence entre annotateurs.

Enfin, plusieurs sessions de test ont été réalisées pour améliorer le protocole. Ces tests nous ont permis d'ajuster les consignes, de mieux comprendre les cas ambigus, et d'obtenir une meilleure cohérence entre les différentes personnes qui annotaient. Une fois le protocole finalisé, nous avons appliqué les nouvelles règles à l'ensemble des tweets pour créer une version annotée propre et fiable, utilisable pour l'entraînement des modèles d'analyse.

## 2.2 Procédure d'annotation manuelle

### 2.2.1 Méthodologie détaillée

L'annotation a été réalisée par un groupe de trois évaluateurs principaux (**Alain, Anton, Tony**) sur un corpus de 1140 tweets. Chaque évaluateur devait initialement attribuer jusqu'à deux émotions par tweet. L'idée était d'être trois pour avoir un vote majoritaire, de commencer par un échantillon (*par exemple 30 tweets*) pour évaluer la concordance, ajuster le protocole si besoin, puis diviser le reste du travail.

### 2.2.2 Mesure de la fiabilité inter-annotateurs (Kappa de Fleiss)

Afin d'évaluer la cohérence de notre annotation manuelle, nous avons calculé le **Kappa de Fleiss**<sup>1</sup>, un indicateur statistique couramment utilisé pour mesurer l'accord entre plusieurs annotateurs. Contrairement à des mesures simples comme le pourcentage d'accord, ce coefficient tient compte du hasard et fournit ainsi une estimation plus fiable de la qualité de l'annotation.

Le calcul du coefficient de Kappa a été réalisé sur la base des annotations effectuées en parallèle, avant toute discussion collective ou harmonisation. Le coefficient de Kappa (*Cohen's Kappa*) est une mesure statistique qui évalue le degré d'accord entre deux annotateurs, en tenant compte de l'accord attendu par hasard. Il varie entre  $-1$  (désaccord total) et  $1$  (accord parfait), une valeur de  $0$  indiquant un accord équivalent à celui obtenu par le hasard.

Le calcul du Kappa a été réalisé sur la base des annotations effectuées en parallèle, avant toute discussion collective ou harmonisation.

Cette mesure nous a permis d'identifier les points de divergence, d'affiner nos consignes et de renforcer la cohérence globale des annotations. Une fois le protocole stabilisé, nous avons utilisé ces résultats pour consolider un jeu de données final, prêt à être exploité pour l'entraînement et l'évaluation des modèles de classification.

### Méthodologie

**Préparation des données :** Nous avons enregistré les annotations dans un fichier CSV avec 15 colonnes, incluant le texte des tweets et les annotations des évaluateurs (par exemple, Alain\_émotions, Alain\_émotions 2, etc.). Un problème de décalage dans les colonnes a été identifié, causé par des virgules et guillemets dans le texte des tweets, perturbant le parsing du fichier. Ce problème a été résolu en remplaçant les virgules par des points (seulement pour le calcul) et un script de nettoyage pour s'assurer que chaque ligne avait exactement 15 colonnes.

---

1. Fleiss, Joseph L. *Measuring nominal scale agreement among many raters*. Consulté le 02 mai 2025.

**Calcul du Kappa de Fleiss :** Nous avons calculé le Kappa de Fleiss sous trois approches principales :

- **Version 1 : Toutes émotions (6 votes par tweet)** : Chaque évaluateur pouvait donner jusqu'à 2 votes par tweet, soit un maximum de 6 votes par tweet (3 évaluateurs  $\times$  2 émotions). Cependant, les colonnes pour l'émotion 2 étaient souvent vides, ce qui rendait le nombre de votes inconstant.
- **Version 2 : Un vote par évaluateur (Émotion 1 uniquement)** : Chaque évaluateur donnait un seul vote (émotion 1), soit exactement 3 votes par tweet.
- **Version optimisée : Maximiser l'accord avec toutes les émotions** : Pour chaque évaluateur, nous avons choisi entre son émotion 1 et son émotion 2 en fonction de celle qui est en meilleur accord avec toutes les émotions des autres évaluateurs (c'est-à-dire émotion 1 et émotion 2 d'Anton et Tony pour Alain, par exemple). Cela prend en compte jusqu'à 6 votes par tweet pour maximiser le consensus.

Le calcul a été effectué avec Python, en utilisant les bibliothèques pandas et statsmodels.

## Résultats

- **Version 1 : Toutes émotions (6 votes par tweet)** : Aucun tweet n'avait exactement 6 votes, car les colonnes pour l'émotion 2 étaient souvent vides (par exemple, Alain\_émotions 2 manquante dans le tweet 0). Résultat : 1140 tweets exclus, aucun Kappa calculé. Cette approche n'était pas viable étant donné la structure des données.
- **Version 2 : Un vote par évaluateur (Émotion 1 uniquement)** : En prenant un vote par évaluateur (l'émotion 1, 3 votes par tweet), le Kappa de Fleiss obtenu est 0.3083. Selon les seuils de Landis & Koch (1977), cela indique un accord léger (0.21–0.40).

- **Version optimisée : Maximiser l'accord avec toutes les émotions** : En comparant chaque émotion (1 et 2) d'un évaluateur avec toutes les émotions (1 et 2) des autres évaluateurs, le Kappa de Fleiss obtenu est 0.3758. Cela représente une amélioration par rapport au Kappa initial de 0.3083, montrant que l'émotion 2, lorsqu'elle est présente, peut refléter un consensus dans certains cas où l'émotion 1 diverge. Selon les seuils de Landis & Koch, ce Kappa indique toujours un accord léger (0.21–0.40), mais il se rapproche de la limite de l'accord modéré (0.41–0.60).

**Discussion sur le Kappa** Suite au calcul du Kappa de Fleiss, notamment avec l'approche optimisée, nous avons obtenu un score de 0,3758. Selon l'échelle d'interprétation de **Landis & Koch (1977)**<sup>2</sup>, cela correspond à un “*accord passable/léger*”. Dans la plupart des domaines, un accord “modéré” (supérieur à 0,40) est généralement recherché, et un bon score serait idéalement supérieur à 0,60. Ce faible score reflète en évidence un manque de cohérence et de la compléxité de la tâche dans nos annotations initiales. Plusieurs facteurs peuvent l'expliquer : une certaine ambiguïté dans les consignes (avant la formalisation complète du protocole), le contenu même des tweets, la barrière de la langue (l'anglais n'étant pas notre point fort), ou encore nos interprétations personnelles des émotions.

Malgré ce faible score, il n'est pas catastrophique, et en restant optimistes, nous pouvons considérer qu'il nous reste une bonne marge de progression pour améliorer notre cohérence, ce qui a été l'objectif des étapes suivantes. Le Kappa initial de 0.3083 (émotion 1 uniquement) indique un accord léger, suggérant que les évaluateurs divergent souvent. L'approche optimisée (0.3758) a montré que l'émotion 2 contient des informations utiles.

## Recommandations et prochaines étapes

À la suite du calcul du Kappa de Fleiss, nous avons organisé une discussion collective entre les annotateurs afin de clarifier les définitions des émotions et d'examiner les tweets pour lesquels il n'y avait pas de consensus clair. Cette phase d'échange a permis d'aligner nos interprétations, en particulier sur les cas ambigus ou sujets à confusion.

Sur la base de ces ajustements, un sous-ensemble de tweets a été réannoté avec les nouvelles consignes. Cette révision a contribué à améliorer la cohérence globale de l'annotation et à valider notre protocole. Le fichier final d'annotations a ainsi été produit à l'issue de ce processus, garantissant une base de données plus fiable pour l'entraînement des modèles.

Pour les prochaines itérations, nous recommandons de formaliser encore davantage les cas d'ambiguïté dans le protocole d'annotation, et d'envisager

---

2. Landis, J. Richard; Koch, Gary G. *The measurement of observer agreement for categorical data*. Consulté le 22 mai 2025.

l'intégration d'une phase de validation croisée entre annotateurs sur de petits lots avant de généraliser le travail à l'ensemble du corpus.

### 2.2.3 Creation du fichier annote final

Cette tape visait  consolider les annotations ralises sur notre corpus de 1 140 tweets en une seule tiquette motionnelle par tweet, afin de produire un jeu de donnees coherent et directement exploitable pour l'entranement des modles.

**Objectif** Une fois les annotations individuelles recueillies, notre priorite a te de garantir l'uniformite des tiquettes finales tout en respectant le plus possible les choix des valuateurs. L'enjeu tait de traiter les cas de desaccord de maniere rigoureuse et transparente, sans reintroduire de biais, et en s'appuyant sur des regles claires de fusion.

**Methode de consolidation** La creation du fichier final s'est deroulee en plusieurs tapes successives, combinant criteres d'accord majoritaire, analyse manuelle et recours  des outils externes lorsque nessaire.

- **1. Accord majoritaire entre valuateurs** : Pour chaque tweet, nous avons d'abord recherche un consensus simple. Si au moins deux annotateurs avaient indique la meme motion principale, celle-ci tait retenue. Cette regle a permis de statuer sur la majorite des cas (environ 80 % du corpus).
- **2. Recours aux emotions secondaires** : Pour les tweets sans majorite claire, nous avons elargi l'analyse en integrant les emotions secondaires renseignes au moment de l'annotation. Lorsque l'une des emotions apparaissait plusieurs fois dans l'ensemble des votes disponibles, elle tait selectionnee comme tiquette finale.
- **3. Reintegration d'une annotation externe** : Un quatrieme valuateur, ayant quitte le projet en cours, avait annote l'ensemble du corpus. Ses annotations ont te temporairement reintegrees pour tenter de degager un consensus dans les cas restants. Cette tape a permis de resoudre une partie supplementaire des tweets non encore attribues.
- **4. Revue finale assistee par IA** : Pour les derniers tweets restant sans consensus, nous avons organise une relecture collective. Afin d'objectiver nos decisions, nous avons utilise deux modles d'intelligence artificielle (Grok et ChatGPT) pour proposer une tiquette motionnelle par tweet. Leurs suggestions ont servi de point de depart pour nos echanges, jusqu'a l'obtention d'un accord.

## Résultats de la création du fichier final

- **79,6 %** (908 tweets) résolus par accord entre au moins deux des trois annotateurs (émotion principale identique), dont **28,7 %** (327 tweets) avec accord complet entre les trois.
- **8,1 %** (92 tweets) résolus en combinant les émotions principales et secondaires renseignées lors de l'annotation.
- **8,4 %** (96 tweets) résolus grâce à la réintégration ponctuelle des annotations d'un quatrième évaluateur.
- **3,9 %** (44 tweets) tranchés lors d'une relecture finale, assistée par les suggestions de Grok et ChatGPT.

**Apport des modèles IA :** L'intervention de Grok et ChatGPT a offert un regard complémentaire sur les cas les plus ambigus. Leurs suggestions ont facilité les discussions et contribué à objectiver la décision finale.

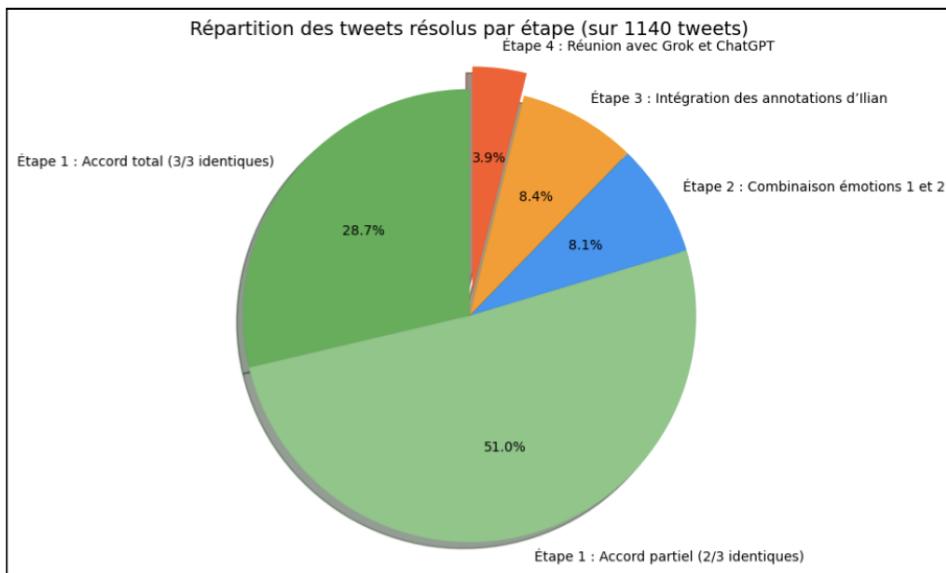


FIGURE 2.2 – Répartition des cas résolus lors de l'annotation des tweets

**Discussion** L'adoption d'une annotation unique par tweet s'est révélée pertinente. Elle a simplifié l'unification des votes, réduit les ambiguïtés et permis de structurer un fichier final plus adapté à l'entraînement de modèles d'apprentissage automatique. Bien que non conservée dans la version finale, l'émotion secondaire a joué un rôle clé dans la résolution de certains désaccords entre annotateurs.

Le traitement automatisé des cas les plus simples a montré une bonne convergence initiale entre évaluateurs, confirmée par les résultats du Kappa de Fleiss. Pour les cas plus complexes, la réintégration ciblée d'annotations supplémentaires et l'usage ponctuel de suggestions issues de modèles d'in-

telligence artificielle se sont avérés efficaces. Quelques tweets ont toutefois nécessité une interprétation collective approfondie, en raison d'un manque de contexte ou d'une formulation particulièrement vague.

**Résultat** Ce processus progressif nous a permis d'attribuer une émotion principale unique à chacun des 1 140 tweets. Le fichier ainsi obtenu constitue une base fiable et harmonisée, conforme à notre protocole, et prête à être utilisée pour l'entraînement et l'évaluation des modèles de classification émotionnelle développés dans la suite du projet.

## 2.3 Analyse exploratoire des données

Une fois le fichier final d'annotations constitué, nous avons procédé à une analyse descriptive pour mieux comprendre la nature de notre corpus. Cette étape est essentielle pour appréhender la répartition des émotions, identifier les tendances et préparer l'interprétation des performances des modèles.

### 2.3.1 Répartition émotionnelle selon les contextes

Notre corpus final, issu du fichier `Annotation_apres_mise_en_accord_a_utiliser_pour_entrainement_modele.tsv`, contient une colonne `science_related` (avec des valeurs 0 ou 1) permettant de distinguer les tweets à caractère scientifique de ceux qui ne le sont pas.

**Distribution Globale des Émotions** Sur l'ensemble des 1140 tweets annotés, la distribution des émotions est la suivante :

- **Neutre** : 484 tweets (42.46%)
- **Joie** : 218 tweets (19.12%)
- **Colère** : 184 tweets (16.14%)
- **Surprise** : 91 tweets (7.98%)
- **Tristesse** : 65 tweets (5.70%)
- **Peur** : 52 tweets (4.56%)
- **Dégoût** : 46 tweets (4.04%)

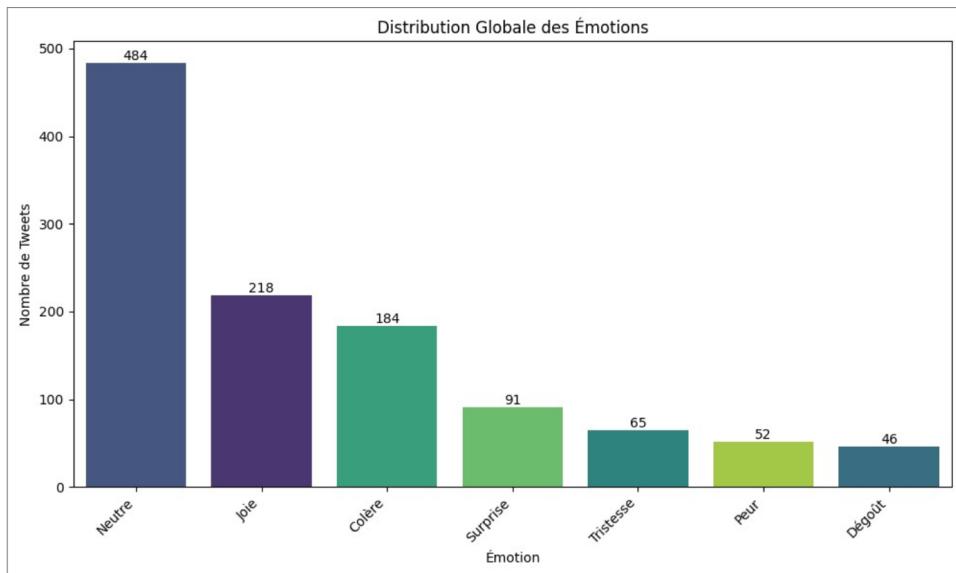


FIGURE 2.3 – Distribution globale des émotions

Comme attendu, l'émotion "*Neutre*" est la plus représentée, ce qui est typique des corpus de communication textuelle. Les émotions "*Joie*" et "*Colère*" sont également significativement présentes. Les émotions "*Tristesse*", "*Peur*" et "*Dégoût*" sont les moins fréquentes, ce qui souligne un déséquilibre naturel des classes. Ce déséquilibre sera une considération importante pour l'entraînement de nos modèles, justifiant potentiellement l'utilisation de techniques de rééquilibrage.

**Analyse Contextuelle : Scientifique vs. Non-Scientifique** Nous avons ensuite analysé la distribution des émotions en séparant les tweets en deux catégories, sur la base de la colonne `science_related` :

- Tweets scientifiques (`science_related == 1`) : 375 tweets.
- Tweets non-scientifiques (`science_related == 0`) : 765 tweets.

#### **Distribution dans les Tweets Scientifiques :**

- Neutre : 168 tweets (44.80%)
- Joie : 61 tweets (16.27%)
- Surprise : 51 tweets (13.60%)
- Peur : 33 tweets (8.80%)
- Colère : 28 tweets (7.47%)
- Tristesse : 24 tweets (6.40%)
- Dégoût : 10 tweets (2.67%)

**Distribution dans les Tweets Non-Scientifiques :**

- Neutre : 316 tweets (41.31%)
- Joie : 157 tweets (20.52%)
- Colère : 156 tweets (20.39%)
- Tristesse : 41 tweets (5.36%)
- Surprise : 40 tweets (5.23%)
- Dégoût : 36 tweets (4.71%)
- Peur : 19 tweets (2.48%)

L'analyse contextuelle révèle des tendances distinctes. Dans les tweets scientifiques, la prédominance de l'émotion "Neutre" est encore plus marquée. Notablement, la "Peur" (8.80% vs 2.48%) et la "Surprise" (13.60% vs 5.23%) y sont proportionnellement plus présentes que dans les tweets non-scientifiques, ce qui peut s'expliquer par des tweets alertant sur des risques ou annonçant des découvertes. Inversement, les tweets non-scientifiques affichent une proportion plus élevée de "Colère" (20.39% vs 7.47%) et de "Joie" (20.52% vs 16.27%).

Pour mieux visualiser quelle proportion de chaque émotion appartient au contexte scientifique ou non-scientifique, le graphique suivant est présenté :

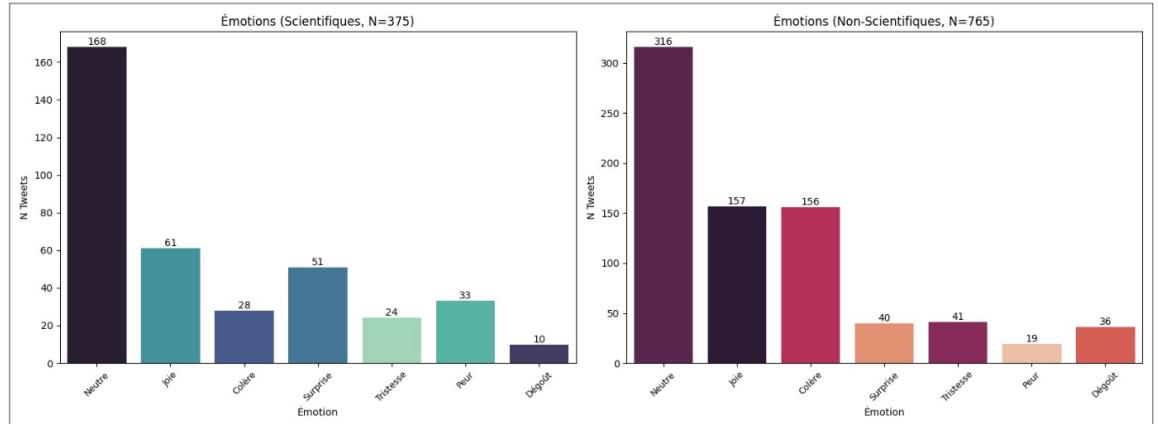


FIGURE 2.4 – Comparaisons des émotions dans les labels

Ce graphique montre clairement que :

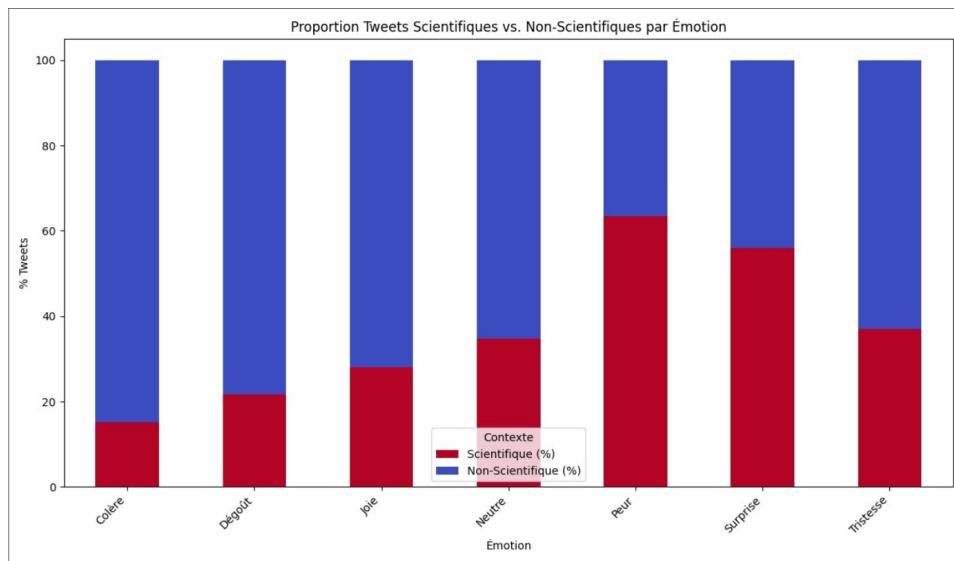


FIGURE 2.5 – Proportion des tweets scientifiques vs non scientifiques par émotions

- La Peur est majoritairement exprimée dans un contexte scientifique (63.5% des tweets de peur sont scientifiques).
- La Surprise est également plus fréquente dans les tweets scientifiques (56.0% des tweets de surprise).
- À l'inverse, la Colère (84.8% non-scientifique), la Joie (72.0% non-

- scientifique) et le Dégoût (78.3% non-scientifique) sont bien plus pré-dominants dans les tweets non-scientifiques.
- La Tristesse (63.2% non-scientifique) et Neutre (65.3% non-scientifique) sont aussi majoritairement non-scientifiques, mais avec une part scientifique plus notable pour "Neutre".

### 2.3.2 Typologie des tweets par émotion

Une analyse qualitative des tweets associés à chaque émotion permet de mieux cerner les thématiques et les formulations typiques.

**Peur :** Souvent liée à des avertissements, des dangers potentiels ou des situations anxiogènes. *Exemple* : "Caution! 3D Printers Could Cause Health Problems..." *Exemple* : "...infection can increase risk of strokes and heart attacks..." *Exemple* : "BLACK KPOP STANS FEEL UNSAFE AND SCARED..."

**Colère :** Exprime fréquemment la frustration, l'indignation, ou une critique véhémente. *Exemple* : "Even Superman had to get a job. Moral of this story is Stop saving everyone for free..." *Exemple* : "Bernie needs to change his strategy... if he doesn't go full gloves off..."

**Joie :** Associée à des succès, des événements positifs, des expressions de soutien ou d'enthousiasme. *Exemple* : "Aegon UK reports rise in earnings..." *Exemple* : "...plant temperaments - how vulnerable they are, but also how strong they can become."

**Surprise :** Déclenchée par des nouvelles inattendues, des découvertes ou des événements sortant de l'ordinaire. *Exemple* : "Going vegetarian 'could save lives and the planet'..." (l'information peut surprendre) *Exemple* : "FBI agents issue statement of support for FBI director..." (un événement potentiellement inattendu)

**Tristesse :** Manifeste la perte, le désespoir, la solitude ou la référence à des événements malheureux. *Exemple* : "Armed conflict disproportionately affects children..." *Exemple* : "@HankAzaria @TheSimpsons oh loneliness and cheeseburgers are a dangerous mix..."

**Dégoût :** Indique un rejet, une aversion ou un jugement moral négatif. *Exemple* : "stupid people in VT support trump..." *Exemple* : "...buy more healthy and less unhealthy food..." (implicitement, un dégoût pour la nourriture malsaine)

**Neutre** : Concerne principalement des messages factuels, des annonces ou des questions informatives. *Exemple* : "#ProjectXVegeta Supports this with immense description." *Exemple* : "This #ImplementationScience article describes an intervention program..."

### 2.3.3 Visualisations : nuages de mots, matrice de confusion, statistiques

Pour visualiser les termes les plus saillants, des nuages de mots (*word-clouds*) ont été générés.

**Wordcloud Global (Texte Original Brut) :** Ce nuage montre les mots les plus fréquents avant tout traitement, incluant potentiellement des éléments comme des URLs (ex : "http", "co", "t").



FIGURE 2.6 – Wordcloud global sur texte brut

**Wordcloud Global (Texte Nettoyé, Stopwords Appliqués) :** Après un premier nettoyage et la suppression des stopwords anglais courants, ce nuage donne une meilleure idée des thèmes sémantiques dominants. Les logs indiquent que les mots les plus fréquents sont : "stop" (192 occurrences), "support" (167), "report" (75), "people" (74), et "new" (69).



FIGURE 2.7 – Wordcloud global sur texte prétraité

**Wordclouds par Émotion (sur texte processé) :** Ces nuages de mots mettent en évidence les vocabulaires spécifiques à chaque émotion. D'après l'analyse TF-IDF des mots les plus discriminants (voir ci-dessous), nous pouvons anticiper certains de ces termes :

- Pour la **Peur** : des termes comme "cause", "report", "risk", "infection", "prevent".
  - Pour la **Colère** : des mots comme "stop", "people", "like", "support", "trump".
  - Pour la **Joie** : "support", "stop", "great", "love", "new".
  - Pour la **Surprise** : "lead", "study", "new", "cancer", "arrest".
  - Pour la **Tristesse** : "cause", "death", "report", "cancer", "stop".
  - Pour le **Dégoût** : "stop", "people", "support", "got", "high".
  - Pour **Neutre** : "support", "stop", "report", "science", "increase", "new".



FIGURE 2.8 – Wordcloud sur l'émotion peur

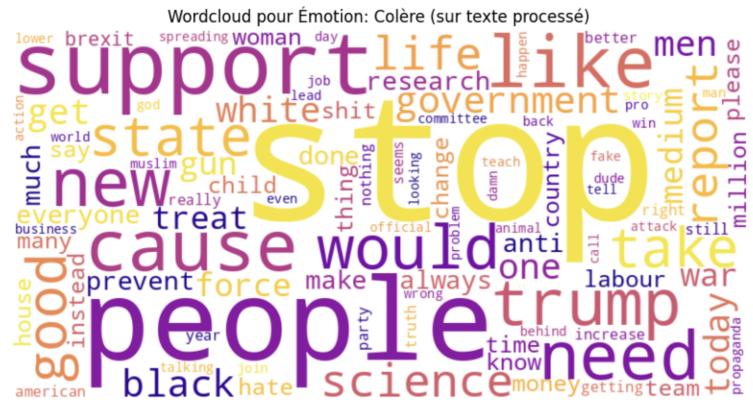


FIGURE 2.9 – Wordcloud sur l'émotion colère



FIGURE 2.10 – Wordcloud sur l’émotion joie

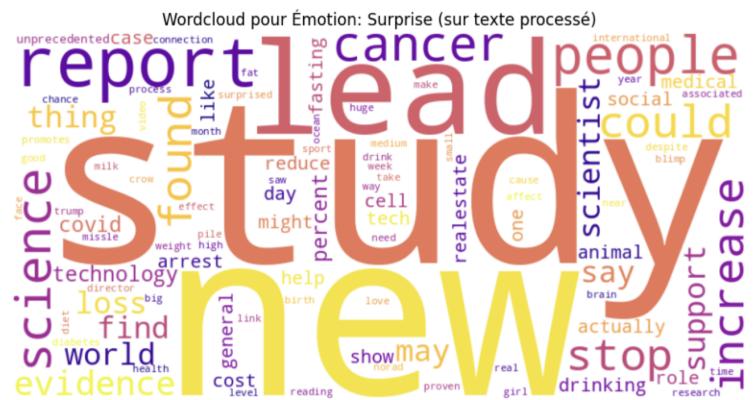


FIGURE 2.11 – Wordcloud sur l’émotion surprise



FIGURE 2.12 – Wordcloud sur l’émotion tristesse



FIGURE 2.13 – Wordcloud sur l’émotion dégoût



FIGURE 2.14 – Wordcloud sur l’émotion neutre

**Matrice de confusion inter-annotateurs** Afin d’analyser les divergences entre annotateurs avant la fusion finale des étiquettes, nous avons générée des matrices de confusion à partir du fichier `Annotation_sans_mise_en_accord_avec_Ilian.csv`. Chaque matrice compare les émotions principales (émotion 1) attribuées par deux annotateurs parmi les trois membres du groupe (Alain, Anton, Tony). Pour faciliter l’analyse, les labels ont été convertis de 1–7 à 0–6 conformément à notre `label_map_analysis`, soit : 0 = Peur, 1 = Colère, 2 = Joie, 3 = Surprise, 4 = Tristesse, 5 = Dégoût, 6 = Neutre.

**Comparaison Alain vs. Anton** L’accord le plus net entre Alain et Anton concerne l’émotion *Neutre* (297 annotations communes), suivie de *Colère* (93), *Joie* (83) et *Surprise* (59). Certaines divergences notables apparaissent toutefois : des tweets annotés *Peur* par Alain ont été considérés *Neutres* par Anton (12 cas). De même, les tweets jugés *Surprise* par Alain sont fréquemment annotés comme *Neutre* (119), *Joie* (33) ou *Colère* (16) par Anton. Enfin, l’émotion *Colère* selon Alain est parfois interprétée comme *Dégoût* (21) ou *Neutre* (35) par Anton.

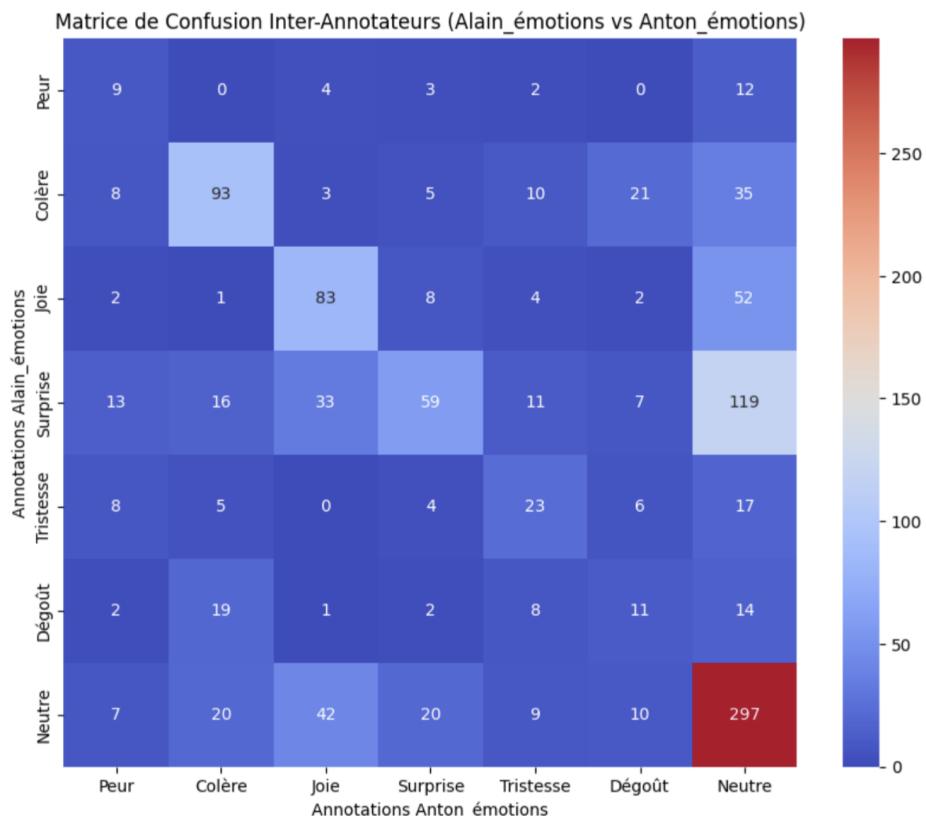


FIGURE 2.15 – Matrice de confusion des annotations : Alain vs. Anton

**Comparaison Alain vs. Tony** Les convergences les plus fortes entre Alain et Tony concernent *Neutre* (174), *Joie* (115) et *Colère* (104). Des écarts importants sont observés sur l’émotion *Surprise*, que Tony interprète souvent comme *Joie* (98), *Neutre* (50), voire *Peur* (36). On observe également une confusion partielle entre *Colère* et *Dégoût* (19 occurrences).

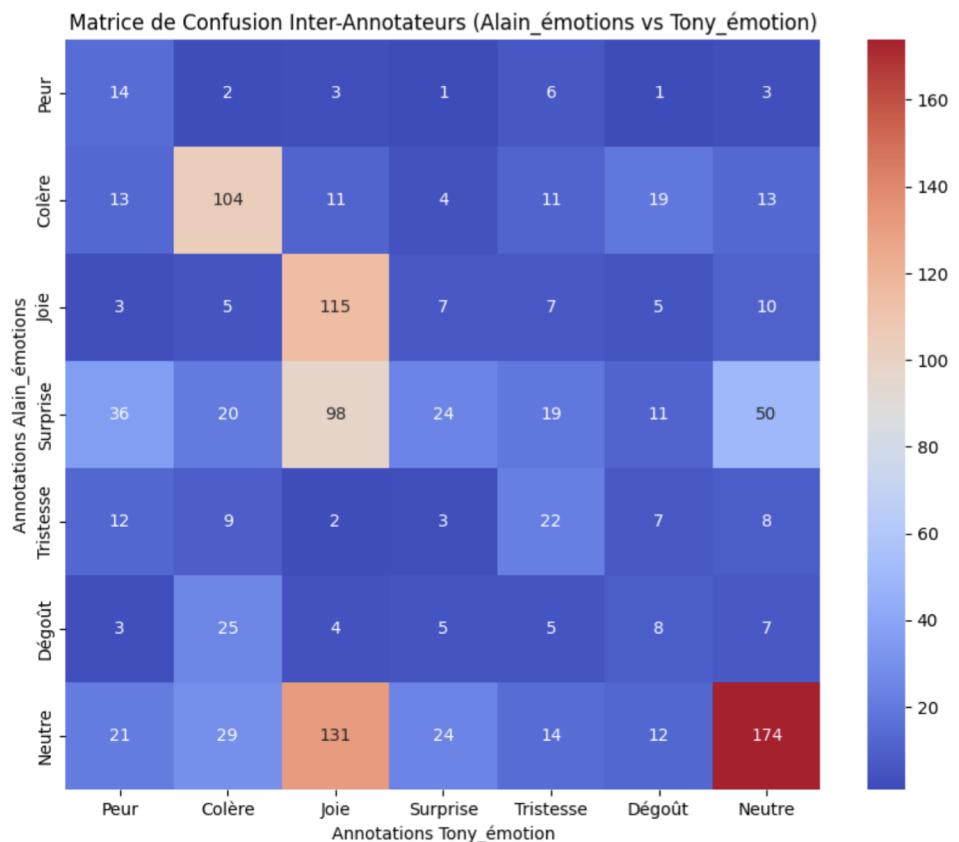


FIGURE 2.16 – Matrice de confusion des annotations : Alain vs. Tony

**Comparaison Anton vs. Tony** Anton et Tony présentent un bon niveau d'accord sur les émotions *Neutre* (210), *Joie* (139) et *Colère* (93). Toutefois, l’émotion *Neutre* attribuée par Anton est parfois perçue différemment par Tony, notamment comme *Joie* (160) ou *Colère* (48). Le *Dégoût* selon Anton est également souvent interprété comme *Colère* par Tony (30 cas).

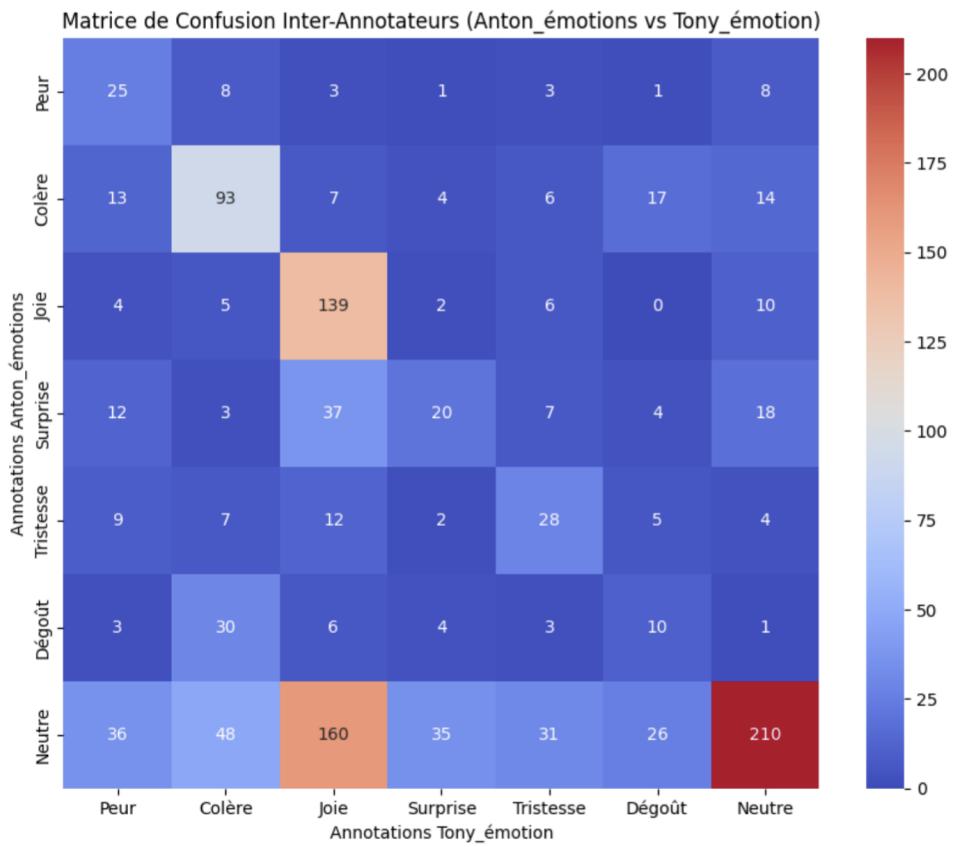


FIGURE 2.17 – Matrice de confusion des annotations : Anton vs. Tony

**Analyse des matrices de confusion** L'examen des matrices de confusion met en évidence plusieurs tendances. L'émotion *Neutre* est celle pour laquelle l'accord est le plus fréquent entre annotateurs, suivie par *Colère* et *Joie*. Les désaccords les plus marqués concernent principalement l'émotion *Surprise*, souvent confondue avec *Neutre* ou *Joie*, ce qui illustre la difficulté à interpréter l'effet de surprise dans des textes courts. Des confusions récurrentes apparaissent également entre *Colère* et *Dégoût*, ou entre *Colère* et *Neutre*, suggérant une proximité sémantique entre ces catégories dans certains contextes.

Ces résultats sont cohérents avec le coefficient de Fleiss obtenu précédemment ( $Kappa = 0,38$ ), indiquant un accord jugé modéré. Ils confirment la nature subjective de la tâche d'annotation émotionnelle, et soulignent la complexité à distinguer certaines émotions lorsqu'elles sont exprimées de façon implicite ou condensée, comme c'est souvent le cas sur les réseaux sociaux. Ces matrices ont constitué un support précieux lors de nos discussions pour la consolidation du fichier final.

**Statistiques descriptives complémentaires** L'analyse de la longueur des tweets selon l'émotion annotée révèle également des tendances intéressantes, aussi bien en nombre de caractères qu'en nombre de mots.

#### Longueur moyenne des tweets (en caractères)

- Peur : 151,2
- Colère : 148,3
- Joie : 147,5
- Neutre : 135,5
- Tristesse : 131,4
- Surprise : 129,3
- Dégoût : 117,5

#### Longueur moyenne des tweets (en mots)

- Colère : 22,1
- Joie : 21,1
- Tristesse : 20,6
- Peur : 19,9
- Neutre : 18,2
- Surprise : 18,0
- Dégoût : 17,3

Ces données suggèrent que certaines émotions, comme la *Colère*, la *Joie* ou la *Peur*, sont en moyenne exprimées dans des tweets plus longs, tant en nombre de mots qu'en caractères. À l'inverse, des émotions comme le *Dégoût*, ou les énoncés *Neutres*, tendent à être formulés de manière plus concise.

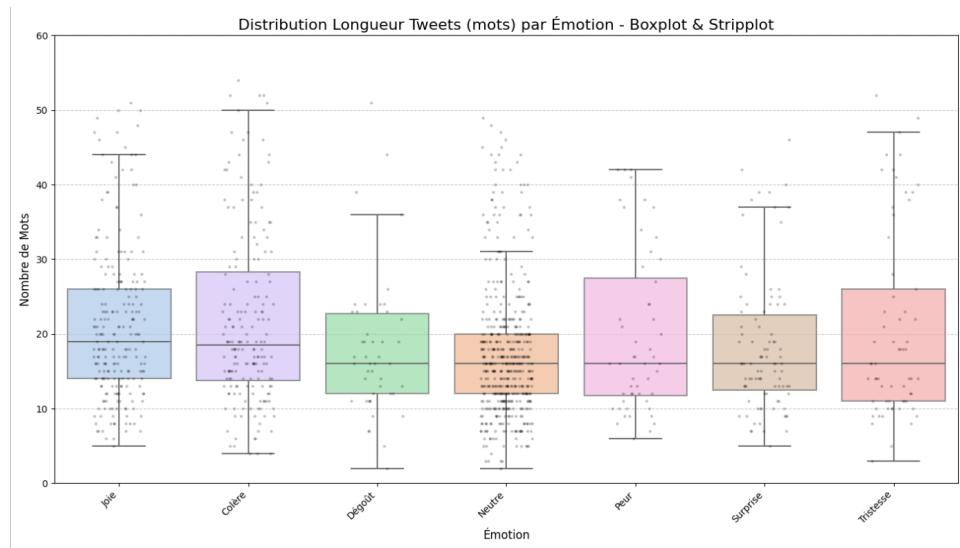


FIGURE 2.18 – Distribution de la longueur des tweets selon l'émotion annotée

Le boxplot présenté ci-dessus illustre non seulement ces moyennes, mais également la variabilité interne à chaque catégorie. Certains cas atypiques (outliers) indiquent qu'un même type d'émotion peut parfois être exprimé très brièvement ou, au contraire, de manière plus développée. Cela pourrait refléter une nécessité de contextualisation plus grande pour certaines émotions, comme la *Colère* ou la *Joie*, alors que d'autres, comme le *Dégoût*, peuvent être formulées de manière plus directe et brève.

**Mots les plus discriminants par émotion (basés sur le score TF-IDF moyen)** Une analyse exploratoire des scores moyens de TF-IDF (Term Frequency–Inverse Document Frequency) permet d'identifier les mots les plus caractéristiques de chaque émotion. Ce type d'analyse met en évidence les termes dont la fréquence est particulièrement marquée dans une classe émotionnelle donnée, tout en étant moins fréquente dans les autres.

- **Peur** : *cause* (0,0547), *report* (0,0497), *risk* (0,0387), *infection* (0,0353), *prevent* (0,0324)
- **Colère** : *stop* (0,1089), *people* (0,0326), *like* (0,0258), *support* (0,0254), *trump* (0,0192)
- **Joie** : *support* (0,0510), *stop* (0,0345), *great* (0,0211), *love* (0,0196), *new* (0,0189)
- **Surprise** : *lead* (0,0475), *study* (0,0338), *new* (0,0290), *cancer* (0,0238), *arrest* (0,0224)
- **Tristesse** : *death* (0,0432), *cause* (0,0378), *report* (0,0364), *cancer* (0,0337), *stop* (0,0317)
- **Dégoût** : *stop* (0,0844), *people* (0,0578), *support* (0,0402), *got* (0,0308), *high* (0,0164)
- **Neutre** : *support* (0,0390), *stop* (0,0269), *report* (0,0233), *science* (0,0211), *increase* (0,0163)

Ces résultats confirment plusieurs observations qualitatives. Le mot *stop* apparaît fortement dans les émotions *Colère* et *Dégoût*, traduisant une volonté de rejet ou d'opposition. À l'inverse, des mots comme *risk*, *infection* ou *death* sont étroitement associés aux émotions *Peur* et *Tristesse*, soulignant leur charge émotionnelle forte. Le mot *support*, présent dans plusieurs classes, illustre quant à lui la dépendance au contexte pour déterminer le ton émotionnel d'un tweet.

**Conclusion de la phase d'annotation et d'analyse du corpus** La constitution d'un corpus annoté de qualité a constitué une étape centrale dans notre démarche, visant à explorer la détection automatique des émotions dans des contenus courts issus des réseaux sociaux.

Ce travail s'est appuyé sur un cadre théorique robuste (modèle d'Ekman), un protocole d'annotation itératif, et un protocole rigoureux pour garantir la cohérence des étiquetages. L'évaluation initiale via le Kappa de Fleiss

a révélé un accord modéré entre annotateurs (0,3758), mettant en lumière la subjectivité propre à ce type de tâche, mais aussi les défis posés par la brièveté et le contexte réduit des tweets.

La création du fichier final, à travers un processus en plusieurs étapes, a permis d'obtenir une annotation consolidée pour 1140 tweets. Cette base fiable a ensuite été explorée de manière descriptive. Les émotions *Neutre* (42,5 %), *Joie* (19,1 %) et *Colère* (16,1 %) sont les plus représentées, tandis que d'autres comme *Dégoût* ou *Surprise* restent plus marginales.

Les analyses menées révèlent également des différences marquées selon le type de tweet : les tweets à caractère scientifique tendent à être plus *neutres*, mais expriment davantage de *peur* ou de *surprise*, souvent en lien avec des découvertes ou alertes. Les tweets non scientifiques présentent une part plus importante de *colère* ou de *joie*, reflétant un style plus expressif ou engagé.

L'examen des longueurs de tweets et des termes spécifiques à chaque émotion a permis d'approfondir cette caractérisation. Les matrices de confusion entre annotateurs, quant à elles, ont mis en lumière les zones les plus sujettes à interprétation, notamment entre émotions sémantiquement proches comme *Surprise*, *Joie* et *Neutre*.

Malgré les difficultés liées à la nature du support étudié, cette phase a abouti à un corpus robuste, annoté de manière fiable et bien documentée. Il constitue désormais un socle essentiel pour les expérimentations de la partie suivante, consacrée à l'évaluation et à l'optimisation de modèles de classification émotionnelle, notamment à travers des architectures fondées sur les Transformers.

# Chapitre 3

## Classification automatique des émotions

### 3.1 Préparation des données et modèles

#### 3.1.1 Démarche méthodologique et traitement du corpus

La réussite de notre projet repose sur le choix du **modèle de classification le plus pertinent**. Pour y parvenir, nous avons adopté une démarche rigoureuse, consistant à tester plusieurs méthodes différentes, puis à comparer leurs performances afin de retenir celle qui donne les meilleurs résultats.

*Mais qu'est-ce qu'un modèle de classification ?* Il s'agit d'un outil informatique capable de ranger automatiquement des données dans des *catégories*. Par exemple, un modèle peut analyser un message et décider s'il s'agit d'un tweet offensant ou non, simplement en étudiant les mots et expressions qu'il contient.

Il existe une *grande variété de modèles de classification*, chacun avec ses propres avantages. Parmi les plus connus, on retrouve :

- la régression logistique,
- le K-plus proches voisins (KNN),
- les machines à vecteurs de support (SVM).

Ces modèles sont dits **linéaires**, c'est-à-dire qu'ils sont efficaces lorsque les différences entre les catégories peuvent être décrites par des règles assez simples.

D'autres modèles, dits **non linéaires**, peuvent gérer des situations plus complexes. Il s'agit par exemple :

- des arbres de décision,
- des forêts aléatoires,
- des réseaux de neurones, capables d'identifier des schémas plus subtils dans les données.

Dans la suite de ce chapitre, nous présentons ces modèles, expliquons leur fonctionnement de manière accessible, et montrons comment nous les avons testés sur notre *corpus de tweets annotés*. Enfin, nous comparerons leurs résultats pour déterminer lequel est le plus adapté à notre problématique.

### 3.1.2 Transformation des tweets en données exploitable

Avant de comparer les modèles, il est essentiel de **préparer correctement les données**, une étape appelée *prétraitement*. Cette phase est cruciale, car une mauvaise préparation peut fausser totalement les résultats.

**Rééquilibrage des classes** Un problème fréquent est le **déséquilibre des classes**. Par exemple, si 90 % des tweets sont “neutres” et seulement 10 % sont “offensants”, un modèle pourrait apprendre à prédire uniquement la classe majoritaire sans réelle intelligence.

Pour corriger cela, nous avons utilisé des techniques de rééquilibrage :

- **Oversampling** : augmentation artificielle du nombre d’exemples dans la classe minoritaire,
- **Undersampling** : réduction du nombre d’exemples dans la classe majoritaire.

**Spécificités du langage sur X (Twitter)** Comme mentionné dans la section 0.3.3, notre jeu de données contient des tweets, un format de texte très particulier. On y trouve notamment :

- des liens (URLs),
- des hashtags (#),
- des mentions (@),
- des emojis,
- un usage souvent non standard de la langue (abréviations, fautes, etc.).

Pour mieux comprendre la composition de notre corpus, nous avons utilisé une visualisation appelée **Treemap**, représentant les éléments les plus fréquents du corpus sous forme de rectangles, dont la taille indique leur importance.

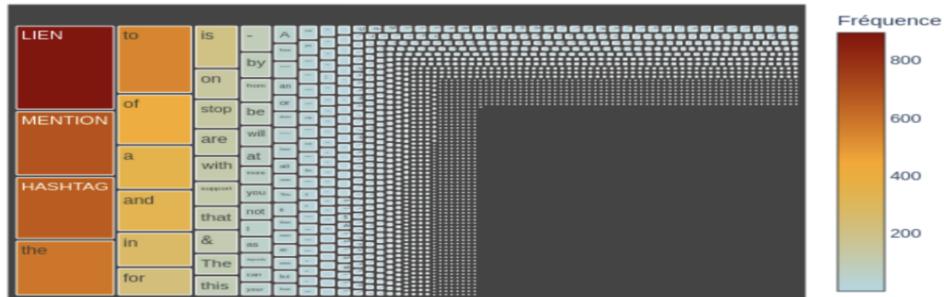


FIGURE 3.1 – Treemap du dataset SciTweets

L’analyse du Treemap a révélé que les hashtags, liens, emojis et stop-words (mots très fréquents comme *the*, *and*, *is*) sont très présents dans les tweets. Ces éléments peuvent parasiter l’analyse, et il est donc pertinent de les supprimer ou de les traiter spécifiquement.

#### Étapes de nettoyage

Nous avons testé différentes configurations de nettoyage et de transformation des tweets, parmi lesquelles :

- suppression des URLs, mentions, hashtags, emojis,
- mise en minuscules,
- suppression des stopwords,
- lemmatisation (réduction des mots à leur forme de base, ex. : *running* → *run*),
- stemmatisation (réduction à la racine du mot, ex. : *learning* → *learn*).

Dans une des configurations, nous avons également expérimenté la transcription des emojis en texte (par exemple, l’emoji smiley jaune a été converti en son équivalent textuel *smiling face*), afin d’en conserver la valeur émotionnelle. Toutefois, cette approche a conduit à des résultats moins satisfaisants en classification, probablement en raison d’une perte de structure sémantique ou d’un vocabulaire trop artificiel pour les modèles. Cette expérimentation nous a conduits à privilégier la suppression pure et simple des emojis dans le jeu de données final.

Les modèles de *machine learning* ne traitant que des valeurs numériques, il est nécessaire de transformer les mots en vecteurs numériques à l’aide d’un processus appelé vectorisation.

Nous avons choisi deux méthodes efficaces et adaptées aux textes courts comme les tweets :

- **CountVectorizer** : compte le nombre d’occurrences de chaque mot dans le texte,
- **TF-IDF** (Term Frequency - Inverse Document Frequency) : pondère la fréquence d’un mot selon sa rareté dans le corpus, ce qui permet de réduire le poids des mots trop fréquents.

**Utilisation des n-grams** Pour mieux capturer le contexte des mots, nous avons utilisé les **n-grams**, c'est-à-dire des groupes de  $n$  mots consécutifs :

- un **unigramme** : un seul mot (ex. *happy*),
- un **bigramme** : deux mots (ex. *very happy*),
- un **trigramme** : trois mots (ex. *not very happy*).

Cette technique permet de repérer des expressions ou associations fréquentes de mots, ce qui peut améliorer les performances des modèles.

Nous avons testé différentes combinaisons :

- unigrammes seuls,
- unigrammes + bigrammes,
- et dans certains cas, jusqu'aux trigrammes.

### 3.1.3 Sélection et configuration des modèles

Afin d'identifier le modèle le plus adapté à notre tâche, nous avons testé plusieurs **modèles issus de la littérature**, repérés lors de notre revue des travaux similaires (section 1).

Cela nous a permis de retenir un ensemble de modèles complémentaires, incluant :

#### Modèles classiques

- Naives Bayes,
- SVM (Support Vector Machine).

Dans le cadre d'un projet universitaire portant sur une problématique similaire, nous avons choisi d'importer et de tester un troisième modèle de classification : la **régression logistique (Logistic Regression)**.

Ces modèles sont *rapides à entraîner* et donnent de bons résultats avec une bonne vectorisation.

**Modèles de type transformer** Nous avons également testé des modèles plus récents, basés sur **BERT**, un modèle de traitement du langage pré-entraîné sur de grandes quantités de textes. Plus précisément, nous avons utilisé :

- **DistilRoBERTa** : une version allégée de RoBERTa, plus rapide mais très performante,
- **EmoRoBERTa** : spécialisée dans la détection des émotions,
- **EmoBERTa** : fine-tunée sur des corpus émotionnels, idéale pour détecter des tonalités ou nuances dans les tweets.

Ces modèles modernes permettent une *compréhension plus fine du langage*, en prenant en compte le contexte autour des mots.

**Hyperparamètres et fine-tuning** Pour chaque modèle, une attention particulière a été portée à la sélection et à l'ajustement des hyperparamètres, tels que le taux d'apprentissage, la taille des *batchs*, le nombre d'époques ou encore les coefficients de régularisation. Ces paramètres influencent directement la qualité de l'apprentissage et les performances finales du modèle.

Afin de garantir une sélection rigoureuse, nous avons utilisé la méthode **GridSearch**, qui permet d'explorer de manière exhaustive différentes combinaisons de valeurs d'hyperparamètres. Cette recherche est couplée à une validation croisée afin de choisir les configurations les plus performantes tout en limitant le risque de surapprentissage (*overfitting*).

Concernant les modèles de type *Transformer*, un *fine-tuning* a été réalisé sur notre corpus spécifique de tweets. Cette étape consiste à réentraîner les modèles pré-entraînés (comme BERT) sur nos propres données annotées, afin d'adapter leurs représentations linguistiques à notre tâche de classification d'émotions. Le *fine-tuning* devrait nous permettre d'atteindre des performances supérieures à celles obtenues avec des représentations génériques non spécialisées.

Nous avons par ailleurs évalué chaque modèle avec différentes configurations de données et de vectorisation, dans le but d'identifier celui qui s'adapte le mieux à notre problématique.

## 3.2 Résultats et analyse comparative des modèles

### 3.2.1 Méthodologie d'évaluation et critères de performance

Pour évaluer de manière rigoureuse les performances des modèles testés, nous avons adopté la méthode de validation croisée k-fold. Le jeu de données est automatiquement divisé en  $k$  sous-ensembles (ou *folds*) ; à chaque itération, un *fold* différent est utilisé comme jeu de test, les autres servant à l'entraînement. Cette approche permet à chaque observation de contribuer à la fois à l'entraînement et à l'évaluation, assurant ainsi une estimation plus fiable des performances et limitant les biais liés à une répartition arbitraire des données. Dans notre cas, nous avons utilisé une validation croisée à cinq *folds*.

Comme évoqué précédemment, notre corpus présente un fort déséquilibre entre les classes, avec une surreprésentation de l'émotion «Neutre». Ce déséquilibre peut fausser les résultats, car un modèle pourrait obtenir de bons scores en se contentant de prédire la classe majoritaire. Pour y remédier, nous avons intégré des techniques de sur-échantillonnage dans nos pipelines de modélisation. Deux méthodes principales ont été explorées :

- *RandomOverSampler*, qui duplique aléatoirement des exemples de la classe minoritaire ;

- *SMOTE* (*Synthetic Minority Over-sampling Technique*), qui génère des exemples synthétiques à partir d'instances existantes de cette même classe.

Ces techniques sont appliquées uniquement sur les ensembles d'entraînement, à chaque itération de la validation croisée, afin de préserver l'intégrité des évaluations et d'éviter toute fuite d'information vers les jeux de test.

Pour interpréter les performances, plusieurs métriques ont été mobilisées : l'accuracy, la précision, le rappel et le score F1. Étant donné le déséquilibre important entre les classes, nous nous sommes principalement appuyés sur le **score F1-macro**, qui calcule la moyenne des F1-scores obtenus pour chaque classe, sans tenir compte de leur fréquence. Cela permet d'évaluer plus équitablement les performances du modèle, même sur les classes minoritaires. Pour rappel :

- **L'accuracy** mesure la proportion de prédictions correctes par rapport à l'ensemble des prédictions. Bien qu'intuitive, cette métrique peut être trompeuse lorsque les classes sont déséquilibrées, car un modèle pourrait obtenir une bonne accuracy simplement en prédisant la classe majoritaire.
- **La précision** indique, parmi les prédictions faites pour une classe donnée, combien sont réellement correctes. Elle est utile pour évaluer le taux de faux positifs, c'est-à-dire les cas où le modèle prévoit une classe alors qu'elle est incorrecte.
- **Le rappel** (ou sensibilité) mesure la capacité du modèle à identifier correctement tous les exemples d'une classe, ce qui permet de réduire les faux négatifs (les cas oubliés).
- **Le score F1**, et plus particulièrement ici le **F1-macro**, est la moyenne harmonique entre précision et rappel, calculée classe par classe et moyennée sans pondération. Il offre une vision plus équilibrée des performances que l'accuracy seule.

### 3.2.2 Analyse comparative des modèles

Dans cette section, nous comparons les performances des différents modèles de classification testés dans le cadre de notre étude. L'objectif est de mettre en évidence leurs forces et leurs limites respectives dans le contexte de l'analyse d'émotions à partir de texte, en tenant compte du déséquilibre important des classes au sein du corpus.

#### Approches basées sur les modèles classiques

Nous examinons tout d'abord les résultats des modèles classiques — *Naive Bayes*, *SVM linéaire* et *Logistic Regression* — évalués avec et sans techniques de rééquilibrage. Par la suite, une comparaison sera faite avec les modèles à base de transformateurs pré-entraînés, tels que *DistilRoBERTa*,

*EmoRoBERTa* et *EmoBERTa*, afin d'évaluer leur capacité à dépasser les approches traditionnelles. L'évaluation comparative des modèles Naive Bayes, SVM linéaire et Logistic Regression révèle des écarts notables. Naive Bayes, bien que relativement stable (42,6% d'Accuracy), affiche une précision et un F1-score très faibles (respectivement 25,1% et 25,7%), signe d'une forte dépendance à la classe dominante. En revanche, le SVM linéaire obtient les meilleurs résultats globaux, avec 48,2% d'Accuracy et un F1-score de 41,4%, indiquant une meilleure discrimination inter-classes. Logistic Regression se situe entre les deux, avec 45,3% d'Accuracy, mais un F1-score plus modeste (31,9%).

Modèle	Rééquilibrage	Accuracy ( $\pm$ )	F1-score ( $\pm$ )
Naive Bayes	Aucun	$0.426 \pm 0.004$	$0.257 \pm 0.008$
SVM linéaire	Aucun	$0.482 \pm 0.026$	<b><math>0.414 \pm 0.029</math></b>
Logistic Regression	Aucun	$0.453 \pm 0.015$	$0.319 \pm 0.022$
Naive Bayes	SMOTE	$0.277 \pm 0.023$	$0.282 \pm 0.036$
SVM linéaire	SMOTE	$0.448 \pm 0.026$	$0.409 \pm 0.023$
Logistic Regression	SMOTE	$0.464 \pm 0.005$	<b><math>0.420 \pm 0.010</math></b>
Naive Bayes	RandomOverSampler	$0.275 \pm 0.019$	$0.288 \pm 0.023$
SVM linéaire	RandomOverSampler	$0.453 \pm 0.026$	$0.413 \pm 0.024$
Logistic Regression	RandomOverSampler	$0.459 \pm 0.019$	<b><math>0.414 \pm 0.016</math></b>

TABLE 3.1 – Performances des modèles avec et sans rééquilibrage (validation croisée à 5 folds)

L'intégration des techniques de rééquilibrage a permis d'améliorer les performances, notamment pour les modèles linéaires. Avec *SMOTE*, *Logistic Regression* atteint 46,4% d'Accuracy et un score F1 de 42,0%, dépassant légèrement le *SVM* (44,8% et 40,9%). Les résultats obtenus avec *RandomOverSampler* sont comparables : 45,9% d'Accuracy pour *Logistic Regression* contre 45,3% pour le *SVM*. En revanche, *Naive Bayes* ne tire aucun bénéfice notable du sur-échantillonnage, ses performances restant faibles (environ 27% d'Accuracy et 28% de score F1). Ces résultats confirment l'intérêt du rééquilibrage dans un contexte de classes déséquilibrées et positionnent *Logistic Regression* comme le modèle le plus performant parmi ceux testés après application de ces techniques.

Pour aller plus loin dans l'optimisation, une **recherche des hyperparamètres** a été réalisée à l'aide de la méthode **GridSearch**, combinée à une validation croisée à cinq folds. Cette approche systématique a permis de tester plusieurs combinaisons de paramètres et d'identifier la configuration optimale pour chaque modèle. Au total, 240 configurations ont été évaluées pour Naive Bayes, et 40 configurations pour chacun des modèles SVM et Logistic Regression.

Les meilleurs résultats obtenus (en score F1-macro) sont les suivants :

Modèle	F1-macro
Naive Bayes	0,357
SVM linéaire	0,415
Régression logistique	0,415

TABLE 3.2 – Meilleurs F1-macro obtenus pour les modèles classiques

Nous pouvons constater que le score F1-Macro pour *Naive Bayes* a nettement été amélioré, passant de 0,288 à 0,357, grâce à l'utilisation des hyperparamètres. En revanche, pour les deux autres modèles, les changements sont très peu perceptibles.

Les meilleurs hyperparamètres identifiés pour chaque modèle sont :

Modèle	Meilleurs hyperparamètres
Naive Bayes	alpha = 0,1, fit_prior = True, ngram_range = (1,1), max_df = 0,9, min_df = 1
SVM linéaire	C = 1, ngram_range = (1,1)
Régression logistique	C = 1, ngram_range = (1,1)

TABLE 3.3 – Meilleurs hyperparamètres identifiés pour chaque modèle

Cette étape de réglage fin a permis de valider la robustesse des modèles linéaires, en particulier **Logistic Regression**, dans le cadre de cette tâche de *classification émotionnelle de tweets*.

### Approches basées sur les transformateurs

Au-delà des modèles classiques de machine learning, nous avons évalué des architectures plus avancées reposant sur des transformateurs pré-entraînés, spécifiquement conçues ou adaptées pour la classification d'émotions. Ces modèles bénéficient d'un apprentissage préalable sur de larges corpus textuels, ce qui leur permet de mieux capturer la sémantique et la structure du langage naturel.

Dans cette étude, nous avons testé trois variantes : *DistilRoBERTa*, une version allégée et optimisée de RoBERTa, ainsi que *EmoRoBERTa* et *EmoBERTa*, deux modèles spécialisés dans la détection d'émotions. Ces modèles ont été fine-tunés sur notre corpus pour permettre une comparaison équitable avec les approches traditionnelles. Nous présentons ci-après les performances obtenues et discutons leur apport en termes de précision et de généralisation.

Les trois modèles BERT que nous avons testés reposent sur une même base technique, mais présentent des différences importantes dans leur conception et leur usage prévu.

*DistilRoBERTa* est une version allégée du modèle RoBERTa : il est plus rapide et moins coûteux à utiliser. Pour notre projet, nous utiliserons sa version *finetunée*, à savoir *j-hartmann/ emotion-english-distilroberta-base*. Ce modèle a également été entraîné sur des jeux de données comportant les mêmes 7 émotions que notre dataset.

*EmoRoBERTa*, quant à lui, a été entraîné sur des textes contenant également ces mêmes 7 émotions. Cela lui permet de mieux identifier les sentiments exprimés dans une phrase.

*EmoBERTa* va encore plus loin : il a été entraîné sur un plus grand nombre de textes et dans plusieurs langues, ce qui peut le rendre plus robuste face à des situations variées.

Modèle	Précision	Rappel	F1-score	Accuracy
EmoRoBERTa	0.4491	0.4667	0.4110	0.4667
DistilRoBERTa	0.4976	0.4781	0.4752	0.4781
EmoBERTa	<b>0.5581</b>	<b>0.5719</b>	<b>0.5235</b>	<b>0.5719</b>

TABLE 3.4 – Comparaison des performances globales des modèles de classification d’émotions

Les résultats obtenus mettent en évidence des différences notables entre les trois modèles testés. **EmoBERTa** obtient les meilleures performances globales, notamment grâce à sa capacité à reconnaître correctement différentes émotions, ce qui montre qu’un entraînement sur un grand volume de données variées peut améliorer la qualité des prédictions. **DistilRoBERTa**, bien qu’il n’ait pas été conçu spécifiquement pour l’analyse des émotions, son modèle *finetuné* fournit des résultats corrects et relativement équilibrés, tout en étant plus léger et rapide à utiliser.

En revanche, **EmoRoBERTa** obtient des résultats plus modestes. Cela peut s’expliquer par le fait qu’il a été entraîné sur un jeu de données contenant 28 émotions différentes, bien plus nombreuses que les 7 catégories que nous avons retenues pour notre projet. Afin de rendre les résultats comparables, nous avons donc regroupé les émotions fines d’EmoRoBERTa dans ces 7 grandes catégories. Ce regroupement, appelé *mapping*, a été réalisé manuellement à l’aide d’un dictionnaire de synonymes en français, en associant chaque émotion spécifique à l’une des émotions principales (par exemple, « tristesse », « chagrin » et « mélancolie » ont été rassemblées sous « sadness »). Bien que cette méthode permette une comparaison entre les modèles, elle peut aussi entraîner une perte de précision, en particulier pour les émotions plus rares ou plus nuancées, qui se retrouvent fondues dans des catégories plus générales.

Émotion	Modèle	Précision	Rappel	F1-score
Fear	EmoRoBERTa	0,17	0,04	0,06
	DistilRoBERTa	0,26	0,38	0,31
	EmoBERTa	0,00	0,00	0,00
Anger	EmoRoBERTa	0,63	0,18	0,29
	DistilRoBERTa	0,59	0,43	0,50
	EmoBERTa	0,62	0,67	0,64
Joy	EmoRoBERTa	0,49	0,38	0,42
	DistilRoBERTa	0,60	0,37	0,46
	EmoBERTa	0,56	0,53	0,54
Surprise	EmoRoBERTa	0,15	0,11	0,13
	DistilRoBERTa	0,21	0,21	0,21
	EmoBERTa	0,83	0,05	0,10
Sadness	EmoRoBERTa	0,50	0,14	0,22
	DistilRoBERTa	0,19	0,37	0,25
	EmoBERTa	0,44	0,31	0,36
Disgust	EmoRoBERTa	0,00	0,00	0,00
	DistilRoBERTa	0,18	0,04	0,07
	EmoBERTa	0,43	0,07	0,11
Neutral	EmoRoBERTa	0,48	0,82	0,61
	DistilRoBERTa	0,57	0,66	0,61
	EmoBERTa	0,57	0,80	0,67

TABLE 3.5 – Résultats par émotion pour chaque modèle (précision, rappel, F1-score)

Les résultats obtenus montrent des différences notables entre les performances des trois modèles sur les différentes émotions. **EmoBERTa** se distingue particulièrement sur les émotions *anger* et *neutral*, atteignant respectivement des F1-scores de 0,64 et 0,67, ce qui témoigne de sa capacité à identifier efficacement ces émotions. **DistilRoBERTa**, bien que plus léger, offre un bon compromis entre précision et rappel, notamment sur les émotions *fear* ( $F1 = 0,31$ ) et *joy* ( $F1 = 0,46$ ).

En revanche, **EmoRoBERTa** présente des résultats très variables selon les émotions. Il parvient à bien détecter *neutral* ( $\text{rappel} = 0,82$ ), mais échoue complètement sur *disgust* ( $F1 = 0,00$ ) et reste très faible sur *fear* ( $F1 = 0,06$ ). De plus, un problème de *mapping* incorrect des étiquettes émotionnelles a été identifié lors du fine-tuning d'EmoRoBERTa, faussant une partie des résultats. En conséquence, ce modèle a été écarté de la suite de l'étude.

Nous avons ainsi choisi de nous concentrer uniquement sur **EmoBERTa** et **DistilRoBERTa**, qui, au-delà de leurs bonnes performances respectives, sont également plus adaptés à notre projet. Leur comportement est plus cohérent, et leur capacité à traiter les émotions majoritaires comme minori-

taires, malgré les déséquilibres dans les données, en fait des candidats plus pertinents pour le fine-tuning final.

Par ailleurs, l'analyse des performances montre une forte corrélation entre le *support* (nombre d'exemples par classe) et les scores obtenus, notamment en F1-score. Les émotions les mieux reconnues sont celles disposant d'un plus grand nombre d'exemples, telles que *neutral* (484 exemples, F1 = 0,67), *anger* (184 exemples, F1 = 0,64) ou *joy* (218 exemples, F1 = 0,54). À l'inverse, les classes les moins représentées, comme *fear* (52 exemples, F1 = 0,00), *disgust* (46 exemples, F1 = 0,11) ou *sadness* (65 exemples, F1 = 0,36), obtiennent des performances nettement inférieures. Ce déséquilibre impacte la capacité des modèles à généraliser sur les émotions rares, qui tendent à être confondues avec des classes plus fréquentes.

### Sélection du modèle optimal

À la suite de l'analyse comparative détaillée dans cette section, la sélection des modèles les plus adaptés à notre tâche de classification des émotions sur des tweets s'est précisée.

Les modèles classiques, bien qu'améliorés par le rééquilibrage (notamment *SMOTE* pour la Régession Logistique) et l'optimisation des hyperparamètres, ont montré leurs limites face à la complexité linguistique et émotionnelle des messages sur les réseaux sociaux.

Les modèles basés sur l'architecture *Transformer*, et en particulier **EmoBERTa** et **DistilRoBERTa**, ont démontré une capacité supérieure à généraliser et à capturer les nuances sémantiques du langage des tweets. Ce constat se traduit par des performances nettement supérieures en termes de F1-score, précision et rappel.

En revanche, **EmoRoBERTa** a été écarté de la suite de l'étude en raison d'un problème identifié lors du fine-tuning : un *mapping* incorrect des étiquettes émotionnelles, faussant les résultats obtenus et rendant les comparaisons non pertinentes.

Nous avons donc retenu **EmoBERTa** et **DistilRoBERTa** comme modèles de référence pour la suite du projet. Si DistilRoBERTa obtient de meilleures performances globales, EmoBERTa conserve un avantage sur certaines émotions spécifiques, notamment *anger* et *neutral*, ce qui justifie leur complémentarité dans l'évaluation finale.

Modèle	Accuracy	F1-score	Précision	Rappel
EmoBERTa	0,5482	0,5321	0,5188	0,5482
DistilRoBERTa	<b>0,5789</b>	<b>0,5592</b>	<b>0,5524</b>	<b>0,5789</b>

TABLE 3.6 – Comparaison des performances *finetuné* entre EmoBERTa et DistilRoBERTa.

Les résultats obtenus à partir des meilleurs modèles entraînés, tels que DistilRoBERTa et EmoBERTa, montrent des performances moyennes, avec une précision globale qui reste inférieure à 60%. Bien que ces modèles soient réputés dans le domaine de l'analyse des émotions, ils peinent à identifier certaines, notamment celles qui sont peu fréquentes ou exprimées de manière implicite.

Ces limites peuvent s'expliquer par plusieurs facteurs :

- **La difficulté du sujet** : les tweets sont très courts, souvent ambigus, et contiennent peu de contexte.
- **La spécificité du corpus** : les messages portent sur des sujets scientifiques, où les émotions sont souvent exprimées de manière plus subtile que dans d'autres domaines (comme la publicité ou les relations personnelles).
- **Le déséquilibre des données** : certaines émotions sont beaucoup plus représentées que d'autres, ce qui désavantage les modèles dans leur apprentissage.

Ces résultats suggèrent que les méthodes existantes, même parmi les plus avancées, ne sont pas entièrement adaptées à cette nouvelle tâche. Il ne suffit pas de réutiliser des outils génériques : pour bien détecter les émotions dans des tweets à contenu scientifique, il faudrait concevoir de nouveaux modèles spécifiquement pensés pour ce contexte.

Composant	Détail
Modèle utilisé	j-hartmann/emotion-english-distilroberta-base (DistilRoBERTa pré-entraîné pour la classification d'émotions)
Tokenisation	Troncature et padding, longueur maximale fixée à 64 tokens pour limiter l'utilisation mémoire
Entraînement	Trainer de Hugging Face avec Accelerator pour la gestion de l'entraînement sur GPU
Taille du batch	8 pour l'entraînement, 16 pour l'évaluation
Époques	3 (valeur minimale pour limiter le surapprentissage et tester rapidement les performances)
Taux d'apprentissage	2e-5
Poids de régularisation	0.01 (weight decay)
Stratégie d'évaluation	Évaluation tous les 50 pas d'entraînement (eval_strategy = "steps")
Sauvegarde	Modèle sauvegardé automatiquement tous les 50 pas (save_strategy = "steps")
Critère pour meilleur modèle	Meilleur score F1 pondéré (metric_for_best_model = "f1")
Arrêt anticipé	Oui, avec EarlyStoppingCallback(patience = 2)
Optimisation mémoire	Activation de fp16 si GPU compatible, suppression explicite des variables inutiles, libération de la mémoire GPU
Évaluation finale	Mesures de précision, rappel, F1-score global et par classe, accuracy
Chemin de sauvegarde	./distilroberta-emotion-resource-friendly

TABLE 3.7 – Résumé des paramètres et stratégies de fine-tuning du modèle DistilRoBERTa.

### 3.2.3 Validation des modèles sur un corpus équilibré généré

Pour compléter notre étude, nous avons évalué les modèles sur un second jeu de données, composé à 30 % de tweets réels et à 70 % de tweets générés automatiquement. Ce corpus présente deux particularités :

- Il est **parfaitement équilibré**, c'est-à-dire qu'il contient autant d'exemples pour chaque émotion.
- Il est en grande partie **artificiel**, ce qui réduit le bruit, les fautes ou les ambiguïtés que l'on retrouve souvent dans les tweets réels.

### Pourquoi utiliser ce corpus ?

L'objectif est d'observer les capacités *pures* des modèles à distinguer les émotions dans un environnement **idéal**, sans être perturbés par les déséquilibres ou la complexité du langage naturel. Ce test en laboratoire permet de comprendre ce que chaque modèle peut théoriquement atteindre.

### Résultats obtenus

Les performances des modèles sur ce corpus équilibré sont nettement supérieures à celles observées sur le corpus réel, en particulier pour les modèles classiques. Le tableau 3.8 résume l'accuracy moyen obtenu :

TABLE 3.8 – Accuracy moyenne des modèles sur corpus équilibré et réel

Modèle	Corpus équilibré	Corpus réel
Naive Bayes	0.77	0.43
SVM Linéaire	0.77	0.48
Régression Logistique	0.77	0.45
EmoRoBERTa	0.59	0.47
DistilRoBERTa	0.63	0.63
EmoBERTa	0.55	0.57

On observe que les modèles classiques comme Naive Bayes, SVM et régression logistique affichent des accuracy autour de 0.77 sur ce corpus équilibré, bien au-dessus de leurs résultats sur le corpus réel. Les modèles basés sur des transformateurs montrent aussi une amélioration, mais dans une moindre mesure.

Le tableau 3.9 illustre la différence de performance entre le corpus équilibré et le corpus réel, mettant en évidence la perte d'efficacité des modèles dans des conditions réelles :

TABLE 3.9 – Performances des modèles sur corpus réel et équilibré avec interprétation

Modèle	F1 réel	F1 équilibré	Interprétation
Naive Bayes	0.26	0.77	Forte baisse sur réel
SVM Linéaire	0.41	0.77	Baisse notable
Régression Logistique	0.32	0.77	Forte baisse
EmoRoBERTa	0.41	0.60	Baisse modérée
DistilRoBERTa	0.48	0.62	Performance stable
EmoBERTa	0.52	0.48	Légère amélioration

### Interprétation

Ces résultats montrent clairement que les modèles réussissent mieux dans un contexte idéal où chaque émotion est représentée équitablement et où le langage est peu bruité. Toutefois, cette performance ne se transpose pas directement dans le contexte réel où :

- Le langage est plus complexe, avec des fautes, des expressions idiomatiques, et du bruit.
- La distribution des émotions est naturellement déséquilibrée, certaines émotions étant rares.

En particulier, les modèles classiques perdent jusqu'à la moitié de leur performance, ce qui indique une forte sensibilité au déséquilibre et au bruit des données. Les transformateurs sont globalement plus robustes, EmoBERTa montrant même une légère amélioration sur le corpus réel, probablement grâce à son pré-entraînement spécifique sur des données émotionnelles authentiques.

Le test sur corpus équilibré généré est une étape utile pour évaluer les capacités fondamentales des modèles dans un cadre contrôlé. Cependant, il met également en évidence les limites pratiques de ces modèles face à la complexité du langage réel. Il souligne la nécessité d'une évaluation systématique sur des données authentiques pour garantir la validité des performances dans des applications réelles.

## Chapitre 4

# Discussion et perspectives

Dans cette dernière partie, nous revenons sur les principaux enseignements du projet, en mettant en lumière ses limites, puis en ouvrant sur des pistes d'amélioration et de recherche future. L'objectif est d'apporter un regard critique, accessible à toute personne curieuse, même sans connaissances avancées en intelligence artificielle ou en traitement automatique du langage.

### 4.1 Limites du projet

#### Contraintes contextuelles et ressources

Le projet s'est déroulé sur quatre mois, en parallèle de nos études et d'autres obligations. Ce contexte a limité le temps que nous pouvions consacrer à chaque étape. Sur le plan technique, l'utilisation de la version gratuite de Google Colab a imposé des restrictions (sessions limitées, mémoire, accès aux processeurs graphiques), ce qui a influencé la durée des entraînements.

#### Corpus d'étude : taille et spécificité

Notre base de données finale contenait 1140 tweets, ce qui est suffisant pour une première exploration, mais reste trop modeste pour entraîner des modèles d'intelligence artificielle vraiment performants. De plus, nous avons choisi de travailler sur des tweets en anglais, principalement liés à la science. Cela limite la possibilité de généraliser nos résultats à d'autres langues ou à d'autres sujets. Enfin, certaines émotions (comme la peur ou le dégoût) étaient très peu représentées, ce qui a compliqué l'apprentissage des modèles, malgré l'utilisation de techniques pour rééquilibrer les données.

## Processus d'annotation et fiabilité

Pour attribuer une émotion à chaque tweet, plusieurs membres du groupe ont annoté les messages. Le coefficient Kappa de Fleiss (0,3758) montre que l'accord entre les annotateurs a été faible à passable. Cela s'explique par la difficulté d'interpréter des textes courts, parfois ambigus, et par la subjectivité de chacun. Malgré l'élaboration d'un protocole d'annotation et des discussions collectives, il restait des divergences. Le départ d'un membre du groupe en cours de projet a aussi perturbé l'organisation. Enfin, nous avons choisi de n'attribuer qu'une seule émotion principale par tweet (mono-label), ce qui simplifie l'analyse mais ne reflète pas toujours la complexité réelle des émotions exprimées.

## Biais thématiques et influence des sujets

Certains sujets, comme la COVID-19, étaient très présents dans notre corpus et souvent associés à des émotions spécifiques (par exemple, la peur). Cela a pu biaiser les modèles, qui risquent d'associer certains mots à une émotion sans en saisir le contexte réel. Faute de temps, nous n'avons pas pu explorer en profondeur l'influence des thématiques à l'aide de méthodes d'analyse avancée.

## Modélisation et évaluation

Les résultats obtenus doivent être interprétés avec prudence. La classification des émotions sur Twitter reste une tâche difficile, et certaines méthodes d'augmentation de données peuvent fausser les résultats. Par exemple, un modèle SVM a vu son score F1 passer de 0,77 à 0,41 après apprentissage sur un corpus artificiellement enrichi, montrant que certaines techniques peuvent introduire des effets indésirables.

## 4.2 Améliorations possibles et perspectives futures

Les limites identifiées ouvrent de nombreuses pistes d'amélioration, tant sur le plan de la méthode que de l'organisation.

### Renforcement du corpus et amélioration de l'annotation

- *Mieux annoter* : Revenir sur les cas où les annotateurs n'étaient pas d'accord, organiser des sessions de calibration et améliorer la formation pour augmenter la cohérence.
- *Élargir le corpus* : Collecter plus de tweets, en visant surtout les émotions peu représentées, afin de mieux équilibrer les données.

- *Prendre en compte l'intensité* : Annoter non seulement la catégorie d'émotion, mais aussi son intensité (par exemple, « légèrement en colère » ou « très joyeux »), pour enrichir l'analyse.
- *Analyser les biais* : Identifier les biais possibles dans l'annotation (subjectivité, différences culturelles), dans la répartition des émotions (classes surreprésentées), ou dans les performances des modèles. Cette analyse permettrait d'adapter les méthodes pour les rendre plus justes et représentatives.

## **Analyse approfondie du lien science–émotions–désinformation**

- *Émotions et viralité* : Étudier comment certaines émotions influencent la diffusion des tweets (nombre de partages, de likes, etc.).
- *Émotions et manipulation* : Analyser comment les émotions sont utilisées dans la désinformation scientifique, en identifiant les émotions qui servent à orienter l'opinion ou à amplifier la diffusion de fausses informations.

## **Affinement des méthodes de modélisation**

- *Techniques d'équilibrage plus avancées* : Tester des méthodes d'augmentation de données plus fines, comme l'utilisation de synonymes contextualisés ou la traduction automatique.
- *Analyse thématique* : Intégrer des outils d'analyse de sujets (comme LDA ou NMF) pour mieux comprendre et contrôler l'impact des thèmes dominants sur la détection des émotions.
- *Validation statistique* : Appliquer des tests statistiques pour appuyer les différences observées entre groupes ou émotions.
- *Fine-tuning optimisé* : Reprendre l'entraînement des modèles avancés (transformers) avec un corpus enrichi et une validation rigoureuse des paramètres.

## **Améliorations organisationnelles**

- Rédiger un protocole d'annotation détaillé dès le début du projet, avec des phases pilotes pour former et calibrer les annotateurs.
- Utiliser des outils collaboratifs dédiés (comme Doccano ou Label Studio) pour centraliser et contrôler le processus d'annotation, ce qui faciliterait le suivi et la cohérence du travail.

# Conclusion

Le projet *AI Emotion* s'est donné pour objectif d'explorer une question originale : dans quelle mesure les émotions exprimées sur Twitter peuvent-elles être liées à la nature scientifique ou non des messages partagés ? Dans un contexte où la diffusion d'informations scientifiques est souvent brouillée par la désinformation et les *fake news*, nous avons cherché à comprendre si les émotions pouvaient offrir des indices utiles sur la perception, la réception et la diffusion du discours scientifique en ligne.

Notre démarche s'est articulée autour de plusieurs étapes complémentaires. La première a consisté à constituer un corpus de tweets annotés manuellement selon les six émotions fondamentales du modèle d'Ekman, enrichies d'une catégorie *Neutre*. Malgré les difficultés liées à la subjectivité de l'annotation, mises en évidence par un accord inter-annotateurs initialement modeste (Kappa de Fleiss), un processus itératif d'ajustement et de consolidation nous a permis d'aboutir à un jeu de données final composé de 1 140 tweets annotés de manière cohérente.

L'analyse exploratoire de ce corpus a révélé des tendances significatives. L'émotion *Neutre* domine largement, mais des différences marquées apparaissent entre les tweets à caractère scientifique (associés plus fréquemment à la *Peur* ou à la *Surprise*) et ceux non scientifiques (davantage liés à la *Colère* ou à la *Joie*). Ces premiers résultats soulignent l'importance du contexte thématique dans l'expression émotionnelle.

Dans la phase de modélisation, nous avons comparé des modèles classiques (**SVM**, **Naive Bayes**, **Régression Logistique**) et des modèles pré-entraînés basés sur l'architecture Transformer (**DistilRoBERTa**, **EmoRoBERTa**, **EmoBERTa**). Après un prétraitement rigoureux, la gestion du déséquilibre des classes et l'optimisation des hyperparamètres, les modèles de type *Transformer* ont démontré de meilleures performances globales, notamment **EmoBERTa**, plus sensible aux nuances émotionnelles. La Régression Logistique, bien qu'appartenant à une famille plus simple, s'est révélée compétitive après ajustements.

Nos expérimentations ont également mis en évidence la fragilité des modèles classiques face aux déséquilibres du corpus, tandis que les Transformers se montrent plus robustes. La validation sur un sous-ensemble équilibré a confirmé ces écarts de comportement.

Ce travail a ainsi permis de cartographier, de façon préliminaire, les émotions associées aux discours scientifiques et non scientifiques sur les réseaux sociaux. Il montre que les émotions ne sont pas des perturbations à écarter, mais des éléments porteurs de sens, qui participent à la manière dont l'information est perçue et relayée.

Certaines limites subsistent. La taille du corpus, bien que suffisante pour une première exploration, limite la portée des généralisations. La focalisation sur des tweets principalement en anglais, la faible représentation de certaines émotions, ainsi que les difficultés inhérentes à l'annotation humaine réduisent également la précision des résultats. Malgré l'élaboration d'un protocole d'annotation détaillé, la subjectivité reste un défi.

Ces limites ouvrent néanmoins de nombreuses perspectives. L'enrichissement du corpus (en volume, en langues, en thématiques), l'intégration d'informations complémentaires (comme l'intensité émotionnelle ou les réactions des utilisateurs), et l'exploration de modèles plus avancés permettraient d'approfondir ces premières observations. L'étude plus fine du lien entre émotion et viralité, ou encore l'analyse ciblée des discours de désinformation scientifique, constituent également des pistes prometteuses.

En conclusion, le projet *AI Emotion* a posé les fondations d'une démarche interdisciplinaire mêlant traitement automatique du langage, analyse des émotions et communication scientifique. En démontrant la pertinence de cette approche pour mieux comprendre les dynamiques entre science, émotions et société, il ouvre la voie à des travaux futurs, susceptibles de contribuer à un espace informationnel plus lucide, nuancé et résilient.

# Bibliographie

- CROWDFLOWER. *Emotion in Text Dataset*. 2016. URL : <https://data.world/crowdflower/emotion-in-text>.
- EKMAN, Paul. "An Argument for Basic Emotions". In : *Cognition and Emotion* 6.3-4 (1992), p. 169-200. DOI : 10.1080/02699939208411068.
- FLEISS, Joseph L. "Measuring nominal scale agreement among many raters". In : *Psychological Bulletin* 76.5 (1971), p. 378-382. URL : <https://www.jstor.org/stable/2538836>.
- JOUKKI, Jukka et al. "The Future of Journalism : Fake News, Misinformation and Fact-Checking". In : *Journalism Practice* 10.7 (2016), p. 879-889. DOI : 10.1080/17512786.2016.1208056.
- MOHAMMAD, Saif et al. "SemEval-2018 Task 1 : Affect in Tweets". In : *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, Louisiana : Association for Computational Linguistics, 2018, p. 1-17. DOI : 10.18653/v1/S18-1001. URL : <https://aclanthology.org/S18-1001/>.
- PLUTCHIK, Robert. "A General Psychoevolutionary Theory of Emotion". In : *Theories of Emotion*. Sous la dir. de Robert PLUTCHIK et Henry KELLERMAN. Academic Press, 1980, p. 3-33.
- SCHELLHAMMER, Sebastian, Jonas PFEIFFER et Iryna GUREVYCH. *SciTweets – A Dataset and Annotation Framework for Detecting Scientific Online Discourse*. <https://arxiv.org/abs/2206.07360>, consulté le 22 mai 2025. 2022.
- SCIENCE ET VIE. *Réseaux sociaux et propagation des émotions : qu'est-ce que la "conscience collective numérique"*. Consulté le 22 mai 2025. 2024. URL : <https://www.science-et-vie.com/article-magazine/reseaux-sociaux-et-propagation-des-emotions-quest-ce-que-la-conscience-collective-numerique>.