

Analyse exploratoire des métriques d'annotations dans MOBILES

Madjid Sadallah, Marie Lefevre

▶ **To cite this version:**

Madjid Sadallah, Marie Lefevre. Analyse exploratoire des métriques d'annotations dans MOBILES. LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/École Centrale de Lyon. 2023. ⟨hal-05129539⟩

HAL Id: hal-05129539

https://hal.science/hal-05129539v1

Submitted on 25 Jun 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Analyse exploratoire des métriques d'annotations dans MOBILES

Madjid Sadallah & Marie Lefevre

Juillet 2023

Cette étude vise à caractériser en profondeur les annotations (posts) générées par les utilisateurs au sein de l'application d'exploration urbaine MOBILES. Pour ce faire, un ensemble riche et multidimensionnel de métriques a été défini et calculé pour chaque annotation, couvrant ses aspects textuels (volume, diversité, complexité, richesse lexicale), son utilisation des icônes (diversité, emphase sociale, dynamique), l'usage des tags (influence, sentiment, esthétique), ainsi que des indicateurs contextuels et d'expressivité émotionnelle ou graphique. Des analyses descriptives, des études de corrélation, des modèles de régression linéaire multiple (pour prédire l'harmonie lexico-syntaxique, l'activité dynamique perçue des icônes, et les sentiments exprimés par les tags) ainsi que des techniques de modélisation prédictive du volume lexical (comparant régression linéaire, arbres de décision, forêts aléatoires, boosting de gradient et réseaux de neurones) ont été menées. De plus, une analyse de clustering a été effectuée pour identifier des regroupements naturels d'annotations et les caractéristiques qui les discriminent le plus. Les résultats révèlent des relations significatives entre différentes facettes des contributions, comme l'influence négative de la profondeur syntaxique sur l'harmonie lexico-syntaxique ou l'impact de l'esthétique des tags sur le sentiment positif. L'étude démontre également la faisabilité de prédire certains aspects des annotations et d'identifier des profils de contributions distincts, offrant ainsi des pistes précieuses pour une meilleure compréhension des comportements utilisateurs et pour l'amélioration continue de l'application MOBILES.

1. Introduction

L'application MOBILES est une plateforme numérique conçue pour faciliter l'exploration urbaine et l'échange d'expériences par les étudiants internationaux. Au cœur de cette plateforme se trouvent les annotations, ou "posts", créées par les utilisateurs. Ces contributions ne sont pas de simples messages textuels ; elles constituent des enregistrements riches et multiformes des découvertes, des perceptions et des interactions des étudiants avec leur nouvel environnement. Chaque annotation encapsule potentiellement du contenu textuel, des choix iconographiques, l'utilisation de tags spécifiques, des indicateurs émotionnels, des données géographiques précises et des éléments visuels. Comprendre la nature et la diversité de ces annotations est donc essentiel pour appréhender la manière dont les utilisateurs vivent et documentent leurs expériences au sein de MOBILES.

Face à la richesse et à la complexité potentielle des annotations générées, un simple survol qualitatif s'avère insuffisant pour saisir pleinement les dynamiques de contribution et d'interaction. Il existe un besoin manifeste de caractériser ces annotations de manière plus objective et systématique. La présente étude se propose donc d'aborder cette problématique en "augmentant" chaque annotation d'un ensemble de métriques calculables, afin d'offrir une vision quantitative et structurée des contributions.

Les objectifs principaux de cette analyse exploratoire sont les suivants :

1. **Définir et calculer un ensemble complet de métriques** décrivant les différentes facettes d'une annotation : textuelles, iconographiques, liées aux tags, à l'expressivité émotionnelle et graphique, ainsi que des indicateurs contextuels.
2. **Réaliser des analyses descriptives** de ces métriques pour comprendre leurs distributions et leurs tendances centrales.
3. **Explorer les relations et corrélations** entre les différentes métriques afin d'identifier d'éventuelles interdépendances ou synergies.
4. **Évaluer la possibilité de modéliser et prédire certains aspects** des annotations, tels que le volume lexical, en utilisant les métriques comme variables explicatives.
5. **Identifier des regroupements naturels ou "clusters" d'annotations** sur la base de leurs caractéristiques métriques, et déterminer les variables qui contribuent le plus à leur différenciation. In fine, cette étude vise à établir une base analytique solide pour une meilleure compréhension des comportements de contribution des utilisateurs de MOBILES.

Pour atteindre ces objectifs, une approche méthodologique en plusieurs étapes a été adoptée. Elle débute par une phase de définition et d'extraction d'un large éventail d'attributs natifs et de métriques calculées pour chaque annotation. S'ensuivent des analyses statistiques descriptives pour chaque catégorie de métriques (textuelles, iconographiques, liées aux tags). Les interrelations entre ces métriques sont ensuite examinées via des analyses de corrélation et des modèles de régression linéaire multiple. La capacité à prédire une caractéristique clé des annotations – le volume lexical – est explorée en comparant les performances de plusieurs algorithmes de régression (allant de la régression linéaire simple aux réseaux de neurones). Enfin, une analyse de clustering non supervisée, précédée d'une réduction de dimensionnalité par Analyse en Composantes Principales (PCA), est employée pour tenter de découvrir des structures latentes ou des profils typiques d'annotations.

Le présent rapport est structuré comme suit : la Section 2 détaillera l'ensemble des métriques d'annotation définies et utilisées dans cette étude. La Section 3 présentera les résultats des analyses descriptives et exploratoires pour chaque grande catégorie de métriques, incluant les études de corrélation et les premiers modèles de régression. La Section 4 se concentrera sur les résultats de la modélisation prédictive du volume lexical. La Section 5 exposera les résultats de l'analyse de clustering. Enfin, la Section 6 proposera une discussion générale des résultats, soulignera les limites de l'étude et conclura sur les implications et les perspectives de travaux futurs.

2. Définition et Calcul des Métriques d'Annotation

Au cœur de cette étude se trouve le concept d'une "annotation augmentée de métriques". Chaque post utilisateur est ainsi enrichi d'un ensemble d'attributs natifs et de métriques calculées, conçues pour capturer diverses dimensions de la contribution. Ces métriques fournissent une base quantitative pour l'analyse des interactions et des contenus générés au sein de MOBILES. Cette section détaille ces attributs et métriques.

2.1. Attributs Natifs et Contextuels d'une Annotation

Ces attributs sont directement extraits des données brutes associées à chaque post ou à son contexte immédiat

- **postId**: Identifiant unique du post.
- **userId**: Identifiant de l'utilisateur ayant créé le post.

- **institution-<institution>** (par exemple, institution-Polytech) : Affiliation institutionnelle de l'utilisateur, binarisée par institution.
- **type**: Type du post, indiquant l'action sous-jacente (ex: addPost, editPost).
- **shared**: Indicateur binaire précisant si le post a été partagé.
- **icons**: Liste des identifiants ou noms des icônes utilisées dans le post.
- **tags**: Liste des tags textuels associés au post.
- **timing**: Fréquence d'occurrence perçue ou déclarée du post (valeurs possibles : OCCASIONAL, REGULAR, UNIQUE).
- **placeType**: Type de lieu associé au post (ex: restaurant, parc).
- **message**: Contenu textuel brut du post.
- **message_length**: Longueur du message textuel brut (nombre de caractères ou de mots).
- **clean_message**: Message textuel après des étapes de nettoyage (suppression de la ponctuation, stopwords, etc.).
- **clean_message_length**: Longueur du message textuel nettoyé.
- **imageId**: Identifiant de l'image potentiellement associée au post.
- **has_image**: Indicateur binaire signalant la présence d'une image dans le post.
- **emoticon**: Émoticône principale utilisée dans le post.
- **emoticon-<emotion>** (par exemple, emoticon-happy) : Indicateurs spécifiques liés à l'utilisation d'émoticônes particulières, binarisés par type d'émotion.
- **begin**: Horodatage marquant le début de la création ou de l'événement lié au post.
- **position (de l'utilisateur)** :
    - lat : Latitude de la position géographique de l'utilisateur au moment de la création du post.
    - lon : Longitude de la position géographique de l'utilisateur.
- **coordinates (du post)** :
    - lat : Latitude des coordonnées géographiques associées au lieu du post.
    - lon : Longitude des coordonnées géographiques associées au lieu du post.
- **place**: Indicateurs spécifiques liés à la catégorisation du type de lieu :
    - city : Indique si le lieu est de type "ville" (contexte général).
    - close : Indique si le lieu est de type "proche" (par rapport à l'utilisateur ou un point de référence).
    - district : Indique si le lieu est de type "district" ou quartier.
    - here : Indique si le lieu est de type "ici" (localisation actuelle).
    - tour : Indique si le lieu est associé à un parcours.
- **tag-<tag>** (par exemple, tag-food) : Indicateurs spécifiques liés à l'utilisation de tags particuliers, binarisés par tag.
- **icon-<icon>** (par exemple, icon-restaurant) : Indicateurs spécifiques liés à l'utilisation d'icônes particulières, binarisés par icône.

2.2. Métriques Calculées à partir des Attributs du Post

Ces métriques sont dérivées des attributs natifs pour quantifier des aspects plus spécifiques des contributions. Elles sont regroupées sous un attribut conteneur metrics dans chaque post.

2.2.1. Métriques Liées à l'Utilisation des Icônes (iconography)

Ces mesures évaluent différents aspects de l'utilisation des icônes dans les annotations :

- **emphasis_social**: Mesure de l'accentuation sociale véhiculée par les icônes choisies.
- **diversity_types_icons_used**: Mesure de la variété des types ou catégories d'icônes utilisées.
- **atmosphere_environment**: Mesure de la capacité des icônes à dépeindre une atmosphère ou un environnement particulier.
- **activity_dynamics**: Mesure de la dynamique d'activité ou du mouvement suggéré par les icônes.
- **resonance_memory**: Mesure de la capacité des icônes à évoquer une résonance mémorielle.
- **frequency_sensationnel**: Mesure de la fréquence d'utilisation d'icônes liées à un aspect sensationnel ou marquant.
- **efficiency_icons_used**: Mesure de l'efficacité perçue des icônes à transmettre une information.

2.2.2. Métrique d'Expressivité Émotionnelle (emoticonality)

- **emoticonality**: Niveau d'expressivité émotionnelle binaire, indiquant la présence (1) ou l'absence (0) d'émoticônes dans le post.

2.2.3. Métriques Liées à l'Utilisation des Tags (tagging)

Ces mesures analysent l'usage et l'impact des tags :

- **social_interaction_index**: Indice reflétant le potentiel d'interaction sociale des tags utilisés.
- **positive_sentiment_ratio**: Proportion de tags exprimant un sentiment positif par rapport au nombre total de tags.
- **aesthetic_tag_index**: Indice évaluant la connotation esthétique des tags.
- **negative_sentiment_ratio**: Proportion de tags exprimant un sentiment négatif par rapport au nombre total de tags.
- **tag_influence_index**: Indice mesurant l'influence ou la popularité potentielle des tags utilisés.

2.2.4. Métrique d'Expressivité Graphique (graphical)

- **graphical**: Niveau d'expressivité graphique binaire, indiquant la présence (1) ou l'absence (0) d'images dans le post. (Note : redondant avec has_image ; à clarifier si ce sont deux concepts distincts ou le même).

2.2.5. Métriques Textuelles sur une Annotation Spécifique (comment et textual)

Ces métriques évaluent la richesse et la complexité du contenu textuel d'un post unique :

- **Volume Lexical**: Mesure la taille du vocabulaire utilisé, exprimée par le nombre total de mots dans le commentaire.
- **Diversité Lexicale**: Évalue la variété de mots utilisés en calculant le rapport entre le nombre de mots uniques et le nombre total de mots dans le commentaire.
- **Rapport de Richesse Lexicale**: Indique la diversité lexicale ajustée à la longueur du texte, calculée en divisant le nombre de mots uniques par la racine carrée du nombre total de mots.
- **Complexité Syntaxique (complexite_syntaxique)**: Mesure la complexité des structures de phrases en évaluant la moyenne des unités syntaxiques par mot dans le commentaire.

2.2.6. Métriques Globales et Tendancielles sur l'Ensemble des Textes d'un Utilisateur (global et tendancy)

Ces métriques, calculées sur l'ensemble des contributions textuelles d'un utilisateur, offrent une vision de son style d'écriture global et de son évolution :

- **Métriques Globales (global) :**
    - **harmonie_lexico_syntaxique_utilisateur**: Mesure de l'harmonie lexico-syntaxique globale de l'utilisateur, potentiellement un ratio entre sa diversité lexicale cumulative et sa profondeur syntaxique cumulative.
    - **ratio_diversite_lexicale_cumulative**: Ratio cumulatif de diversité lexicale sur l'ensemble de ses contributions.
    - **profondeur_syntaxique_cumulative**: Profondeur syntaxique cumulative observée sur l'ensemble de ses contributions.
- **Métriques Tendancielles (tendancy) :**
    - **evolution_complexite_syntaxique**: Mesure l'évolution de la complexité syntaxique de l'utilisateur au fil du temps et de ses posts.
    - **progression_diversite_lexicale**: Mesure la progression de la diversité lexicale de l'utilisateur au fil du temps.
    - **dynamique_harmonie_lexico_syntaxique**: Mesure la dynamique de l'harmonie lexico-syntaxique de l'utilisateur au fil du temps.

La définition précise de ces métriques est fondamentale car elles constituent les variables d'entrée pour toutes les analyses statistiques et de modélisation présentées dans les sections suivantes de ce rapport.

3. Analyses Descriptives et Exploratoires des Métriques

Après avoir défini l'ensemble des métriques caractérisant les annotations au sein de MOBILES, cette section présente les résultats des premières analyses descriptives et exploratoires. L'objectif est de comprendre la distribution de ces métriques, d'identifier leurs tendances centrales, d'explorer leurs interrelations et de mener des analyses de régression initiales pour élucider certaines dynamiques sous-jacentes aux contributions des utilisateurs.

3.1. Analyse des Métriques Textuelles sur une Annotation Individuelle

Les premières métriques examinées concernent les caractéristiques textuelles intrinsèques à chaque annotation. Elles incluent :

- **Volume Lexical**: Nombre total de mots.
- **Diversité Lexicale**: Rapport mots uniques / mots totaux.
- **Rapport de Richesse Lexicale**: Mots uniques / mots totaux.
- **Complexité Syntaxique**: Moyenne des unités syntaxiques par mot.

3.1.1. Distributions des Métriques Textuelles Individuelles

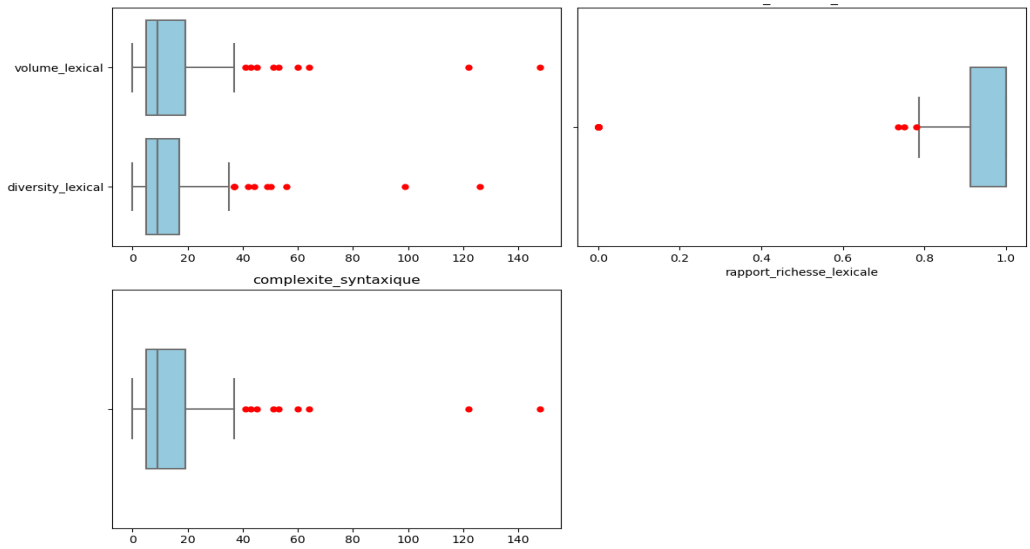


Figure 1. *Boxplots du Volume Lexical, de la Diversité Lexicale et de la Complexité Syntaxique par annotation. Ces graphiques illustrent la dispersion, la tendance centrale et les valeurs atypiques pour chaque métrique textuelle de base.*

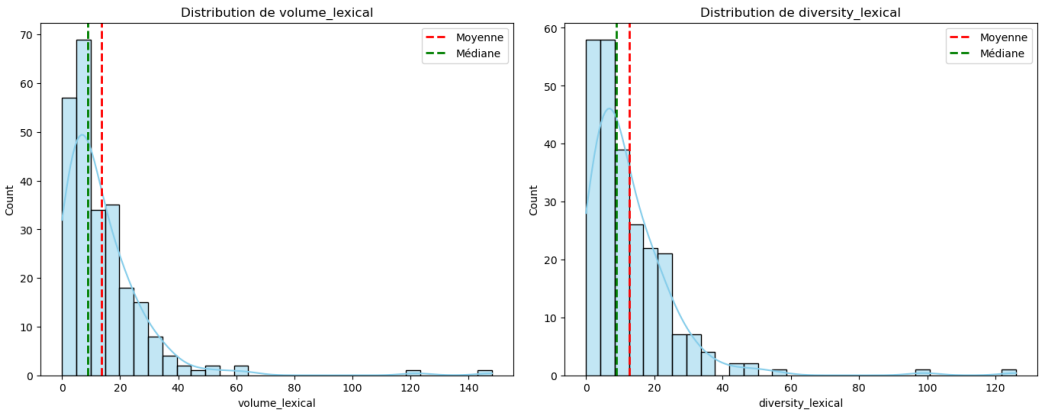


Figure 2. *Distribution du Volume Lexical et Diversité Lexicale des annotations. L'histogramme à gauche montre la fréquence des différentes longueurs de texte (nombre de mots) dans les annotations. Les lignes pointillée et tiretée représentent respectivement la moyenne et la médiane. L'histogramme à droite illustre la fréquence des différents ratios de mots uniques par rapport au nombre total de mots. Les lignes pointillée et tiretée représentent respectivement la moyenne et la médiane.*

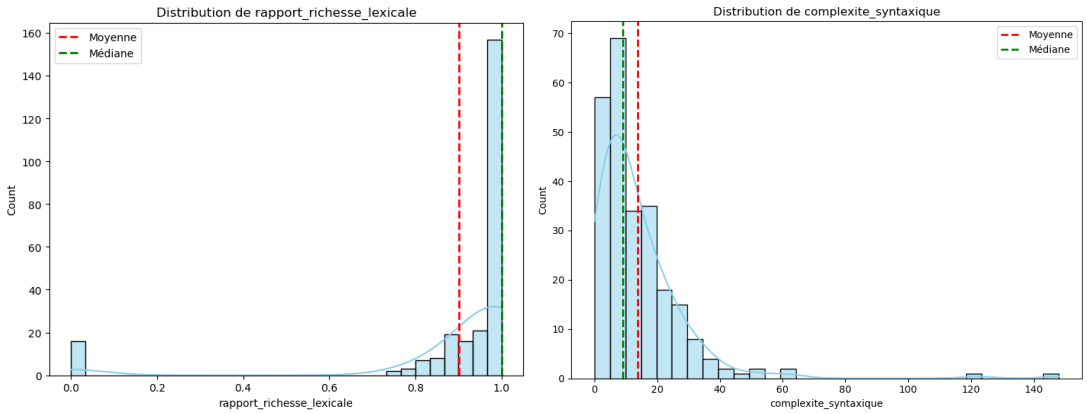


Figure 3. *Gauche : Distribution du Rapport de Richesse Lexicale des annotations. L'histogramme montre la fréquence des différents ratios de diversité lexicale ajustée à la longueur du texte. Les lignes pointillée et tiretée représentent respectivement la moyenne et la médiane. Droite : Distribution de la Complexité Syntaxique des annotations. L' histogramme illustre la fréquence des différentes moyennes d'unités syntaxiques par mot. Les lignes pointillée et tiretée représentent respectivement la moyenne et la médiane.*

Interprétation des Résultats

- **Volume Lexical :** L'histogramme de distribution du *volume_lexical* montre une distribution fortement asymétrique à droite (*positive skew*). La majorité des annotations semblent contenir un nombre relativement faible de mots, avec une queue de distribution s'étendant vers des volumes plus importants. Le boxplot correspondant confirme cette tendance, avec une médiane et une boîte interquartile concentrées vers les faibles valeurs, et la présence de nombreux points atypiques (outliers) indiquant des annotations occasionnellement très longues. La moyenne (indiquée par la ligne pointillée rouge sur l'histogramme) est vraisemblablement tirée vers la droite par ces outliers, étant supérieure à la médiane (ligne verte). Cela suggère que si la plupart des posts sont concis, certains utilisateurs produisent des textes substantiels.

- **Diversité Lexicale :** La distribution de la diversity_lexical apparaît également asymétrique à droite, bien que de manière moins prononcée que le volume lexical. La plupart des annotations ont un ratio de diversité lexicale modéré. La présence d'une concentration de valeurs vers la droite pourrait indiquer que, même pour des textes courts, les utilisateurs tendent à utiliser un vocabulaire varié proportionnellement à la longueur. Le boxplot montre une dispersion significative, avec une médiane indiquant un certain niveau de variété.

- **Rapport de Richesse Lexicale :** L'histogramme du rapport_richesse_lexicale présente une distribution qui semble tendre vers une forme plus normale, possiblement légèrement asymétrique à gauche ou bimodale. Cette métrique, en ajustant la diversité par la racine carrée du volume, vise à normaliser l'effet de la longueur du texte. La position de la moyenne et de la médiane suggère une tendance centrale autour de valeurs élevées (proches de 0.8-1.0), ce qui pourrait indiquer une bonne richesse lexicale relative pour la majorité des posts une fois la longueur prise en compte.

- **Complexité Syntaxique :** La distribution de la complexite_syntaxique est, comme le volume lexical, fortement asymétrique à droite. La grande majorité des posts présentent une faible complexité syntaxique, avec quelques posts affichant des valeurs nettement plus élevées. Cela suggère un style d'écriture majoritairement simple et direct, avec des exceptions.

- **Relations Visuelles (Boxplots Groupés) :** Les boxplots "Volume Lexical & Diversity Lexical" et "complexite_syntaxique" montrent une variabilité importante pour chaque métrique, avec la présence d'outliers, en particulier pour le volume et la complexité.
- **Graphique "Diversity Lexical within Volume Lexical" :** Ce graphique à barres n'est pas immédiatement clair sans légende expliquant "Non-Diversity Lexical". S'il représente la proportion de mots non uniques par rapport aux mots uniques pour chaque post (indexé en abscisse), il pourrait illustrer comment la répétition de mots varie en fonction du volume lexical (représenté par la hauteur des barres "Volume Lexical"). Une analyse plus poussée serait nécessaire pour interpréter pleinement ce graphique spécifique.

3.1.2. Corrélations entre les Métriques Textuelles

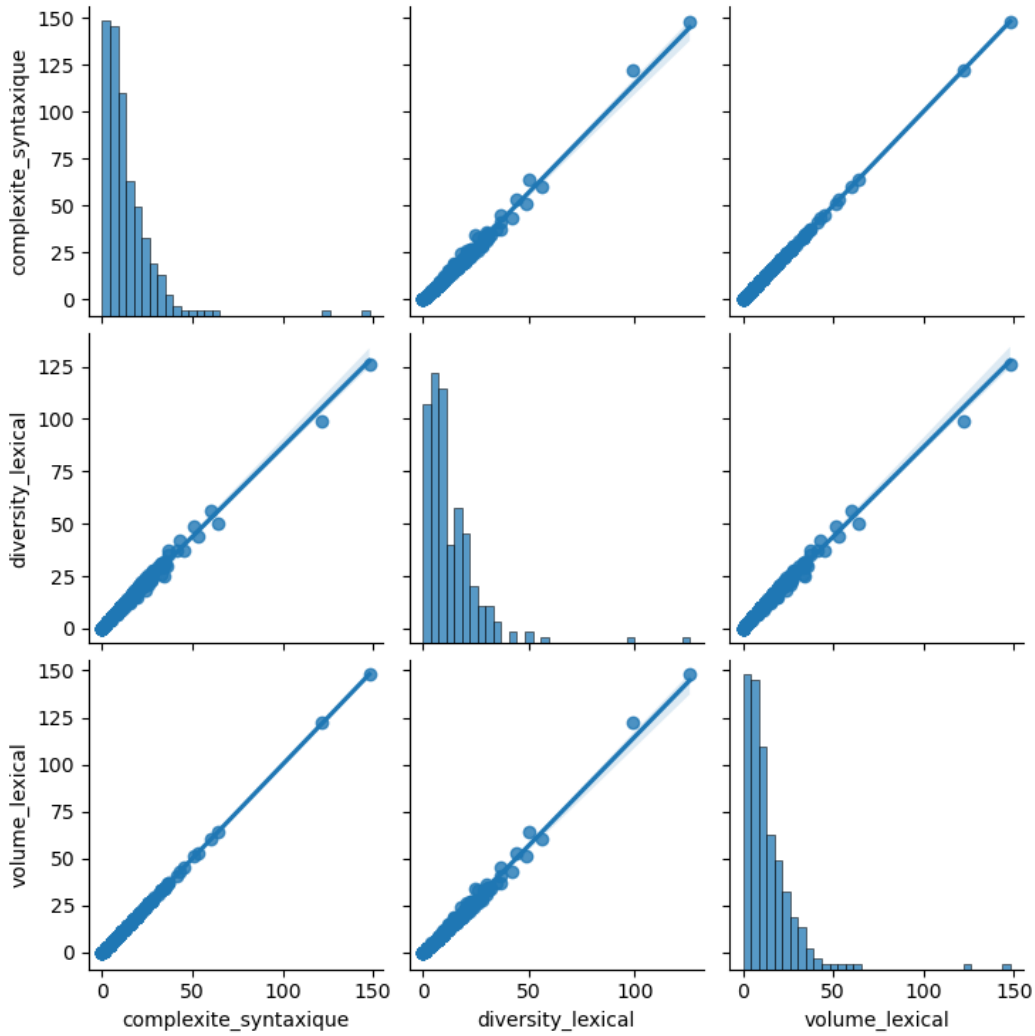


Figure 4. *Matrice de corrélation et de distribution des métriques textuelles de base par annotation. Les graphiques sur la diagonale représentent les histogrammes de distribution de chaque métrique. Les graphiques hors diagonale montrent les nuages de points illustrant les relations par paires entre ces métriques, avec une droite de régression linéaire superposée.*

Interprétation des résultats

L'examen des nuages de points et des droites de régression superposées dans la matrice permet de déduire les relations suivantes entre les métriques textuelles individuelles. Contrairement à une affirmation initiale d'une corrélation parfaite, les graphiques suggèrent des tendances fortes mais avec une dispersion notable, ce qui est attendu pour des données réelles.

- **Relation entre Volume Lexical et Diversité Lexicale :**

Le nuage de points montre une **forte corrélation positive** entre le volume lexical d'une annotation et sa diversité lexicale. Les points tendent à s'aligner le long d'une pente ascendante, indiquant que les posts plus longs (volume lexical plus élevé) ont une propension significative à contenir un plus grand nombre de mots uniques (diversité lexicale plus élevée). Cela est intuitivement logique : un texte plus long offre davantage d'opportunités d'introduire de nouveaux termes. Cependant, la dispersion des points autour de la droite de régression indique que d'autres facteurs influencent également la diversité lexicale, et que tous les textes longs ne sont pas uniformément diversifiés.

- **Relation entre Volume Lexical et Complexité Syntaxique :**

Une **corrélation positive, probablement modérée à forte,** est également observable entre le volume lexical et la complexité syntaxique. Les annotations plus longues semblent associées à une complexité syntaxique accrue. Cela peut s'expliquer par le fait que l'expression d'idées plus développées, typique des textes plus longs, nécessite souvent des structures de phrases plus élaborées, avec davantage de clauses subordonnées ou de constructions complexes. Là encore, la dispersion suggère que la longueur n'est pas le seul déterminant de la complexité.

- **Relation entre Diversité Lexicale et Complexité Syntaxique :**

Le nuage de points suggère une **corrélation positive** entre la diversité lexicale et la complexité syntaxique. Les posts qui utilisent un vocabulaire plus varié tendent également à présenter une structure syntaxique plus complexe. Cette relation pourrait indiquer un style d'écriture globalement plus "riche" ou "sophistiqué" : les utilisateurs capables d'employer un large éventail de mots pourraient également être plus enclins à construire des phrases plus complexes pour articuler leurs pensées.

Synthèse

Les analyses visuelles des corrélations indiquent que les trois principales métriques de la richesse textuelle d'une annotation – son volume, sa diversité lexicale et sa complexité syntaxique – ne sont pas indépendantes. Elles tendent à évoluer conjointement : une annotation plus longue est susceptible d'être à la fois plus diversifiée dans son vocabulaire et plus complexe dans sa syntaxe. Cette interdépendance est importante à considérer lors d'analyses ultérieures, car elle suggère que ces métriques pourraient capturer des facettes d'un construit sous-jacent lié à "l'élaboration" ou à la "richesse" d'une contribution textuelle. Il serait pertinent, dans des travaux futurs, de quantifier ces corrélations à l'aide de coefficients (par exemple, le coefficient de Pearson) et de tester leur significativité statistique pour confirmer ces observations visuelles.

3.2. Analyse des Métriques Textuelles sur Toutes les Annotations d'un Utilisateur

Au-delà des caractéristiques d'une annotation individuelle, il est instructif d'examiner des métriques qui reflètent le style textuel global d'un utilisateur ou son évolution au fil de ses contributions. Les métriques considérées ici sont :

- **Harmonie Lexico-Syntaxique de l'Utilisateur**: Mesure la cohérence entre le vocabulaire et la syntaxe de l'utilisateur (calculée en divisant la Diversité Lexicale Cumulative par la Profondeur Syntaxique Cumulative).
- **Ratio Cumulatif de Diversité Lexicale**: Évalue la diversité lexicale de l'utilisateur au fil du temps (rapport entre le nombre total de mots différents utilisés et le nombre cumulatif de mots).
- **Profondeur Syntaxique Cumulative**: Évalue l'évolution cumulative de la complexité syntaxique (nombre moyen de clauses par unité en fonction du nombre cumulatif d'unités syntaxiques).

3.2.1. Distributions des Métriques Textuelles Utilisateur Globales

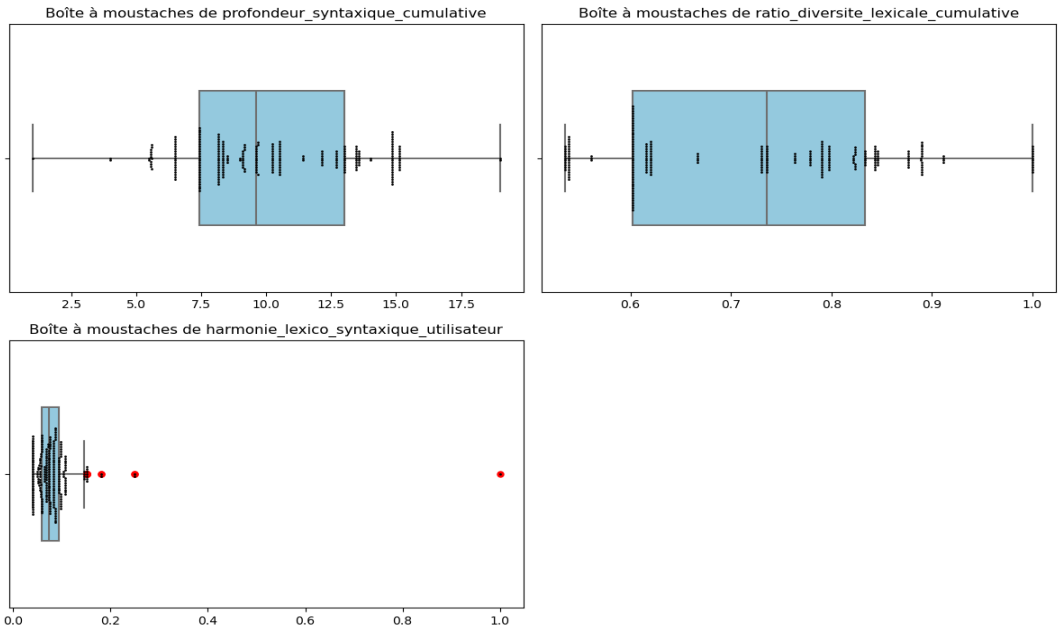


***Figure 5**. Distributions des métriques textuelles cumulatives au niveau utilisateur, représentée comme Boxplots de la Profondeur Syntaxique Cumulative, du Ratio de Diversité Lexicale Cumulative et de l'Harmonie Lexico-Syntaxique Utilisateur. Ces graphiques illustrent la dispersion, la tendance centrale et les valeurs atypiques pour chaque métrique textuelle globale par utilisateur.*

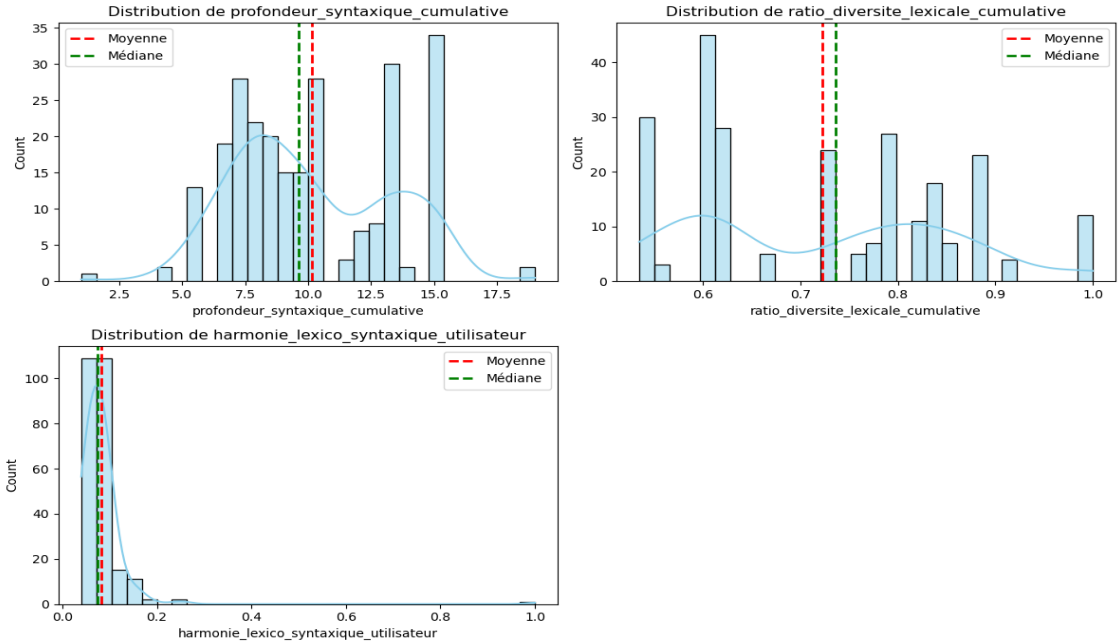


***Figure 6**. Distributions des métriques textuelles cumulatives au niveau utilisateur, représentée comme des Histogrammes de distribution de la Profondeur Syntaxique Cumulative, du Ratio de Diversité Lexicale Cumulative et de l'Harmonie Lexico-Syntaxique Utilisateur. Ces graphiques montrent la fréquence des différentes valeurs pour chaque métrique globale. Les lignes pointillée et tiretée représentent respectivement la moyenne et la médiane.*

Interprétation

- **Profondeur Syntaxique Cumulative :**

L'histogramme de la profondeur_syntaxique_cumulative montre une distribution qui semble multimodale ou, du moins, pas clairement unimodale, avec plusieurs pics. La majorité des utilisateurs se situent dans une plage de valeurs de profondeur syntaxique moyenne (autour de 7.5 à 15, si l'on se fie aux pics les plus élevés). Le boxplot correspondant montre une médiane située autour de 10-11, avec une boîte interquartile assez large et la présence de quelques outliers, notamment vers les valeurs plus élevées, indiquant des utilisateurs avec une complexité syntaxique cumulative significativement plus grande.

- **Ratio de Diversité Lexicale Cumulative :**

La distribution du ratio_diversite_lexicale_cumulative est concentrée vers les valeurs élevées (principalement entre 0.6 et 0.9), avec un pic marqué autour de 0.7-0.8. Cela suggère que, cumulativement, la plupart des utilisateurs maintiennent un bon niveau de diversité dans leur vocabulaire. Le boxplot confirme cette tendance, avec une médiane élevée et une dispersion relativement contenue, bien que des outliers existent des deux côtés.

- **Harmonie Lexico-Syntaxique Utilisateur :**

L'histogramme de l'harmonie_lexico_syntaxique_utilisateur est fortement asymétrique à droite, avec une très grande majorité d'utilisateurs présentant des valeurs d'harmonie faibles (proches de 0.0 à 0.2). Quelques utilisateurs se distinguent avec des valeurs plus élevées. Le boxplot illustre également cette concentration vers les faibles valeurs avec de nombreux outliers positifs. Étant donné que cette métrique est un ratio de la diversité lexicale cumulative (souvent élevée) par la profondeur syntaxique cumulative (plus variable mais pouvant être élevée), des valeurs faibles d'harmonie pourraient indiquer soit une profondeur syntaxique cumulative proportionnellement beaucoup plus grande que la diversité lexicale, soit une manière spécifique dont ces deux aspects interagissent.

3.2.2. Corrélations entre les Métriques Textuelles Utilisateur Globales

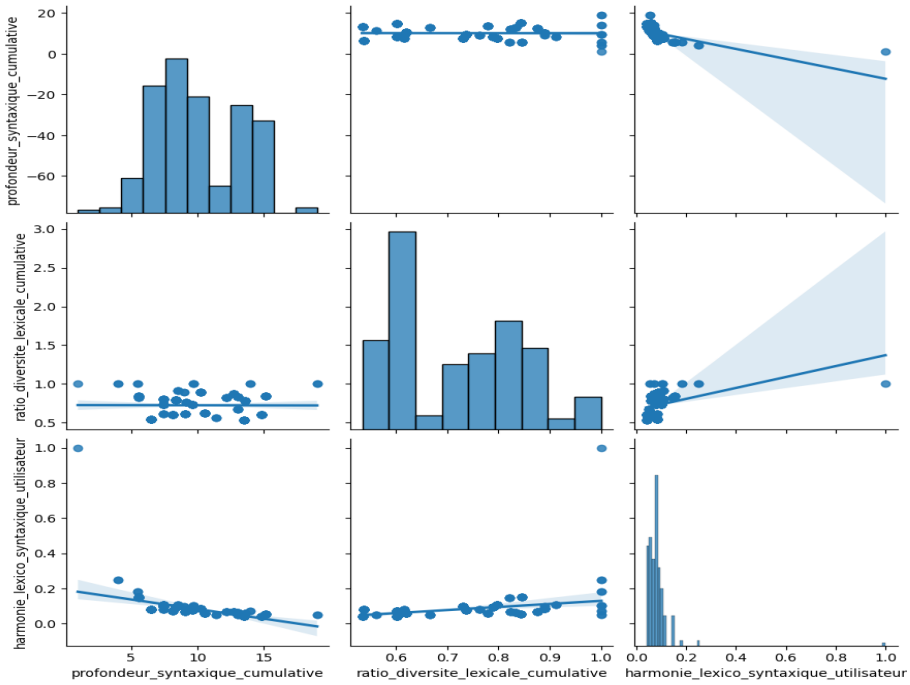


Figure 7. *Matrice de corrélation et de distribution des métriques textuelles cumulatives au niveau utilisateur. Les graphiques sur la diagonale représentent les histogrammes de distribution de chaque métrique. Les graphiques hors diagonale montrent les nuages de points illustrant les relations par paires entre ces métriques, avec une droite de régression linéaire superposée*

1. Relation entre Profondeur Syntaxique Cumulative et Ratio de Diversité Lexicale Cumulative :

L'analyse n'a révélé aucune corrélation statistiquement significative (au seuil de confiance de 95%) entre la profondeur syntaxique cumulative d'un utilisateur et son ratio de diversité lexicale cumulative. Cette absence de lien est d'ailleurs corroborée par l'inspection visuelle du nuage de points, qui montre une dispersion des données sans directionnalité marquée. Concrètement, cela suggère que, pour les utilisateurs de cet échantillon, l'évolution vers une plus grande complexité syntaxique dans leurs écrits ne s'accompagne pas nécessairement d'un enrichissement proportionnel de leur vocabulaire, et inversement. Ces deux facettes de l'évolution stylistique – la complexification de la structure des phrases et l'expansion du lexique – semblent donc progresser de manière indépendante l'une de l'autre.

2. Relation entre Profondeur Syntaxique Cumulative et Harmonie Lexico-Syntaxique

Une corrélation de Pearson de -0.52, statistiquement significative (au seuil de confiance de 95%), a été observée entre la profondeur syntaxique cumulative et l'harmonie lexico-syntaxique de l'utilisateur. Cette corrélation négative, de force modérée à notable, indique une tendance claire : à mesure que la complexité syntaxique cumulative d'un utilisateur augmente, son score d'harmonie lexico-syntaxique tend à diminuer. Le nuage de points illustre bien cette relation inverse, malgré une certaine dispersion. Étant donné que l'harmonie est calculée comme le ratio Diversité Lexicale Cumulative / Profondeur Syntaxique Cumulative, cet effet est mathématiquement cohérent : si la profondeur syntaxique (au dénominateur) croît plus rapidement ou sans une augmentation proportionnelle de la diversité lexicale (au numérateur), le score d'harmonie s'en trouvera mécaniquement réduit. Ce résultat suggère que l'accroissement de la complexité structurelle des phrases se fait fréquemment sans un enrichissement lexical équivalent, impactant ainsi négativement la mesure d'harmonie telle que définie.

3. Relation entre Ratio de Diversité Lexicale Cumulative et Harmonie Lexico-Syntaxique de l'Utilisateur :

L'étude a mis en évidence une corrélation de Pearson positive et statistiquement significative (au seuil de confiance de 95%) de 0.35 entre le ratio de diversité lexicale cumulative et l'harmonie lexico-syntaxique de l'utilisateur. Bien que de force modérée, cette corrélation, visible sur le nuage de points par une tendance ascendante (quoique plus diffuse que la précédente), suggère que les utilisateurs qui parviennent à enrichir leur vocabulaire de manière cumulative tendent également à présenter un score d'harmonie lexico-syntaxique plus élevé. En se référant à la définition de l'harmonie (Diversité / Profondeur), cela s'explique si l'augmentation de la diversité lexicale (au numérateur) n'est pas annulée par une croissance excessive de la profondeur syntaxique (au dénominateur). Ainsi, un vocabulaire cumulativement plus riche semble contribuer positivement à l'harmonie, potentiellement parce que la complexité syntaxique de ces utilisateurs reste relativement maîtrisée ou évolue moins rapidement.

3.2.3. Modélisation de l'Harmonie Lexico-Syntaxique de l'Utilisateur par Régression OLS

OLS Regression Results

Dep. Variable:	harmonie_lexico_syntaxique_utilisateur	R-squared:	0.387
Model:	OLS	Adj. R-squared:	0.382
Method:	Least Squares	F-statistic:	77.65
Date:	Sun, 28 Jan 2024	Prob (F-statistic):	7.22e-27
Time:	14:24:21	Log-Likelihood:	384.34
No. Observations:	249	AIC:	-762.7
Df Residuals:	246	BIC:	-752.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0674	0.021	3.180	0.002	0.026	0.109
profondeur_syntaxique_cumulative	-0.0109	0.001	-10.301	0.000	-0.013	-0.009
ratio_diversite_lexicale_cumulative	0.1727	0.025	6.957	0.000	0.124	0.222

Omnibus:	522.081	Durbin-Watson:	1.918
Prob(Omnibus):	0.000	Jarque-Bera (JB):	404446.702
Skew:	13.335	Prob(JB):	0.00
Kurtosis:	198.631	Cond. No.	102.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

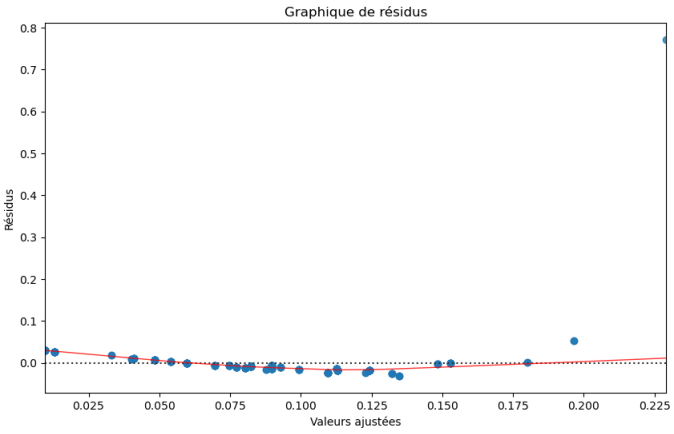


***Figure 8**. Graphique des résidus du modèle de régression OLS pour l'Harmonie Lexico-Syntaxique Utilisateur. Ce graphique représente les résidus (erreurs de prédiction) en fonction des valeurs ajustées (prédites) par le modèle*

Interprétation:

Le modèle de régression vise à expliquer les valeurs de la variable *harmonie_lexico_syntaxique_utilisateur* en utilisant les valeurs des variables *profondeur_syntaxique_cumulative* et *ratio_diversite_lexicale_cumulative* comme prédicteurs.

- **Ajustement du Modèle (R-carré) :** Le R-carré est de 0.39 (R-carré ajusté de 0.382). Cela signifie que le modèle explique environ 39% de la variance observée dans l'harmonie lexico-syntaxique des utilisateurs. C'est une part modérée, indiquant que d'autres facteurs non inclus dans le modèle influencent également cette harmonie, mais que les deux prédicteurs retenus ont un pouvoir explicatif certain.
- **Significativité Globale du Modèle (Statistique F) :** La statistique F est de 77.65 avec une p-value très faible (7.22e-27). Cela indique que le modèle dans son ensemble est statistiquement très significatif ; les prédicteurs combinés ont une relation significative avec la variable dépendante.
- **Coefficients des Prédicteurs :**
    - **profondeur_syntaxique_cumulative :** Le coefficient est de -0.0109 (arrondi à -0.01 dans le texte). Il est statistiquement très significatif (P>|t| = 0.000). Cela signifie qu'une augmentation d'une unité de la profondeur syntaxique cumulative est associée à une diminution d'environ 0.011 unités de l'harmonie lexico-syntaxique, toutes choses étant égales par ailleurs. Cela confirme la corrélation négative observée précédemment.
    - **ratio_diversite_lexicale_cumulative :** Le coefficient est de 0.1727 (arrondi à 0.17). Il est également statistiquement très significatif (P>|t| = 0.000). Une augmentation d'une unité du ratio de diversité lexicale cumulative est associée à une augmentation d'environ 0.173 unités de l'harmonie lexico-syntaxique, toutes choses étant égales par ailleurs. Cela confirme la corrélation positive.
- **Constante (const) :** Le coefficient de 0.0674 est significatif, représentant la valeur attendue de l'harmonie lorsque les deux prédicteurs sont à zéro (ce qui peut ne pas avoir d'interprétation pratique directe si des valeurs nulles pour les prédicteurs ne sont pas réalistes).
- **Graphique des Résidus :** Le graphique des résidus (différence entre les valeurs observées et prédites) par rapport aux valeurs ajustées semble montrer une dispersion relativement aléatoire autour de zéro, sans structure évidente (comme un entonnoir ou une courbe), ce qui est un bon signe pour la validité des hypothèses du modèle de régression linéaire (homoscédasticité, linéarité). Il y a quelques points qui s'écartent davantage, ce qui est normal.

Conclusion :
L'analyse des métriques textuelles globales au niveau de l'utilisateur montre que la profondeur syntaxique cumulative et la diversité lexicale cumulative sont des indicateurs importants du style d'écriture. Tandis qu'elles semblent évoluer indépendamment l'une de l'autre, elles ont des effets opposés et significatifs sur "l'harmonie lexico-syntaxique" : une plus grande complexité syntaxique tend à la diminuer, tandis qu'une plus grande diversité lexicale tend à l'augmenter. Le modèle de régression confirme ces influences et suggère que ces deux facteurs expliquent une part non négligeable des variations dans cette mesure d'harmonie.

3.3. Analyse des Métriques Relatives à l'Utilisation des Icônes

Les icônes constituent un élément visuel important des annotations dans MOBILES, potentiellement porteuses de sens et d'expressivité. Cette section examine un ensemble de métriques conçues pour évaluer différents aspects de leur utilisation :

- **Efficiency Icons Used**: Efficacité globale des icônes à transmettre une information.
- **Diversity Types Icons Used**: Variété des catégories ou types d'icônes.
- **Emphasis Social**: Accentuation des aspects sociaux par les icônes.
- **Atmosphere Environment**: Capacité des icônes à dépeindre une ambiance ou un environnement.
- **Activity Dynamics**: Intensité visuelle de l'activité ou du mouvement suggéré.
- **Resonance Memory**: Capacité des icônes à susciter un impact mémoriel.
- **Frequency Sensationnel**: Fréquence d'utilisation d'icônes liées à un aspect marquant.

3.3.1. Distributions des Métriques Iconographiques

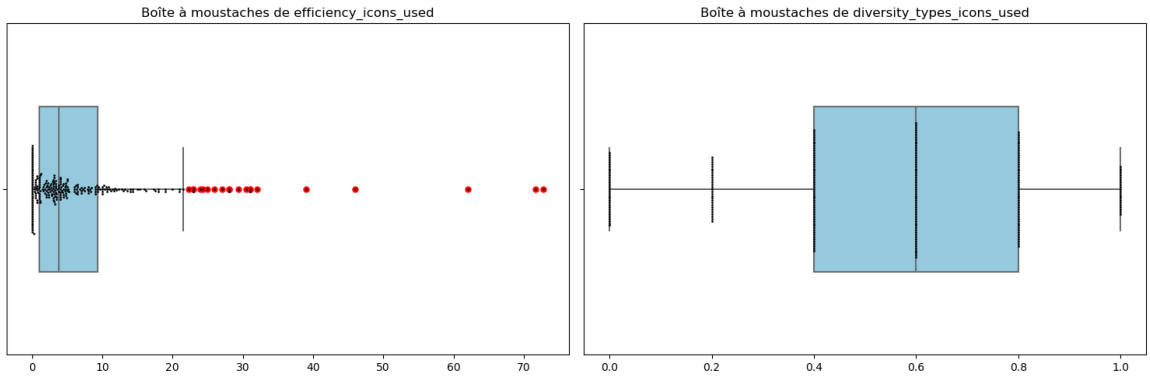


Figure 9. *Distributions des métriques iconographiques par annotation. Gauche : Boxplot de l'Efficacité des Icônes Utilisées. Droite : Boxplot de la Diversité des Types d'Icônes Utilisées.*

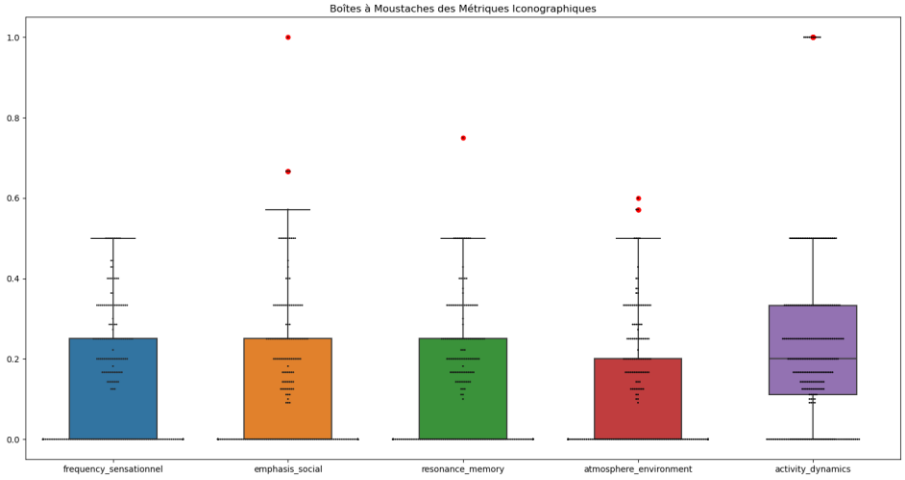


Figure 10. *Boxplots comparatifs des métriques iconographiques qualitatives.*

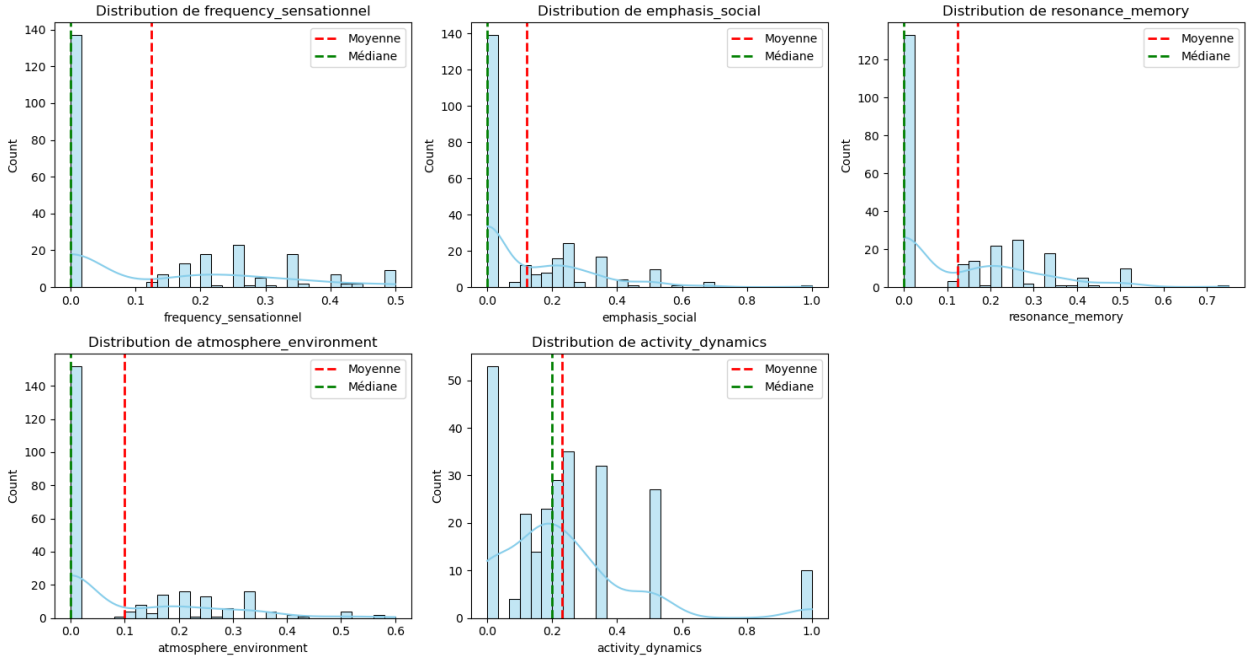


Figure 11. *Histogrammes de distribution des métriques iconographiques qualitatives. Les lignes pointillée et tiretée représentent respectivement la moyenne et la médiane pour chaque métrique.*

Interprétation des résultats

- **Efficiency Icons Used (Boxplot) :** Ce boxplot montre une très forte concentration des valeurs à zéro, avec une longue série de points atypiques (outliers) s'étendant vers des valeurs plus élevées. Cela suggère que pour la grande majorité des annotations, l'efficacité perçue des icônes (telle que mesurée) est nulle ou très faible, et que seule une petite fraction des posts utilise des icônes jugées "efficaces". Il serait intéressant de comprendre comment cette efficacité est calculée pour mieux interpréter cette distribution.
- **Diversity Types Icons Used (Boxplot) :** Le boxplot de la diversité des types d'icônes montre une médiane proche de 0.5, avec une boîte interquartile indiquant une certaine variabilité. Il y a moins d'outliers extrêmes que pour l'efficacité. Cela suggère que les utilisateurs emploient une variété modérée de types d'icônes, sans tendre vers une ultra-spécialisation (très faible diversité) ni une exploration exhaustive de tous les types (très haute diversité) pour la plupart des posts.
- **frequency_sensationnel :** L'histogramme et le boxplot montrent une distribution fortement asymétrique à droite, avec la majorité des valeurs concentrées près de zéro. Cela indique que les icônes à caractère "sensationnel" sont utilisées peu fréquemment.
- **emphasis_social :** Distribution également asymétrique à droite, mais peut-être un peu moins que la précédente. La plupart des posts ont une faible emphase sociale via les icônes, mais il y a une plus grande proportion de posts avec des valeurs modérées.

- **resonance_memory :** Similaire aux précédentes, avec une concentration des valeurs faibles, suggérant un impact mémoriel généralement bas pour les icônes utilisées dans la majorité des posts.
- **atmosphere_environment :** Encore une distribution asymétrique à droite, indiquant que la capacité des icônes à dépeindre une atmosphère est souvent faible, mais avec une queue de distribution plus étalée.
- **activity_dynamics :** La distribution la plus étalée parmi ces cinq, bien que toujours asymétrique à droite. Il y a une plus grande variabilité dans la perception de la dynamique d'activité véhiculée par les icônes, avec un nombre non négligeable de posts obtenant des scores modérés.

Les boxplots groupés confirment ces tendances : pour frequency_sensationnel, emphasis_social, resonance_memory, et atmosphere_environment, les médianes sont basses et les boîtes sont relativement compactes vers le bas, avec des outliers vers le haut. activity_dynamics a une médiane légèrement plus élevée et une dispersion plus grande.

3.3.2. Corrélations entre les Métriques Iconographiques

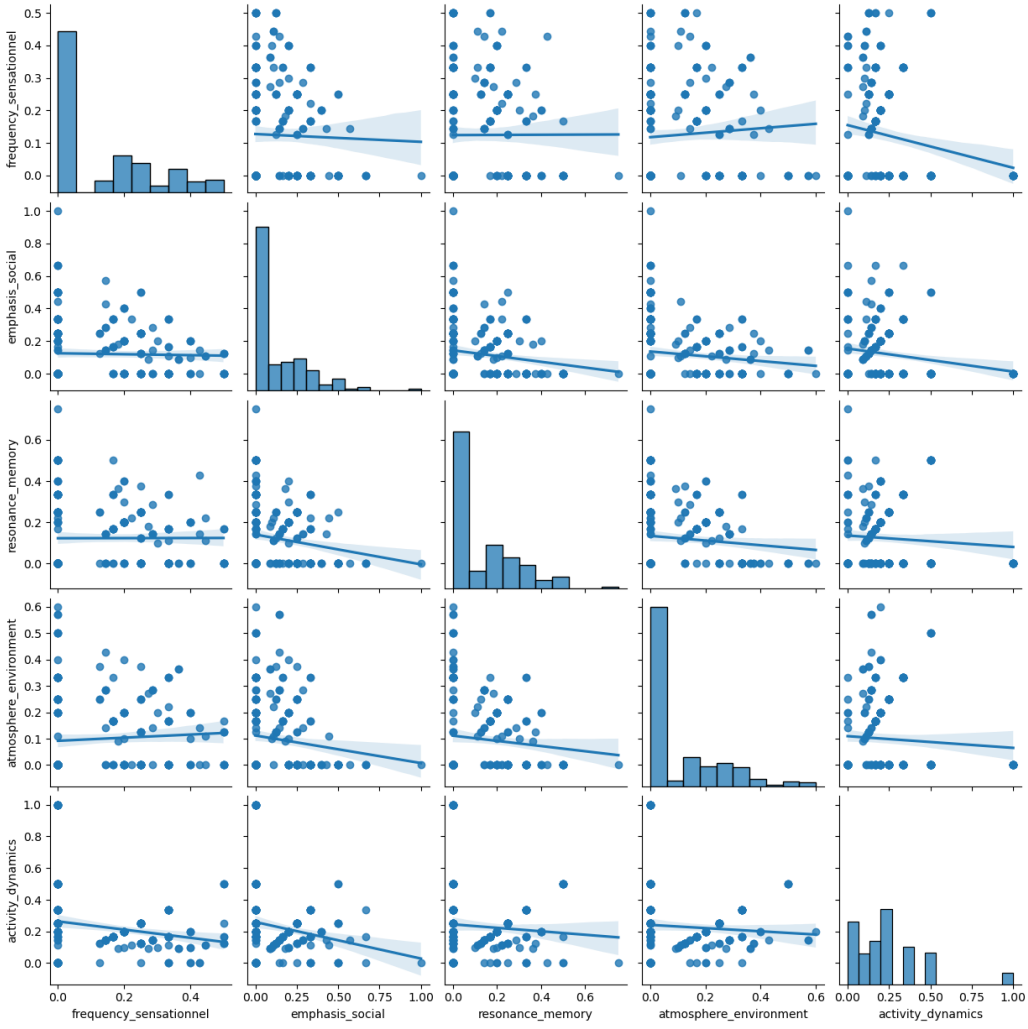


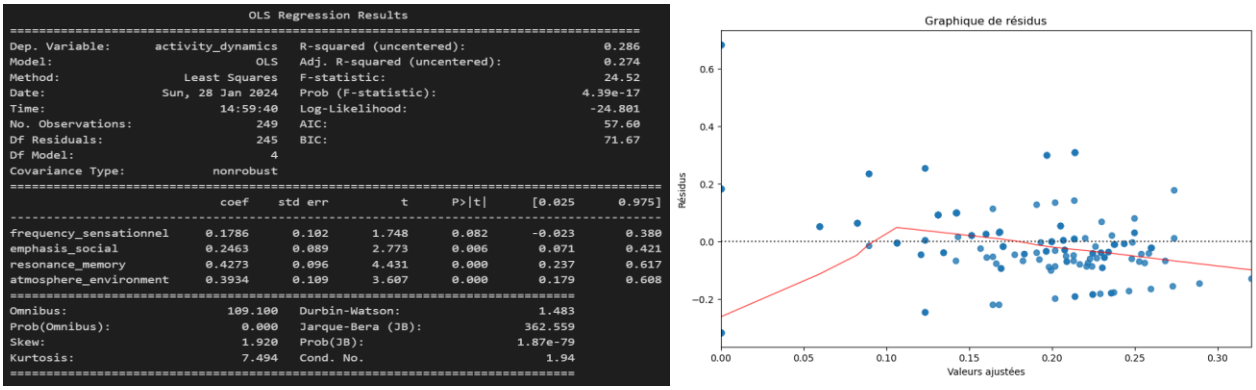
Figure 12. *Matrice de corrélation et de distribution des métriques iconographiques qualitatives. Les graphiques sur la diagonale représentent les histogrammes de distribution. Les graphiques hors diagonale montrent les nuages de points illustrant les relations par paires, avec une droite de régression linéaire superposée.*

Interprétation des Corrélations:

Nous pouvons tirer plusieurs observations sur les corrélations, que nous allons reprendre et analyser :

- **Efficacité Globale des Icônes (Efficiency Icons Used) et Variété des Types d'Icônes (Diversity Types Icons Used) :**
    - *Aucune corrélation significative n'a été observée avec les autres variables pour ces deux métriques.*
    - **Analyse :** Cela suggère que ni l'efficacité perçue des icônes ni la simple variété de leurs types ne sont liées de manière linéaire aux aspects plus qualitatifs comme l'emphase sociale, l'atmosphère, la dynamique, la résonance mémorielle ou la fréquence du sensationnel. L'efficacité et la diversité pourraient être des dimensions orthogonales à ces autres perceptions.
- **Accentuation Sociale (Emphasis Social) vs Resonance Memory et Activity Dynamics :**
    - *Présente une corrélation négative significative avec Resonance Memory (-0.16) et Activity Dynamics (-0.18). Cela suggère que lorsque l'accentuation sociale augmente, la résonance mémorielle et l'intensité visuelle de l'activité tendent à diminuer.*
    - **Analyse :** Ces corrélations négatives, bien que faibles (proches de -0.2), sont intéressantes. Elles pourraient signifier que les icônes qui soulignent fortement les aspects sociaux (par exemple, des icônes de personnes, de groupes, d'interactions) sont peut-être moins marquantes individuellement (moins de résonance mémorielle) ou perçues comme moins dynamiques. Peut-être que les icônes "sociales" sont plus génériques ou moins distinctives visuellement que celles qui évoquent une forte activité ou un souvenir précis.
- **Atmosphère Environnementale (Atmosphere Environment) :**
    - *Aucune corrélation significative n'a été identifiée avec d'autres variables.*
    - **Analyse :** La capacité d'une icône à dépeindre une ambiance semble être une caractéristique indépendante des autres dimensions iconographiques mesurées ici.
- **Dynamique d'Activité (Activity Dynamics) vs Emphasis Social :**
    - *Présente une corrélation négative significative avec Emphasis Social (-0.18).*"
    - **Analyse :** Confirmé. Une plus forte emphase sociale est liée à une perception de moindre dynamique d'activité.
- **Résonance Mémorielle (Resonance Memory) vs Emphasis Social :**
    - *Montre une corrélation négative significative avec Emphasis Social (-0.16).*
    - **Analyse :** Confirmé. Une plus forte emphase sociale est liée à une moindre résonance mémorielle.
- **Fréquence d'Éléments Sensationnels (Frequency Sensationnel) :**
    - *Aucune corrélation significative n'a été observée avec d'autres variables.*
    - **Analyse :** L'utilisation d'icônes à caractère "sensationnel" ne semble pas être liée linéairement aux autres aspects qualitatifs des icônes.

3.3.3. Modélisation de la Dynamique d'Activité (activity_dynamics) par Régression OLS



OLS Regression Results

Dep. Variable:	activity_dynamics	R-squared (uncentered):	0.286
Model:	OLS	Adj. R-squared (uncentered):	0.274
Method:	Least Squares	F-statistic:	24.52
Date:	Sun, 28 Jan 2024	Prob (F-statistic):	4.39e-17
Time:	14:59:40	Log-Likelihood:	-24.801
No. Observations:	249	AIC:	57.60
Df Residuals:	245	BIC:	71.67
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
frequency_sensationnel	0.1786	0.102	1.748	0.082	-0.023	0.380
emphasis_social	0.2463	0.089	2.773	0.006	0.071	0.421
resonance_memory	0.4273	0.096	4.431	0.000	0.237	0.617
atmosphere_environment	0.3934	0.109	3.607	0.000	0.179	0.608

Omnibus:	109.100	Durbin-Watson:	1.483
Prob(Omnibus):	0.000	Jarque-Bera (JB):	362.559
Skew:	1.920	Prob(JB):	1.87e-79
Kurtosis:	7.494	Cond. No.	1.94

Figure 13. *Graphique des résidus du modèle de régression OLS pour la Dynamique d'Activité des icônes.*

Interprétation des Résultats de la Régression:

Le modèle tente d'expliquer *activity_dynamics* à partir des valeurs des autres variables.

- **Ajustement du Modèle (R-carré) :** Le R-carré non centré est de 0.286 (R-carré ajusté non centré de 0.274). Cela signifie que le modèle explique environ 29% de la variance de activity_dynamics. C'est une part relativement faible, indiquant que d'autres facteurs importants ne sont pas pris en compte par ces prédicteurs.
- **Significativité Globale du Modèle (Statistique F) :** La statistique F est de 24.52 avec une p-value très faible (4.39e-17). Le modèle est donc globalement significatif.
- **Coefficients des Prédicteurs :**
    - **frequency_sensationnel :** Coefficient de 0.1786 (0.18 dans le texte). P-value = 0.082. Non statistiquement significatif au seuil strict de 0.05, c'est correct, bien que 0.082 soit parfois considéré comme marginalement significatif. L'effet positif suggère qu'une plus grande fréquence d'éléments sensationnels pourrait être liée à une plus grande dynamique d'activité, mais la relation n'est pas solidement établie dans cet échantillon.
    - **emphasis_social :** Coefficient de 0.2463 (0.25 dans le texte). P-value = 0.006. C'est statistiquement significatif. Une plus grande emphase sociale est associée à une augmentation de la dynamique d'activité. *Attention : cela contredit la corrélation négative simple observée précédemment (-0.18).* En régression multiple, les coefficients reflètent l'effet d'une variable en contrôlant les autres. Il est possible que la relation simple soit masquée ou modifiée en présence des autres variables.
    - **resonance_memory :** Coefficient de 0.4273 (0.43 dans le texte). P-value = 0.000. Très significatif. Une plus grande résonance mémorielle est fortement associée à une augmentation de la dynamique d'activité.
    - **atmosphere_environment :** Coefficient de 0.3934 (0.39 dans le texte). P-value = 0.000. Très significatif. Une plus forte capacité à dépeindre une atmosphère/environnement est également fortement associée à une augmentation de la dynamique d'activité.

- **Graphique des Résidus :** Les résidus semblent dispersés de manière relativement aléatoire autour de la ligne zéro, bien qu'il puisse y avoir une légère tendance à une plus grande variabilité pour les valeurs ajustées plus élevées (hétéroscédasticité potentielle), mais cela reste léger.

Conclusion :
L'utilisation des icônes dans MOBILES est variée. La plupart des annotations utilisent peu d'icônes à caractère "sensationnel" ou à forte "emphase sociale". Les corrélations simples entre les différentes dimensions iconographiques sont généralement faibles ou inexistantes, à l'exception de relations négatives entre l'emphase sociale et la résonance mémorielle/dynamique d'activité. Cependant, le modèle de régression multiple pour expliquer la activity_dynamics montre que, en contrôlant les autres facteurs, l'emphasis_social, la resonance_memory et l'atmosphere_environment ont toutes un impact positif et significatif sur la perception de dynamique. Cela souligne la complexité des interactions entre ces différentes facettes iconographiques et la différence entre les relations bivariées et multivariées. Le modèle n'explique cependant qu'une part limitée de la variance de la dynamique d'activité.

3.4. Analyse des Métriques Relatives à l'Utilisation des Tags

Les tags constituent un moyen important pour les utilisateurs de catégoriser, de décrire et d'ajouter une dimension sémantique à leurs annotations. Les métriques suivantes ont été développées pour analyser leur utilisation :

- **Indice d'Influence des Tags**: Mesure l'influence globale des tags dans une annotation.
- **Indice d'Interaction Sociale pour les Tags**: Évalue l'importance des interactions sociales véhiculées par les tags.
- **Ratio de Sentiment Positif**: Proportion de tags exprimant un sentiment positif.
- **Indice Esthétique des Tags**: Quantifie l'attention portée à l'esthétique à travers les tags.
- **Ratio de Sentiment Négatif**: Proportion de tags exprimant un sentiment négatif.

3.4.1. Distributions des Métriques liées aux Tags

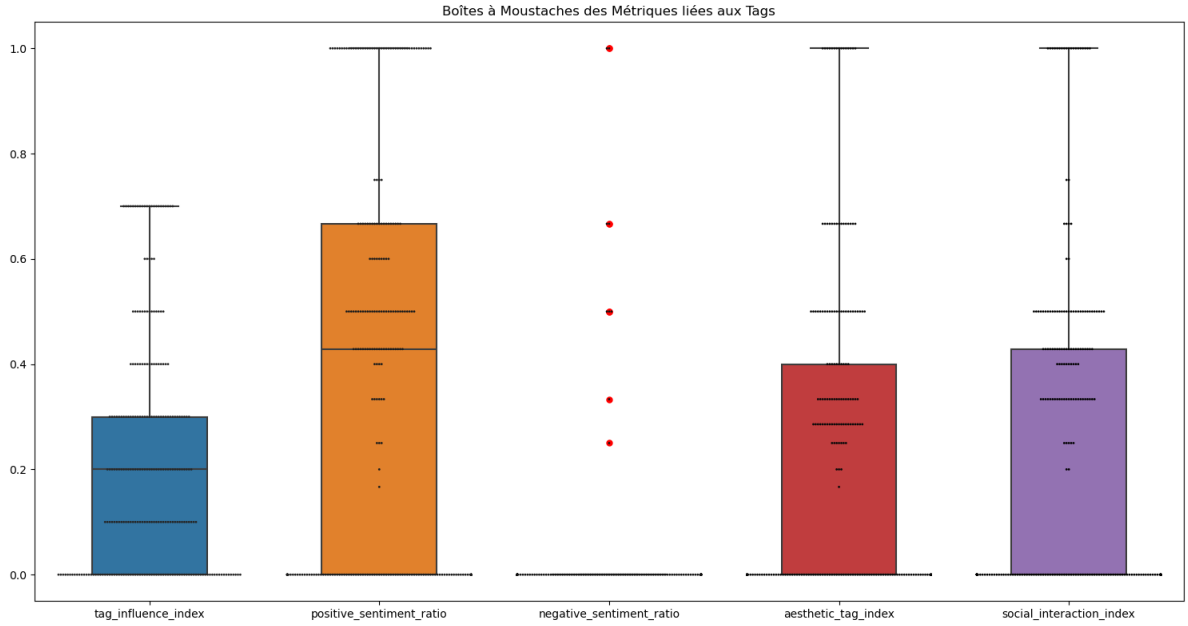


Figure 14. *Boxplots comparatifs des métriques liées à l'utilisation des tags. Illustre la dispersion des différentes variables.*

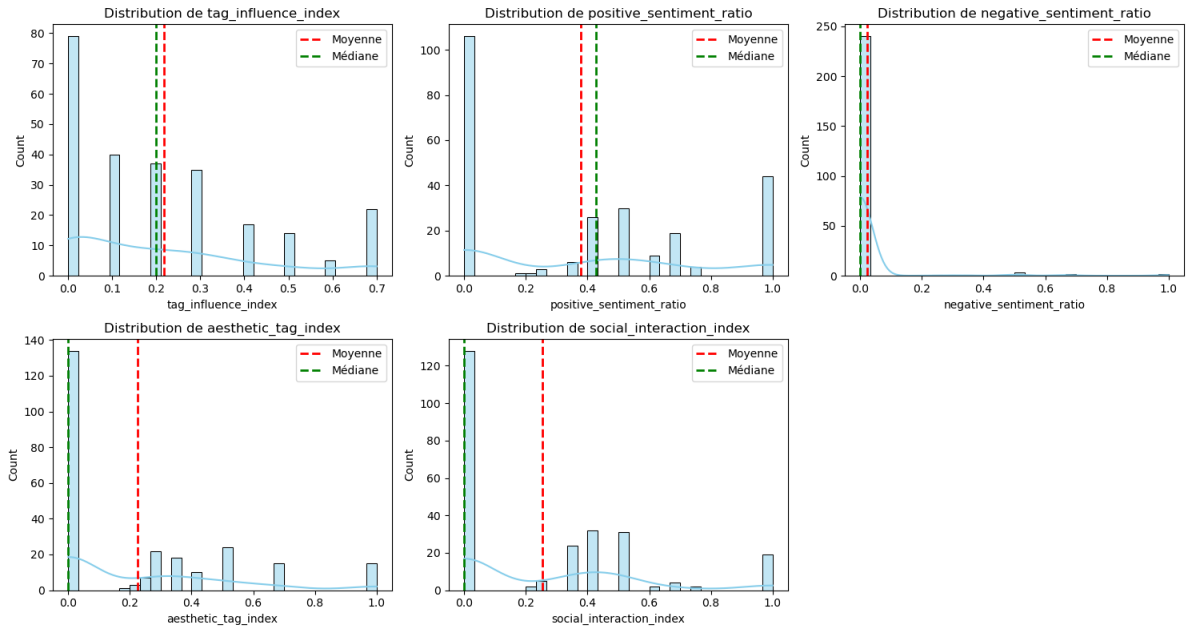


Figure 15. *Histogrammes de distribution des métriques liées à l'utilisation des tags. Montre la fréquence des différentes valeurs pour chaque métrique de tag. Les lignes pointillée et tiretée représentent respect. la moyenne et la médiane.*

Interprétation des Visualisations :

- **tag_influence_index :** L'histogramme montre une distribution avec plusieurs pics, notamment un pic important pour les faibles valeurs (proches de 0.1-0.2) et un autre pour des valeurs plus élevées (autour de 0.5-0.6). Cela suggère des comportements d'utilisation des tags variés : certains utilisateurs utilisent peu de tags par rapport au maximum autorisé, tandis que d'autres en utilisent un nombre plus conséquent. Le boxplot confirme cette dispersion avec une médiane autour de 0.2-0.3 et des outliers vers le haut.
- **positive_sentiment_ratio :** La distribution est fortement concentrée à 0 et à 1, avec très peu de valeurs intermédiaires. Cela indique que pour les tags à connotation émotionnelle, les utilisateurs ont tendance soit à n'utiliser que des tags positifs (ratio de 1), soit aucun tag positif (ratio de 0, ce qui pourrait signifier des tags neutres ou uniquement négatifs), plutôt que des mélanges. Le boxplot montre une médiane à 0.5, ce qui est un peu trompeur vu la nature bimodale (ou trimodale si on inclut les valeurs intermédiaires rares) de la distribution. La boîte est large, reflétant cette polarisation.
- **negative_sentiment_ratio :** L'histogramme montre une distribution très fortement concentrée à 0. La grande majorité des posts n'ont aucun tag à sentiment négatif. Les occurrences de tags négatifs sont rares. Le boxplot confirme cela avec une médiane à 0 et une boîte très écrasée en bas, avec quelques outliers positifs.
- **aesthetic_tag_index :** La distribution est également asymétrique à droite, avec un pic important à 0, indiquant que les tags à connotation esthétique ne sont pas utilisés dans un grand nombre d'annotations. Cependant, il y a une queue de distribution plus étalée que pour le sentiment négatif, suggérant une utilisation plus fréquente, bien que minoritaire, de tags esthétiques. Le boxplot a une médiane basse mais une dispersion plus visible que pour le sentiment négatif.
- **social_interaction_index :** Similaire à l'indice esthétique, la distribution est concentrée à 0, mais avec une présence notable de valeurs non nulles, indiquant que les tags à connotation sociale sont utilisés, mais pas majoritairement. Le boxplot reflète une médiane basse avec une dispersion.

Les boxplots groupés résument bien ces tendances : negative_sentiment_ratio est le plus concentré vers le bas. positive_sentiment_ratio est très dispersé en raison de sa nature bimodale. Les autres indices montrent des médianes basses mais une certaine variabilité.

3.4.2. Corrélations entre les Métriques Liées aux Tags

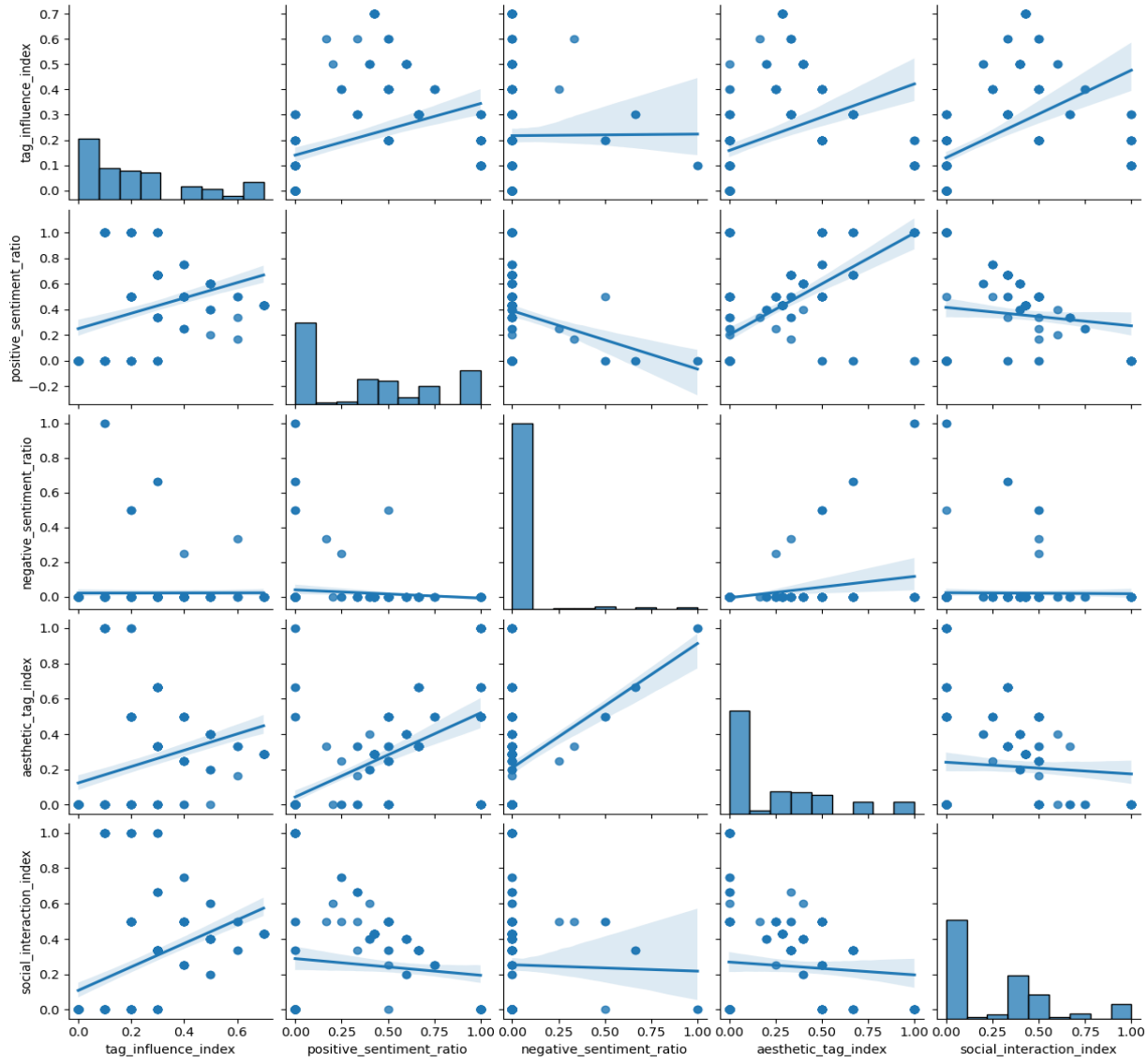
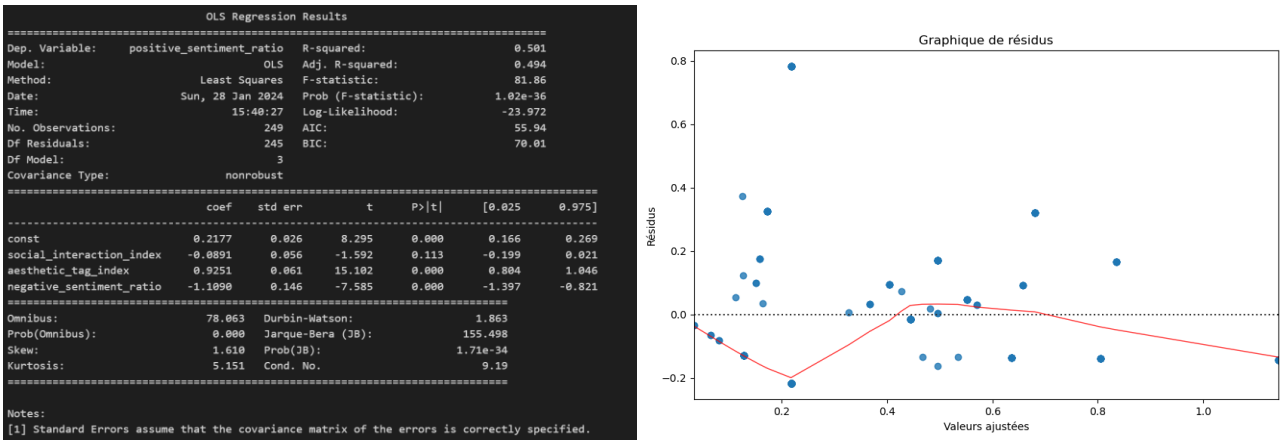


Figure 16. *Matrice de corrélation et de distribution des métriques liées à l'utilisation des tags. Les graphiques sur la diagonale représentent les histogrammes de distribution. Les graphiques hors diagonale montrent les nuages de points illustrant les relations par paires, avec une droite de régression linéaire superposée.*

- Aucune relation statistiquement significative n'a été mise en évidence entre l'Indice d'Influence des Tags et le Ratio de Sentiment Négatif à un niveau de confiance de 95%, suggérant que l'influence des tags n'est pas associée à une tonalité négative. L'influence globale des tags (leur nombre relatif) n'est pas liée à la proportion de tags négatifs. Utiliser beaucoup de tags ne signifie pas qu'on utilise plus (ou moins) de tags négatifs.

- Aucune corrélation statistiquement significative n'a été identifiée entre le Ratio de Sentiment Positif et l'Indice d'Interaction Sociale pour les Tags à un niveau de confiance de 95%, indiquant que la positivité des sentiments n'est pas nécessairement liée à une forte composante sociale. La positivité des sentiments exprimés par les tags n'est pas liée à leur connotation sociale. Des tags positifs peuvent être non sociaux et vice-versa.
- Aucune corrélation statistiquement significative n'a été constatée entre le Ratio de Sentiment Négatif et l'Indice d'Interaction Sociale pour les Tags à un niveau de confiance de 95%, suggérant que la négativité des sentiments ne prédomine pas dans les interactions sociales au sein des tags. la négativité des sentiments n'est pas liée à la dimension sociale des tags.
- Une corrélation positive significative de 0.35 entre l'Indice d'Influence des Tags et le Ratio de Sentiment Positif à un niveau de confiance de 95% suggère que les tags influents sont souvent associés à des sentiments positifs.
- L'Indice d'Influence des Tags montre une corrélation positive significative de 0.35 avec l'Indice Esthétique des Tags à un niveau de confiance de 95%, indiquant qu'une plus grande influence des tags est liée à une composante esthétique plus marquée.
- Une corrélation positive significative de 0.48 entre l'Indice d'Influence des Tags et l'Indice d'Interaction Sociale pour les Tags à un niveau de confiance de 95% suggère que les tags influents sont souvent associés à des interactions sociales importantes.
- Le Ratio de Sentiment Positif montre une corrélation significative de 0.61 avec l'Indice Esthétique des Tags à un niveau de confiance de 95%, indiquant que les sentiments positifs sont souvent liés à une dimension esthétique prononcée. C'est une corrélation forte. Les tags exprimant un sentiment positif sont très souvent ceux qui ont également une connotation esthétique. Cela suggère que l'appréciation esthétique est une source majeure de sentiment positif dans les tags.

3.4.3. Modélisation du Ratio de Sentiment Positif (positive_sentiment_ratio) par Régression OLS



OLS Regression Results

Dep. Variable:	positive_sentiment_ratio	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.494
Method:	Least Squares	F-statistic:	81.86
Date:	Sun, 28 Jan 2024	Prob (F-statistic):	1.02e-36
Time:	15:40:27	Log-Likelihood:	-23.972
No. Observations:	249	AIC:	55.94
Df Residuals:	245	BIC:	70.01
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2177	0.026	8.295	0.000	0.166	0.269
social_interaction_index	-0.0891	0.056	-1.592	0.113	-0.199	0.021
aesthetic_tag_index	0.9251	0.061	15.102	0.000	0.804	1.046
negative_sentiment_ratio	-1.1090	0.146	-7.585	0.000	-1.397	-0.821

Omnibus:	78.063	Durbin-Watson:	1.863
Prob(Omnibus):	0.000	Jarque-Bera (JB):	155.498
Skew:	1.610	Prob(JB):	1.71e-34
Kurtosis:	5.151	Cond. No.	9.19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

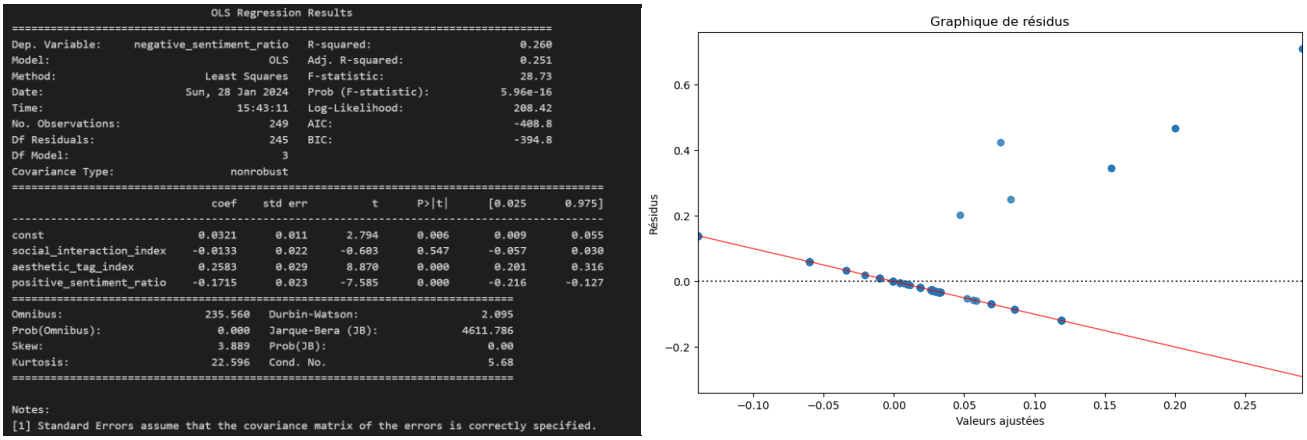
Figure 17. *Graphique des résidus du modèle de régression OLS pour le Ratio de Sentiment Positif des tags.*

Interprétation des Résultats

Le modèle tente d'expliquer positive_sentiment_ratio à partir de social_interaction_index, aesthetic_tag_index, et negative_sentiment_ratio.

- **Ajustement du Modèle (R-carré) :** R-carré de 0.501 (R-carré ajusté de 0.494). Le modèle explique environ 50% de la variance du ratio de sentiment positif. C'est un ajustement modéré à bon.
- **Significativité Globale du Modèle (Statistique F) :** F est de 81.86 avec une p-value très faible (1.02e-36). Le modèle est globalement très significatif.
- **Coefficients des Prédicteurs :**
    - **social_interaction_index :** Coefficient de -0.0891. P-value = 0.113. Peu d'impact. Au seuil de 0.05, il n'est pas statistiquement significatif. Une fois les autres variables prises en compte, l'aspect social des tags ne semble pas influencer significativement la proportion de tags positifs.
    - **aesthetic_tag_index :** Coefficient de 0.9251. P-value = 0.000. Très significatif et de forte magnitude. Une augmentation de l'indice esthétique est fortement associée à une augmentation substantielle du ratio de sentiment positif.
    - **negative_sentiment_ratio :** Coefficient de -1.1090. P-value = 0.000. Très significatif et de forte magnitude négative. Comme attendu, une augmentation de la proportion de tags négatifs est fortement associée à une diminution de la proportion de tags positifs (ils sont en quelque sorte mutuellement exclusifs si un tag ne peut être à la fois positif et négatif, ou si les utilisateurs ne mélangent pas beaucoup).
- **Graphique des Résidus:** La dispersion des résidus semble acceptable, bien qu'il puisse y avoir quelques motifs (les points ne sont pas parfaitement aléatoires), ce qui pourrait suggérer que d'autres facteurs ou des relations non linéaires pourraient être en jeu.

3.4.4. Modélisation du Ratio de Sentiment Négatif (negative_sentiment_ratio) par Régression OLS



OLS Regression Results

Dep. Variable:	negative_sentiment_ratio	R-squared:	0.260
Model:	OLS	Adj. R-squared:	0.251
Method:	Least Squares	F-statistic:	28.73
Date:	Sun, 28 Jan 2024	Prob (F-statistic):	5.96e-16
Time:	15:43:11	Log-Likelihood:	208.42
No. Observations:	249	AIC:	-408.8
Df Residuals:	245	BIC:	-394.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0321	0.011	2.794	0.006	0.009	0.055
social_interaction_index	-0.0133	0.022	-0.603	0.547	-0.057	0.030
aesthetic_tag_index	0.2583	0.029	8.870	0.000	0.201	0.316
positive_sentiment_ratio	-0.1715	0.023	-7.585	0.000	-0.216	-0.127

Omnibus:	235.560	Durbin-Watson:	2.095
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4611.786
Skew:	3.889	Prob(JB):	0.00
Kurtosis:	22.596	Cond. No.	5.68

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 18. *Graphique des résidus du modèle de régression OLS pour le Ratio de Sentiment Négatif des tags.*

Interprétation des Résultats de la Régression :

Le modèle tente d'expliquer negative_sentiment_ratio à partir de social_interaction_index, aesthetic_tag_index, et positive_sentiment_ratio.

- **Ajustement du Modèle (R-carré) :** R-carré de 0.260 (R-carré ajusté de 0.251). Le modèle explique seulement 26% de la variance. C'est un ajustement faible, suggérant que le modèle n'est "pas bien ajusté aux données".
- **Significativité Globale du Modèle (Statistique F) :** F est de 28.73 avec une p-value très faible (5.96e-16). Le modèle est globalement significatif, malgré son faible pouvoir explicatif.
- **Coefficients des Prédicteurs :**
    - **social_interaction_index :** Coefficient de -0.0133. P-value = 0.547. Non significatif. "Peu d'impact".
    - **aesthetic_tag_index :** Coefficient de 0.2583. P-value = 0.000. Significatif. Une augmentation de l'indice esthétique est associée à une augmentation du ratio de sentiment négatif. *C'est un résultat surprenant et contre-intuitif par rapport à la relation positive entre esthétique et sentiment positif.* Cela pourrait indiquer une interaction complexe ou un type spécifique de tags esthétiques qui peuvent aussi avoir une connotation négative (par exemple, "sombre", "mélancolique" peuvent être esthétiques mais négatifs). Ou alors, c'est un artefact dû au faible pouvoir explicatif global du modèle et à la présence d'autres variables.
    - **positive_sentiment_ratio :** Coefficient de -0.1715. P-value = 0.000. Significatif. Une augmentation de la proportion de tags positifs est associée à une diminution de la proportion de tags négatifs (cohérent).
- **Graphique des Résidus:** Les résidus sont très concentrés près de zéro pour les petites valeurs ajustées et plus dispersés pour les valeurs ajustées plus élevées, indiquant une hétéroscédasticité. Cela viole une des hypothèses de la régression OLS et peut affecter la fiabilité des tests de significativité.

Conclusion:

L'utilisation des tags est variée, avec une forte polarisation des sentiments (souvent tout positif ou tout négatif/neutre, et rarement négatif seul). L'influence des tags est positivement liée à leur positivité, leur esthétique et leur dimension sociale. Le sentiment positif est fortement prédit par l'esthétique des tags et l'absence de tags négatifs. La prédiction du sentiment négatif est plus complexe et moins bien expliquée par le modèle actuel, avec un rôle contre-intuitif pour l'indice esthétique qui mériterait une investigation plus poussée. Les problèmes avec le graphique des résidus pour le modèle du sentiment négatif suggèrent également que ce modèle pourrait ne pas être approprié ou nécessiterait des transformations des variables.

4. Modélisation Prédictive du Volume Lexical

Au-delà de l'analyse descriptive et corrélationnelle, cette section explore la capacité de diverses techniques de modélisation à prédire le volume_lexical des annotations. Le volume lexical, en tant qu'indicateur de la richesse et de la longueur du contenu textuel, est une caractéristique importante des contributions. Comprendre les facteurs qui l'influencent et être capable de le prédire pourrait avoir des implications pour, par exemple, l'identification automatique de posts potentiellement riches ou la stimulation de contributions plus détaillées.

4.1. Objectif et Choix des Modèles de Régression

L'objectif principal de cette analyse est de développer et d'évaluer la performance de plusieurs modèles de régression pour prédire la variable cible volume_lexical en se basant sur un ensemble de caractéristiques extraites des annotations.
Le choix des modèles évalués couvre un spectre de complexité et d'approches :

1. **Régression Linéaire :** Modèle simple et interprétable, servant de référence. Il est pertinent si les relations entre les prédicteurs et le volume lexical sont approximativement linéaires.
2. **Arbre de Décision :** Capable de modéliser des relations non linéaires et des interactions complexes par des règles de décision séquentielles. Utile si des seuils dans les prédicteurs influencent le volume lexical.
3. **Forêt Aléatoire :** Modèle d'ensemble basé sur de multiples arbres de décision. Il améliore généralement la stabilité et la performance par rapport à un arbre unique et réduit le risque de surajustement. Efficace pour des données complexes avec de nombreux prédicteurs.
4. **Boosting de Gradient :** Autre modèle d'ensemble qui construit des modèles séquentiellement, chacun corrigeant les erreurs du précédent. Souvent très performant pour capturer des relations subtiles.
5. **Réseau de Neurones :** Modèle non linéaire puissant, capable d'apprendre des représentations complexes et des relations hautement non linéaires, mais nécessitant généralement plus de données et un réglage fin.

Cette sélection permet de comparer des approches allant de la plus simple à la plus sophistiquée pour déterminer le niveau de complexité nécessaire pour modéliser efficacement la relation entre les caractéristiques des annotations et leur volume lexical.

4.2. Variables Explicatives Utilisées

Les variables explicatives (prédicteurs) utilisées pour ces modèles incluent un large éventail de métriques précédemment définies, couvrant :
L'importance sociale des icônes (emphasis_social), la diversité des types d'icônes (diversity_types_icons_used), l'environnement atmosphérique véhiculé (atmosphere_environment), la dynamique de l'activité (activity_dynamics), la résonance mémorielle (resonance_memory), la sensation fréquentielle (frequency_sensationnel), l'efficacité des icônes (efficiency_icons_used), l'expressivité émotionnelle (emoticonality), l'indice d'interaction sociale des tags (social_interaction_index), le ratio de sentiment positif des tags (positive_sentiment_ratio), l'indice esthétique des tags (aesthetic_tag_index), le ratio de sentiment négatif des tags (negative_sentiment_ratio), l'indice d'influence des tags (tag_influence_index), la présence d'une composante graphique (graphical), ainsi que d'autres indicateurs contextuels non explicitement listés ici mais potentiellement inclus dans l'analyse originale.

4.3. Évaluation des Modèles et Résultats

Les modèles sont évalués sur un ensemble de test en utilisant deux métriques principales :

- **Erreur Quadratique Moyenne (MSE) :** Mesure la moyenne des carrés des erreurs de prédiction. Une valeur plus faible indique un meilleur ajustement.

- **Coefficient de Détermination (R-carré) :** Indique la proportion de la variance de la variable cible (volume lexical) qui est expliquée par les prédicteurs du modèle. Une valeur plus proche de 1 indique un meilleur ajustement.

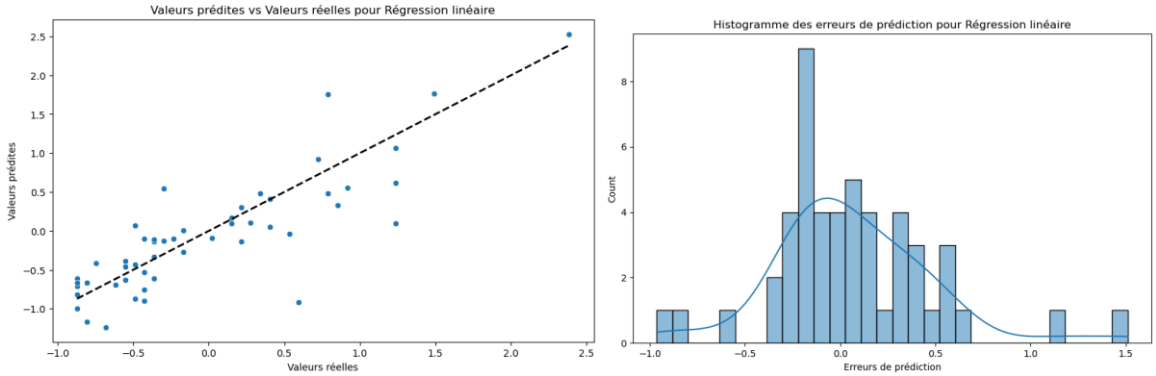


Figure 19. *Performance de la Régression Linéaire pour la prédiction du Volume Lexical.*

Arbre de décision : *Erreur quadratique moyenne : 0.6505473425627639 – R-carré : -0.2153459693994062*

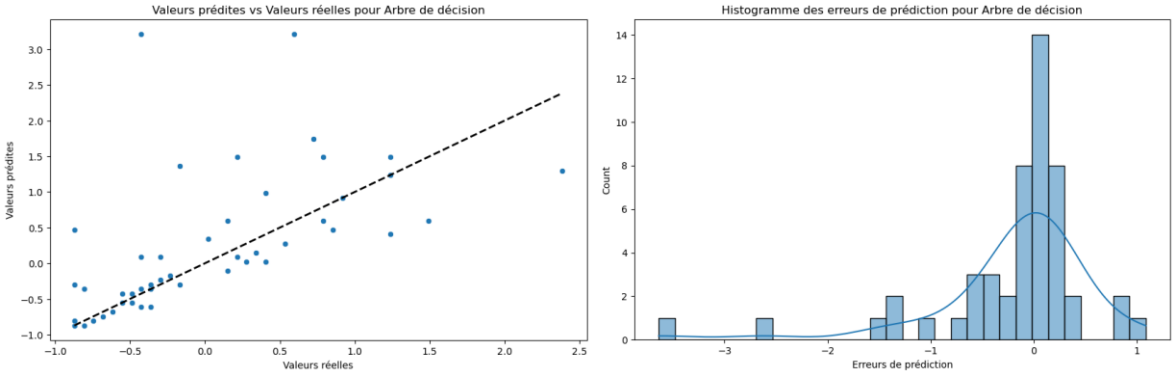


Figure 20. *Performance de l'Arbre de Décision Linéaire pour la prédiction du Volume Lexical.*

Forêt aléatoire : *Erreur quadratique moyenne : 0.17966527266736446 – R-carré : 0.6643511537267404*

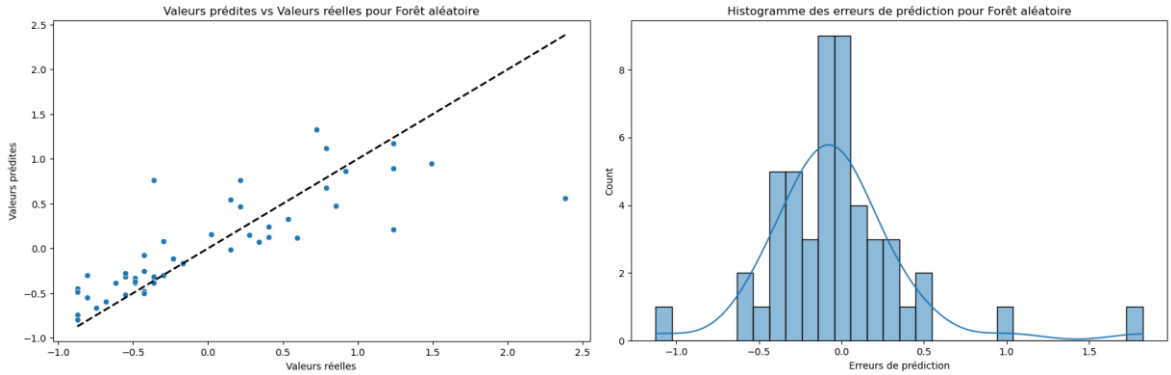


Figure 21. *Performance de la méthode Forêt Aléatoire pour la prédiction du Volume Lexical.*

Boosting de gradient : *Erreur quadratique moyenne : 0.18644800355406396 –R-carré : 0.6516797244465942*

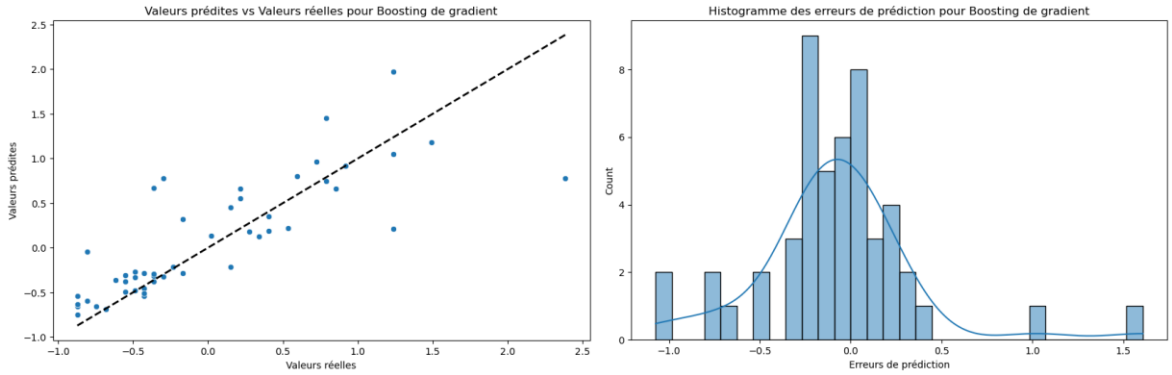


Figure 22. *Performance de la Boosting de gradient Linéaire pour la prédiction du Volume Lexical.*

Réseau de neurones : *Erreur quadratique moyenne : 0.352798649314723 05 – R-carré : 0.3409051295712028*

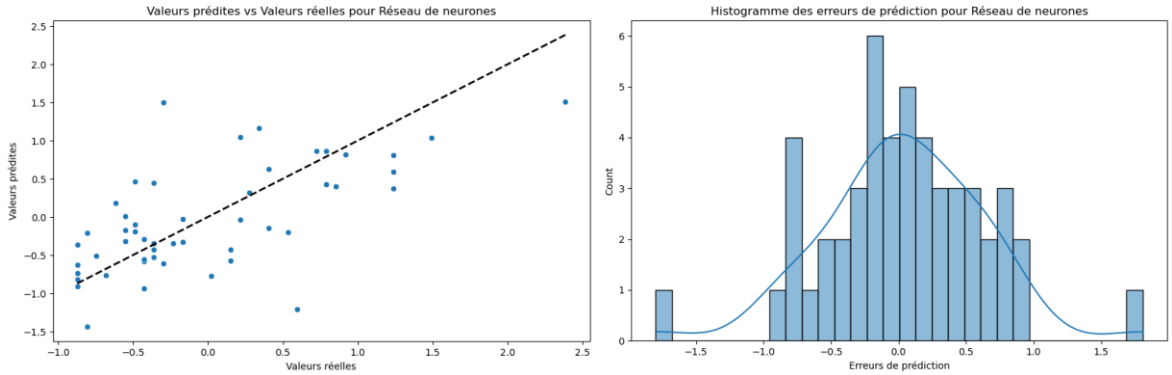


Figure 23. *Performance des Réseaux de Neurones pour la prédiction du Volume Lexical.*

Comparaison

```
     Régression linéaire  Arbre de décision  Forêt aléatoire  \
MSE             0.176985           0.650547         0.179665
R2              0.669359          -0.215346         0.664351

     Boosting de gradient  Réseau de neurones
MSE              0.186448            0.352799
R2               0.651680            0.340905

Le meilleur modèle basé sur l'erreur quadratique moyenne est : Régression linéaire
Le meilleur modèle basé sur le R-carré est : Régression linéaire
```

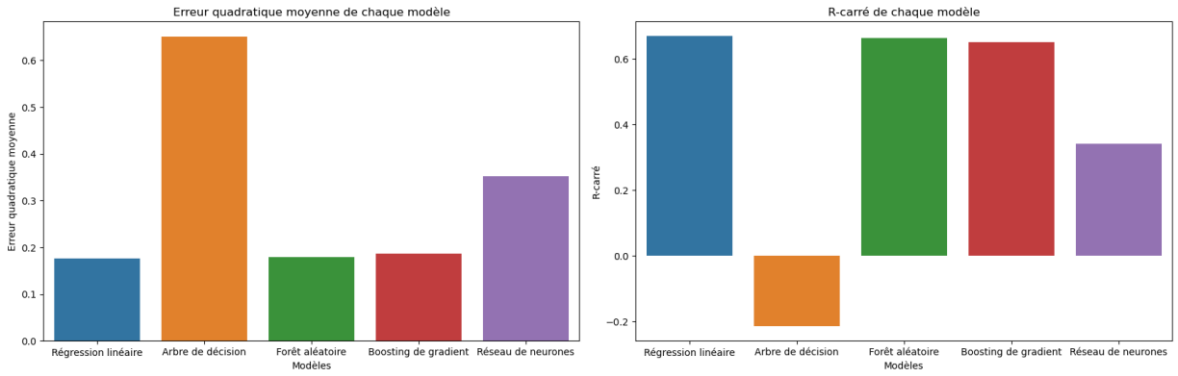


Figure 24. *Comparaison des performances des modèles de prédiction du Volume Lexical. Gauche : Erreur quadratique moyenne pour chaque modèle testé. Droite : Coefficient de détermination (R-carré) pour chaque modèle testé.*

Interprétation des Performances des Modèles:

1. **Régression Linéaire :**
    - MSE : 0.17698
    - R-carré : 0.66936
    - **Analyse :** La régression linéaire obtient le **meilleur R-carré** (environ 67% de la variance expliquée) et la **plus faible MSE** parmi tous les modèles testés. Le graphique des valeurs prédites vs réellesmontre une bonne corrélation, bien que les prédictions semblent légèrement sous-estimer les très hauts volumes et surestimer les très bas volumes. L'histogramme des erreurs est centré autour de zéro et relativement symétrique, ce qui est un bon signe.
2. **Arbre de Décision :**
    - MSE : 0.65055
    - R-carré : -0.21535 (valeur négative très problématique)
    - **Analyse :** L'arbre de décision performe très mal. Un R-carré négatif indique que le modèle est pire qu'un simple modèle qui prédirait toujours la moyenne de la variable cible. Le graphique des valeurs prédites vs réelles montre une très faible corrélation et une structure en "bandes" typique des arbres de décision non optimisés. L'histogramme des erreurs est large et peu centré. Ce modèle est clairement inadapté ou mal configuré.
3. **Forêt Aléatoire :**
    - MSE : 0.17967
    - R-carré : 0.66435
    - **Analyse :** La forêt aléatoire performe presque aussi bien que la régression linéaire, avec un R-carré de 66.4% et une MSE très proche. Le graphique des valeurs prédites vs réelles est similaire à celui de la régression linéaire. C'est une performance solide.
4. **Boosting de Gradient :**

- MSE : 0.18645
- R-carré : 0.65168
- **Analyse :** Le boosting de gradient est légèrement en retrait par rapport à la régression linéaire et à la forêt aléatoire, mais offre toujours une performance respectable avec un R-carré de 65.2%. Les graphiques sont également comparables.

5. **Réseau de Neurones :**
   - MSE : 0.35280
   - R-carré : 0.34091
   - **Analyse :** Le réseau de neurones a la performance la plus faible après l'arbre de décision, expliquant seulement 34% de la variance. Le graphique des valeurs prédites vs réelles montre une plus grande dispersion. Cela pourrait être dû à un manque de données pour entraîner efficacement un modèle aussi complexe, à une architecture non optimale, ou à un besoin de plus de réglage des hyperparamètres.

Comparaison et Conclusion sur la Modélisation :

Le tableau récapitulatif et les graphiques à barres confirment clairement les observations :

- **Meilleur modèle basé sur la MSE : Régression Linéaire** (suivie de très près par la Forêt Aléatoire).
- **Meilleur modèle basé sur le R-carré : Régression Linéaire** (suivie de très près par la Forêt Aléatoire).

Le fait que le modèle le plus simple, la régression linéaire, surpasse (ou égale pratiquement) les modèles plus complexes comme les forêts aléatoires, le boosting de gradient, et surtout les réseaux de neurones, est un résultat notable. Cela suggère plusieurs interprétations possibles :

- Les relations sous-jacentes entre les métriques prédictives et le volume lexical sont majoritairement linéaires ou peuvent être bien approximées par des relations linéaires.
- Les modèles plus complexes pourraient souffrir de surajustement sur l'ensemble d'entraînement (si non évalués sur un ensemble de test distinct et suffisamment grand) ou n'ont pas été suffisamment optimisés (hyperparamètres, architecture pour le réseau de neurones).
- L'ensemble de données n'est peut-être pas assez grand ou ne contient pas assez de signaux non linéaires forts pour que les modèles complexes puissent pleinement démontrer leur supériorité.

Néanmoins, la capacité à expliquer environ deux tiers de la variance du volume lexical avec un modèle simple comme la régression linéaire (ou une forêt aléatoire) est un résultat encourageant.

5. Analyse de Clustering pour les Variables Sélectionnées

Après avoir exploré les caractéristiques individuelles des annotations et tenté de prédire leur volume lexical, cette section s'intéresse à une approche non supervisée : l'analyse de clustering. L'objectif est de regrouper les annotations en clusters (groupes) homogènes, où les annotations au sein d'un même cluster partagent des

caractéristiques similaires, tout en étant distinctes des annotations des autres clusters. Ce processus peut révéler des "types" ou des "profils" d'annotations qui n'étaient pas apparents au préalable.

5.1. Objectif et Méthodologie du Clustering

L'analyse de clustering vise à partitionner l'ensemble des observations (annotations, décrites par leurs métriques) en un certain nombre de groupes. Les étapes clés de cette méthodologie sont :

1. **Sélection des Caractéristiques :** Utilisation des métriques extraites précédemment comme base pour le clustering.
2. **Détermination du Nombre Optimal de Clusters :** Application de techniques heuristiques pour estimer le nombre de clusters le plus pertinent.
3. **Application de l'Algorithme de Clustering :** Utilisation de l'algorithme K-Means.
4. **Réduction de Dimensionnalité pour la Visualisation :** Emploi de l'Analyse en Composantes Principales (PCA) pour visualiser les clusters en 2D.
5. **Analyse Post-Clustering :** Identification des caractéristiques les plus importantes pour la formation des clusters.

5.2. Détermination du Nombre Optimal de Clusters (K)

La détermination du nombre approprié de clusters (K) est une étape cruciale en K-Means. Deux méthodes ont été utilisées :

Méthode du Coude (Elbow Method) : Cette méthode trace la somme des carrés des distances des points à leur centre de cluster le plus proche (inertie) en fonction du nombre de clusters (k). On recherche un "coude" dans la courbe : un point où l'ajout d'un nouveau cluster n'apporte plus une réduction significative de l'inertie.

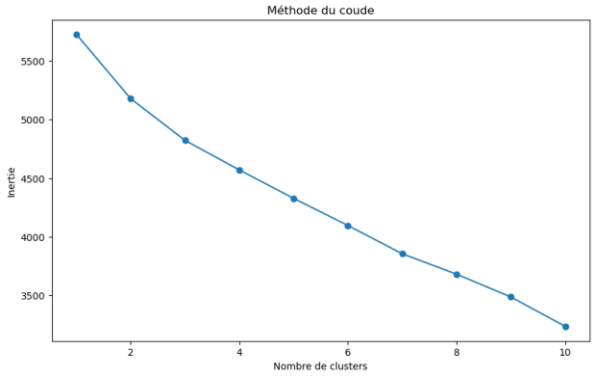


Figure 25. *Détermination du nombre optimal de clusters par la méthode du coude. Ce graphique montre l'évolution de l'inertie intra-cluster en fonction du nombre de clusters.*

Ce graphique trace l'inertie intra-cluster (*somme des carrés des distances des points à leur centroïde de cluster*) en fonction du nombre de clusters. On recherche un "*coude*" dans la courbe : un point où l'ajout d'un nouveau cluster n'apporte plus une réduction significative de l'inertie. La méthode du coude suggère 10 clusters. Visuellement, sur

le graphique, la courbe montre une descente rapide initiale, puis la pente s'adoucit. Il est plausible d'identifier un coude autour de K=3 ou K=4, après quoi la diminution de l'inertie devient plus graduelle.

Méthode de la Silhouette : Cette méthode mesure à quel point chaque point d'un cluster est similaire aux points de son propre cluster par rapport aux points des clusters voisins. Un score de silhouette élevé (proche de 1) indique une bonne cohésion intra-cluster et une bonne séparation inter-cluster. On recherche le nombre de clusters qui maximise le score de silhouette moyen.

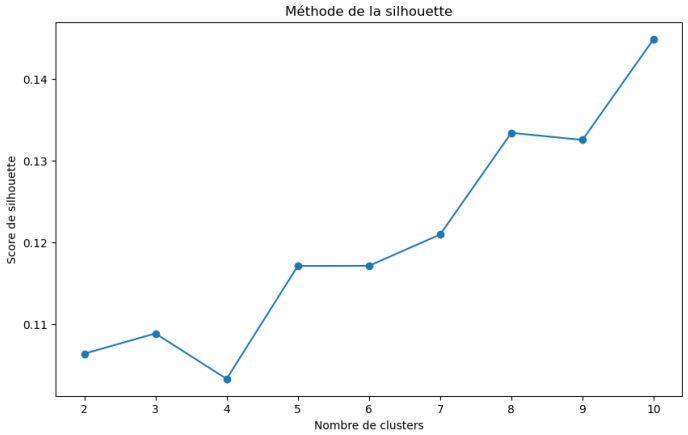


Figure 26. *Détermination du nombre optimal de clusters par la méthode de la silhouette. Ce graphique montre l'évolution du score de silhouette moyen en fonction du nombre de clusters.*

Interprétation : Cette méthode mesure à quel point chaque point est similaire à son propre cluster par rapport aux autres clusters. Un score de silhouette élevé indique une bonne cohésion intra-cluster et une bonne séparation inter-cluster. On cherche le nombre de clusters qui maximise le score de silhouette moyen. La méthode de la silhouette suggère 10 clusters. Le graphique montre que le score de silhouette atteint un pic à K=10, avec une valeur d'environ 0.145. Des scores de silhouette plus élevés sont généralement préférables (plus proches de 1). Un score de 0.145 est relativement bas, ce qui pourrait indiquer que les clusters ne sont pas très denses et/ou bien séparés, ou que la structure des clusters est complexe.

Conclusion : Après avoir utilisé la méthode du coude et la méthode de la silhouette pour évaluer différents nombres de clusters, nous avons obtenu deux suggestions pour le nombre optimal de clusters. La méthode du coude suggère 10 clusters, tandis que la méthode de la silhouette suggère 10 clusters. Ces deux méthodes nous donnent une idée du nombre de clusters à choisir, mais le choix final peut dépendre du contexte spécifique de notre analyse et de notre connaissance du domaine.

5.3. Réduction de Dimension avec PCA et Visualisation des Clusters

Pour visualiser les clusters dans un espace à deux dimensions, une *Analyse en Composantes Principales* (*PCA*) a été réalisée avant le clustering pour réduire les dimensions aux deux premières composantes principales.

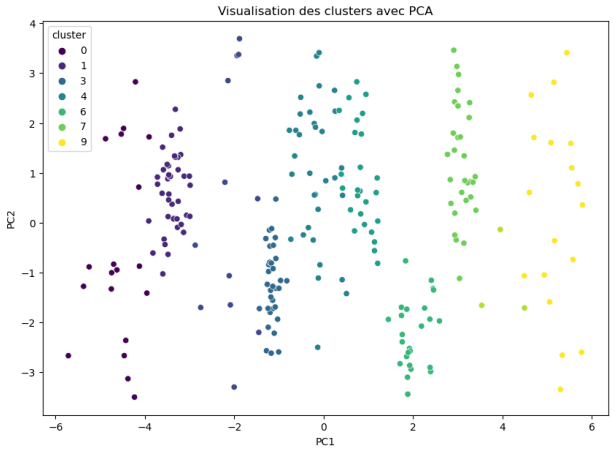


Figure 27. *Visualisation des clusters d'annotations après réduction de dimension par PCA. Ce graphique représente les annotations dans un espace à deux dimensions (PC1 et PC2), colorées selon le cluster auquel elles appartiennent.*

Interprétation de la Visualisation PCA :

Le graphique montre la projection des annotations (colorées par leur cluster assigné pour K=10) sur les deux premières composantes principales.

- **Séparation des Clusters :** On observe une certaine séparation entre certains groupes de couleurs, mais aussi un chevauchement considérable pour beaucoup d'entre eux. Par exemple, les clusters 0 (violet foncé), 1 (violet moyen), et 3 (bleu-vert) semblent occuper des régions distinctes dans le quadrant inférieur gauche, tandis que les clusters 4 (vert clair), 6 (jaune), et 7 (orange) se chevauchent davantage dans la partie droite du graphique. Le cluster 9 (rouge-brun) est plus diffus.
- **Densité des Clusters :** Certains clusters apparaissent plus compacts que d'autres.
- **Limites de la PCA :** Il est important de rappeler que la PCA ne conserve qu'une partie de la variance totale des données. La visualisation en 2D est une simplification et les clusters pourraient être mieux séparés dans l'espace de dimension supérieure original. Néanmoins, cette visualisation suggère que la structure des clusters est complexe et pas toujours nettement délimitée, ce qui est cohérent avec le score de silhouette modeste.

5.4. Importance des Caractéristiques pour la Formation des Clusters

Pour comprendre quelles métriques contribuent le plus à la différenciation des clusters, un classificateur RandomForest a été utilisé pour identifier les caractéristiques les plus importantes.

feature	importance
place-here	0.097385
place-close	0.092038
timing-UNIQUE	0.071023
timing-REGULAR	0.060943
timing-OCCASIONAL	0.060211
place-district	0.056050
positive_sentiment_ratio	0.054525
tag_influence_index	0.049966
diversity_types_icons_used	0.049123
efficiency_icons_used	0.048613
aesthetic_tag_index	0.044635
volume_lexical	0.044434
activity_dynamics	0.038843
place-tour	0.034456
place-city	0.031923
emoticonality	0.028209
resonance_memory	0.023982
emphasis_social	0.022406
graphical	0.021955
frequency_sensationnel	0.021551
social_interaction_index	0.021548
atmosphere_environment	0.013849
negative_sentiment_ratio	0.012331

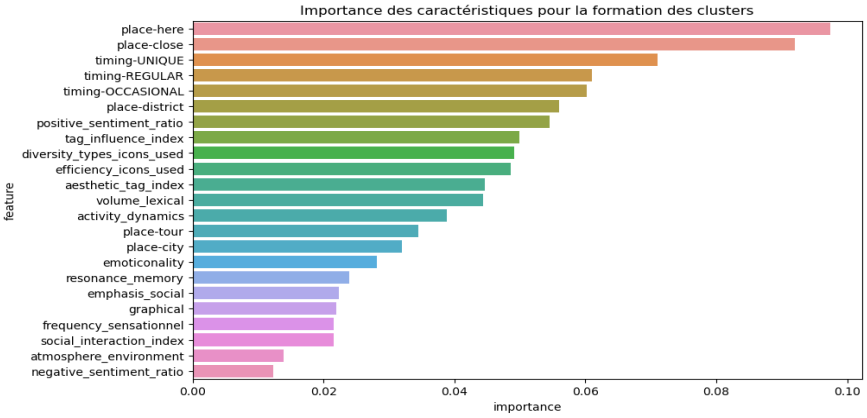


Figure 28. *Importance des caractéristiques pour la formation des clusters d'annotations. Ce graphique à barres horizontales montre le score d'importance de chaque métrique dans la discrimination des clusters, tel que déterminé par un classificateur RandomForest.*

Interprétation de l'Importance des Caractéristiques :

Selon le graphique et le tableau, les caractéristiques par ordre décroissant d'importance sont :

1. **place-here** (0.097)
2. **place-close** (0.092)
3. **timing-UNIQUE** (0.071)
4. **timing-REGULAR** (0.061)
5. **timing-OCCASIONAL** (0.060)
6. **place-district** (0.056)
7. **positive_sentiment_ratio** (0.055)
8. **tag_influence_index** (0.050)
9. **diversity_types_icons_used** (0.049)
10. **efficiency_icons_used** (0.049)
11. **aesthetic_tag_index** (0.045)
12. **volume_lexical** (0.044)

Analyse :

- **Dominance des Indicateurs Contextuels :** Les caractéristiques les plus discriminantes sont majoritairement liées au contexte de la publication : la localisation (place-here, place-close, place-district) et le timing/fréquence du post (timing-UNIQUE, timing-REGULAR, timing-OCCASIONAL). Cela suggère que les clusters pourraient fortement correspondre à des annotations faites "ici et maintenant" (uniques et localisées), des annotations sur des lieux proches, ou des annotations faites de manière régulière ou occasionnelle sur des types de lieux spécifiques.
- **Rôle des Sentiments et Tags :** Le positive_sentiment_ratio et le tag_influence_index suivent, indiquant que l'expression de sentiments positifs via les tags et l'utilisation plus ou moins importante de tags contribuent également à distinguer les groupes.

- **Importance Modérée des Icônes et du Texte :** Les métriques liées aux icônes (diversity_types_icons_used, efficiency_icons_used) et le volume_lexical apparaissent ensuite, avec une importance modérée.
- **Métriques Moins Influentes :** Les autres métriques (comme emoticonality, graphical, resonance_memory, etc., qui ont des scores d'importance plus bas dans le tableau) semblent moins contribuer à la définition des clusters identifiés.

Cela suggère que les "types" d'annotations qui émergent de ce clustering sont fortement influencés par *où* et *quand* (ou avec quelle fréquence perçue) elles sont faites, plus que par la finesse de leur contenu textuel intrinsèque (au-delà du volume) ou par certains aspects plus subtils de l'utilisation des icônes.

Conclusion pour la Section 5 :

L'analyse de clustering, en utilisant K-Means avec K=10 clusters (suggéré par les méthodes du coude et de la silhouette), a permis d'identifier des regroupements d'annotations. La visualisation PCA montre une certaine structure mais aussi des chevauchements, en accord avec un score de silhouette modeste. L'analyse de l'importance des caractéristiques révèle que les indicateurs de contexte spatial (proximité du lieu) et temporel/fréquentiel (caractère unique, régulier ou occasionnel du post) sont les plus déterminants dans la formation de ces clusters, suivis par l'expression de sentiments positifs via les tags et l'influence générale des tags. D'autres aspects comme le volume lexical et l'utilisation des icônes jouent un rôle secondaire. Ces résultats suggèrent que les profils d'annotations identifiés sont fortement ancrés dans le contexte situationnel de leur création.

6. Discussion Générale

L'ensemble des analyses menées sur les métriques d'annotations de l'application MOBILES fournit un éclairage multidimensionnel sur la nature des contributions des utilisateurs. Cette section vise à synthétiser les principaux enseignements, à discuter de la valeur de l'approche méthodologique adoptée, à reconnaître les limites de la présente étude et à envisager les implications pour la conception et l'évolution de MOBILES.

6.1. Synthèse des Principaux Enseignements sur les Annotations MOBILES

L'étude a révélé plusieurs caractéristiques et relations notables concernant les annotations :

- **Caractéristiques Textuelles :** Une majorité d'annotations sont concises et syntaxiquement simples, bien qu'une variabilité existe avec des contributions occasionnellement plus longues et complexes. Le volume, la diversité lexicale et la complexité syntaxique des annotations individuelles tendent à être positivement corrélés. Au niveau de l'utilisateur, la profondeur syntaxique cumulative et la diversité lexicale cumulative influencent de manière opposée l'"harmonie lexico-syntaxique", cette dernière étant modérément expliquée par ces deux facteurs.
- **Utilisation des Icônes :** L'usage des icônes est varié, avec une efficacité perçue et une fréquence d'éléments "sensationnels" généralement faibles. Des relations complexes existent entre les différentes dimensions iconographiques ; par exemple, l'emphase sociale, la résonance mémorielle et la perception de l'atmosphère contribuent positivement à la dynamique d'activité perçue des icônes en analyse multivariée, alors que certaines corrélations simples peuvent suggérer des liens différents.
- **Utilisation des Tags :** L'expression de sentiments négatifs via les tags est rare. L'influence des tags (leur nombre) est positivement corrélée à leur connotation positive, esthétique et sociale. L'esthétique des tags

est un fort prédicteur du sentiment positif. Un résultat plus surprenant est l'association positive entre l'indice esthétique et le ratio de sentiment négatif en régression multiple, suggérant des dynamiques émotionnelles potentiellement complexes véhiculées par les tags "esthétiques".

- **Prédiction du Volume Lexical :** Il est possible de prédire une part significative (environ 67% de la variance) du volume lexical des annotations en utilisant les métriques définies, la régression linéaire s'avérant être le modèle le plus performant parmi ceux testés dans cette étude, surpassant des modèles plus complexes comme les réseaux de neurones.
- **Clustering des Annotations :** Une structure de clustering en 10 groupes a été explorée. Bien que la séparation des clusters ne soit pas toujours très marquée (score de silhouette modeste), l'analyse a révélé que les caractéristiques spatio-temporelles (localisation, type de timing) sont les principaux discriminateurs des groupes, suivies par l'usage des tags et des icônes, puis par les aspects textuels.

6.2. Contribution de l'Approche par Métriques Augmentées

L'approche consistant à "augmenter" chaque annotation d'un riche ensemble de métriques calculées offre plusieurs avantages par rapport à des analyses plus traditionnelles :

- **Objectivation et Quantification :** Elle permet de passer d'une appréciation qualitative des contributions à une évaluation quantitative de multiples facettes (stylistique, sémantique, contextuelle, expressive).
- **Découverte de Relations Complexes :** Elle rend possible l'exploration statistique des interdépendances entre ces différentes facettes, révélant des relations qui ne seraient pas évidentes par simple observation.
- **Base pour la Modélisation :** Ces métriques servent de fondations solides pour des tâches de modélisation prédictive ou de segmentation non supervisée, comme illustré par la prédiction du volume lexical et le clustering.
- **Potentiel d'Automatisation :** Une fois définies, ces métriques peuvent être calculées automatiquement pour chaque nouvelle annotation, ouvrant la voie à des systèmes de feedback ou de personnalisation dynamiques.

6.3. Limites de l'Étude

Il est important de reconnaître les limites de cette analyse exploratoire :

- **Nature de l'Échantillon :** Les résultats sont spécifiques à l'échantillon d'annotations analysé. La généralisabilité à d'autres contextes ou à une population d'utilisateurs plus large nécessiterait des validations supplémentaires. La taille de l'échantillon (249 observations pour certaines régressions) peut également limiter la puissance statistique pour détecter des effets plus faibles ou la robustesse des modèles complexes.
- **Définition des Métriques :** Bien que conçues pour être informatives, certaines métriques sont des proxies ou des simplifications de concepts complexes (par exemple, "l'efficacité des icônes" ou "l'harmonie lexico-syntaxique"). Leur validité de construit pourrait être approfondie.

- **Modèles Explicatifs Partiels :** Les R-carrés obtenus dans les modèles de régression, bien que souvent significatifs, indiquent qu'une part non négligeable de la variance des variables cibles reste inexpliquée. D'autres facteurs, non inclus dans les métriques actuelles, jouent certainement un rôle.
- **Interprétation du Clustering :** Le score de silhouette modeste pour le clustering suggère que les groupes identifiés peuvent avoir des frontières floues. Une analyse qualitative du contenu des annotations au sein de chaque cluster est indispensable pour valider leur pertinence sémantique.
- **Causalité vs Corrélation :** Les analyses de corrélation et de régression identifient des associations statistiques, mais ne prouvent pas de relations de cause à effet.

6.4. Implications pour la Conception de MOBILES et les Systèmes Futurs

Les enseignements tirés de cette étude peuvent avoir plusieurs implications concrètes :

- **Amélioration de l'Interface de Contribution :** Comprendre ce qui caractérise les annotations (par exemple, la tendance à la concision) peut informer la conception des outils de création de posts pour encourager, si souhaité, des descriptions plus riches ou plus structurées.
- **Développement de Systèmes de Feedback :** Les métriques pourraient être utilisées pour fournir un feedback personnalisé aux utilisateurs sur la qualité ou l'impact potentiel de leurs contributions (par exemple, "Votre post pourrait être plus visible si vous ajoutiez des tags esthétiques").
- **Enrichissement des Systèmes de Recommandation :** Les profils d'annotations issus du clustering, ou les métriques individuelles, pourraient servir à affiner les recommandations de contenu ou à suggérer des types de contributions spécifiques aux utilisateurs. Par exemple, si un utilisateur crée souvent des posts dans un cluster "exploration de lieux proches avec peu de texte mais des icônes dynamiques", le système pourrait lui suggérer des lieux similaires à explorer.
- **Identification de Contenus de Haute Qualité ou Uniques :** Les métriques pourraient aider à identifier automatiquement des contributions particulièrement riches, originales ou engageantes pour les mettre en avant.

7. Conclusion et Travaux Futurs

7.1. Conclusion

Cette étude exploratoire a démontré la valeur de l'utilisation d'un ensemble diversifié de métriques pour analyser et caractériser les annotations utilisateurs au sein de l'application MOBILES. En quantifiant des aspects textuels, iconographiques, sémantiques liés aux tags, et contextuels, nous avons pu mettre en lumière des tendances de contribution, des interrelations significatives entre différentes facettes des posts, et identifier des profils d'annotations potentiels. Les analyses de régression ont montré qu'il est possible de modéliser certains aspects des annotations, comme leur volume lexical ou l'expression de sentiments, bien que le pouvoir explicatif des modèles varie. Le clustering a quant à lui révélé que le contexte spatio-temporel joue un rôle prépondérant dans la différenciation des types d'annotations. Bien que des limites aient été identifiées, notamment en ce qui concerne la robustesse de certains modèles ou la clarté de la structure de clustering, cette analyse fournit une base solide et de nombreuses pistes pour une compréhension plus approfondie des dynamiques de création de contenu dans MOBILES.

7.2. Travaux Futurs

Les résultats et les limites de cette étude ouvrent la voie à plusieurs perspectives de recherche et de développement :

1. **Affinement et Validation des Métriques :** Poursuivre le travail sur la définition et la validation des métriques, notamment celles dont l'interprétation est complexe ou dont la contribution explicative est faible. Des études qualitatives pourraient aider à mieux comprendre ce que certaines métriques capturent réellement du point de vue de l'utilisateur.
2. **Enrichissement des Modèles Prédictifs :** Explorer l'ajout de nouvelles variables explicatives (par exemple, des caractéristiques de l'utilisateur, des données d'interaction plus fines) pour améliorer la performance des modèles de régression et de prédiction.
3. **Analyse Qualitative des Clusters :** Mener une analyse de contenu détaillée des annotations appartenant à chaque cluster identifié afin de leur donner une signification sémantique plus riche et de valider leur pertinence.
4. **Études Longitudinales :** Examiner l'évolution des métriques et des profils d'annotation pour un même utilisateur au fil du temps, afin de comprendre les trajectoires d'apprentissage et d'engagement.
5. **Lien avec l'Impact des Annotations :** Corréler les métriques des annotations avec des indicateurs d'impact, tels que le nombre de vues, de "j'aime", de commentaires reçus, ou leur influence sur les parcours d'autres utilisateurs.
6. **Intégration Opérationnelle :** Explorer l'intégration de certaines de ces métriques dans l'application MOBILES pour fournir des outils d'auto-évaluation aux utilisateurs, ou pour alimenter directement des systèmes de recommandation ou de modération de contenu plus sophistiqués.
7. **Comparaison Inter-Groupes :** Si des données démographiques ou de profil utilisateur sont disponibles (par exemple, durée du séjour, nationalité, type d'études), analyser si les métriques d'annotation diffèrent significativement entre ces groupes.

En poursuivant ces pistes, il sera possible de tirer encore davantage de valeur des riches données de contribution générées par les utilisateurs de MOBILES, au bénéfice de l'amélioration continue de l'application et de l'expérience utilisateur.