Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Détection de l'hétérogénéité des messages et leur corrélation avec les profils d'auteurs sur les réseaux sociaux

Shami Thirion Sen[1, 2]
(1) Inalco, ERTIM, 2, rue de Lille, 75007 Paris, France
(2) Arlequin AI, Paris, France
shami.thirionsen@inalco.fr, shami@arlq.ai

RÉSUMÉ

Les réseaux sociaux tels que X (ex-Twitter) sont des plateformes incontournables de discussions des sujets sociaux et politiques. Dans la continuité d'études ayant démontré l'hétérogénéité des idéologies au sein d'un même mouvement politique, nous examinons un jeu de données comportant 7373 messages avec le slogan « *Nicolas qui paie* » issus de X sur les manifestations du 10 septembre 2025 en France. Notre objectif est de détecter l'hétérogénéité des messages publiés ainsi que leurs corrélations avec différents profils d'auteurs, identifiés via leurs biographies. Nous adoptons une approche non supervisée reposant sur la vectorisation et le *clustering* des messages, suivie d'une extraction des termes les plus spécifiques des clusters selon la méthodologie de Lafon et d'une annotation des clusters à l'aide du grand modèle de langue Claude 3.7 Sonnet. Nos premiers résultats révèlent qu'une forte homogénéité des messages sélectionnés entrave la détection de corrélations entre les types de messages et les profils d'auteurs les ayant publiés.

ABSTRACT

Detecting Heterogeneity in Social Media Posts and their Correlation with User Profiles

Social media platforms such as X (formerly Twitter) have now become established spaces for discussing social and political issues online. Studies analysing social media messages alongside author biographies have demonstrated that the same political movement can harbour a range of different political outlooks. Through our initial experiments on a dataset consisting of 7,373 X messages containing the term "*Nicolas Qui Paie*", we aim to detect potential heterogeneity within similar messages posted in relation to the protests held in France on September 10, 2025 and examine their correlation with the types of authors who posted them. We adopt an unsupervised approach combining data vectorization and clustering, followed by the extraction of the most specific terms for each cluster and finally by annotating these clusters using the LLM Claude 3.7 Sonnet. Our initial experiments reveal that our dataset being highly homogeneous, we cannot establish meaningful correlations between message and author types.

MOTS-CLÉS : Réseaux sociaux, détection d'hétérogénéité des données, analyse des profils d'utilisateurs, apprentissage non supervisé, *clustering*, annotation par LLM.

KEYWORDS: Social media, heterogeneity detection, account profiling, LLM annotations, unsupervised learning, clustering.

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 134–150.

1 Introduction

Au-delà des communications interpersonnelles, les réseaux sociaux, comme X, sont également des plateformes incontournables de publication sur des sujets sociaux et politiques (Vainio & Holmberg, 2017). L'analyse de ces phénomènes représente un enjeu important pour des tâches de veille informationnelle, sondages, recherche académique, dans les secteurs publics et privés. Le contenu des messages X fait objet de nombreux études en traitement automatique des langues (Golbeck *et al.*, 2011; Marchand, 2015; Dey *et al.*, 2017). De même, l'utilisation des biographies des auteurs est sollicitée pour des tâches de profilage des utilisateurs (Vainio & Holmberg, 2017; Ghosh *et al.*, 2022; Vandeweerdt *et al.*, 2026). Notre but ici est d'exploiter la combinaison des deux afin d'analyser les différences subtiles parmi des messages similaires, et d'identifier des correspondances éventuelles entre les catégories de messages et les profils d'utilisateurs.

Nous étudions ce phénomène à travers des expérimentations préliminaires sur un jeu de données portant sur l'appel aux manifestations politiques autour du 10 septembre 2025 [1] en France. Contrairement aux approches centrées sur la fouille d'opinion ou analyse de sentiment au sein des données web, notre approche repose sur la sélection d'un corpus composé d'échantillons lexicalement et sémantiquement très similaires. Cette démarche vise non seulement à examiner la diversité des messages publiés sur un même sujet ainsi que les caractéristiques des auteurs, observées à partir de leurs biographies, mais aussi l'étude de la corrélation entre les deux.

Ainsi, pour nos expérimentations, nous sélectionnons un corpus de 7 373 messages contenant l'expression « *Nicolas Qui Paie* » [2], le slogan associé au mouvement, critiquant la pression fiscale et évoquant la figure du contribuable français, parfois associée à des discours anti-immigration. Nous extrayons également les biographies des auteurs ayant publié ces messages. À la différence des approches supervisées dans la littérature (Fraisier, 2018; Chalehchaleh *et al.*, 2024), nous nous intéressons ici à des approches non supervisées, plus adaptées à la volumétrie et l'exploration de nos données non-annotées.

Notre méthodologie repose principalement sur la vectorisation des données textuelles (*embeddings*), leur projection en deux dimensions et l'application d'algorithmes de *clustering*. Afin d'interpréter qualitativement les *clusters* obtenus, nous mobilisons également le calcul de spécificité de Lafon (1980), ainsi qu'une annotation automatique réalisée à l'aide d'un grand modèle de langue (LLM).

2 État de l'art

Avec la popularisation des plateformes en ligne, les données issues des réseaux sociaux constituent une mine d'informations pour l'analyse des comportements des usagers dans des domaines divers tels que le ciblage publicitaire (Liu & Song, 2023) et politique (Boches & Cooney, 2023). Dans sa thèse doctorale, Fraisier (2018) élabore la méthodologie de détection de communauté sur les données issues de X (Twitter) sur les élections présidentielles 2017, en annotant les données ; Thonet (2017) exploite la détection de la fouille non-supervisée d'opinions multiples à partir des données web.

En outre, plusieurs études exploitent les messages ainsi que les éléments contextuels, tels que les biographies des auteurs (Dimitrov *et al.*, 2020; Chalehchaleh *et al.*, 2024) pour l'analyse des données

1. Mouvement du 10 septembre 2025
2. Slogan « Nicolas Qui Paie »

en ligne. Shuster *et al.* (2024) étudient l'hétérogénéité idéologique existant au sein d'un même mouvement politique, de manière non supervisée avec la vectorisation des textes (*embeddings*) et l'application de méthodes de *clustering*. À partir des messages publiés sur X par des supporters de Bernie Sanders lors des primaires démocrates de 2020 aux États-Unis, les auteurs montrent qu'un mouvement apparemment homogène peut en réalité recouvrir une pluralité de positions et de sensibilités.

Face à la diversité et à la volumétrie des messages, à l'émergence constante de nouveaux sujets dans l'industrie comme dans les secteurs publics, nous privilégions une approche non supervisée, à l'aide de la vectorisation des données avec des plongements lexicaux (*embeddings*) permettant une représentation numérique des données tout en gardant les relations syntaxiques et sémantiques, sans données préalablement annotées. Les modèles d'*embeddings* contextuels, tels que BGE M3 (Chen *et al.*, 2024), Qwen3 (Zhang *et al.*, 2025), optimisés pour la tâche d'extraction d'information à partir d'une base de données vectorielle (« *retrieval* » en anglais), permettent de vectoriser des unités textuelles de granularités variables : phrases, paragraphes, voire documents entiers, tout en conservant leurs relations sémantiques.

La visualisation des données vectorisées, à l'aide des méthodes de regroupement ou *clustering*, telles que K-Means, HDBSCAN (Campello *et al.*, 2013), permet d'explorer les regroupements des échantillons, et d'identifier d'éventuelles similarités au sein du corpus. Allant plus loin, l'algorithme H-NNE (Sarfraz *et al.*, 2022) permet une projection des données en deux dimensions et *clustering* hiérarchique simultané, tout en minimisant le coût de calcul. Basé sur des graphes des plus proches voisins, cet algorithme construit une structure hiérarchique des données et ajuste progressivement la projection en fonction de cette structure et ainsi ne nécessite pas de paramétrage tels que le nombre de clusters ou des seuils de densité, ce qui le rend particulièrement adapté à l'analyse de notre jeu de données.

S'inscrivant dans une logique similaire à celle de Shuster *et al.* (2024), nous nous intéressons à la détection automatique de l'hétérogénéité des points de vue au sein d'un même mouvement politique. Notre objectif est de nous appuyer sur les messages ainsi que l'auto-description des auteurs (biographies), issues de la plateforme X pour analyser les diversités observables à partir des *clusterings* des messages ainsi que leurs correspondances aux auteurs. Nous étudions notamment le rôle des biographies d'utilisateurs, conjointement aux messages publiés, comme source d'information pour caractériser cette diversité. Pour ce faire, nous effectuons la vectorisation des données ou *embeddings* avec le modèle BGE M3 (Chen *et al.*, 2024). Ensuite, nous effectuons la projection de nos données en deux dimensions, ainsi qu'un *clustering* hiérarchique simultané proposé par l'algorithme H-NNE Sarfraz *et al.* (2022), pour la visualisation, l'observation des structures globales et l'interprétation des relations entre les messages et les biographies des auteurs.

Pour compléter l'analyse qualitative de nos expérimentations, nous effectuons un calcul des termes spécifiques au sein des *clusters* des messages et des biographies, ainsi qu'une annotation automatique à l'aide du grand modèle de langue `Claude-3.7 Sonnet`. Malgré certaines limitations, l'utilisation des LLM pour des tâches d'annotation suscite un intérêt croissant en traitement automatique des langues (Tan *et al.*, 2024). Plusieurs travaux récents ont notamment mobilisé des modèles de la famille Claude pour l'analyse de contenus : le modèle `Claude-2` a été utilisé pour des tâches de la détection de la violation des droits de l'homme (Nemkova *et al.*, 2025), et `Claude 3.5 Sonnet` pour l'analyse des programmes politiques (Benoit *et al.*, 2025).

3 Méthodologie

Malgré l'avancée des modèles de langue actuels, le traitement du volume de données est un vrai défi pour des acteurs publics et privés. Dans ce contexte, les approches par apprentissage non supervisé sont de plus en plus sollicitées pour le traitement automatique des données. Ainsi, nous exploitons des approches de la vectorisation, la projection à deux dimensions et le *clustering* des données. Notre approche méthodologique est illustrée dans la Figure 1.

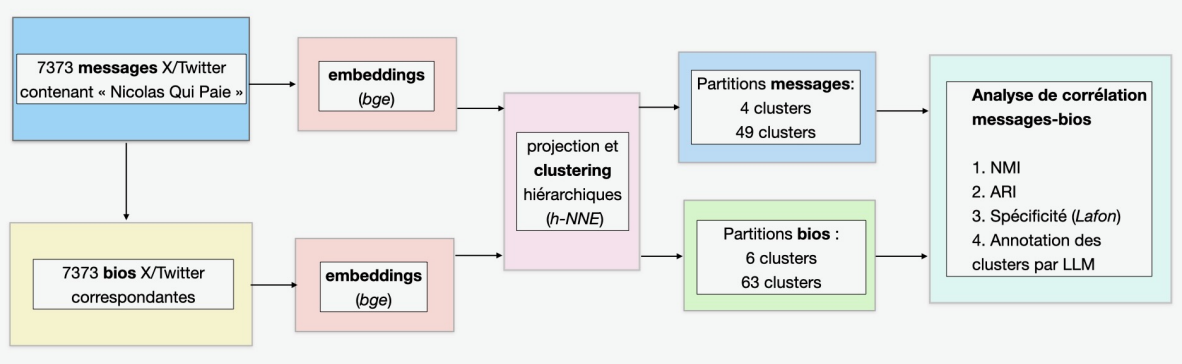

FIGURE 1 – Schéma de notre méthodologie

3.1 Jeu de données : « *10 septembre* »

Les réseaux sociaux comme X (Twitter) sont des plateformes pour la mobilisation des mouvements politiques. Nous exploitons ainsi un jeu de données portant sur l'appel à la manifestation pour le 10 septembre 2025 en protestation contre les décisions budgétaires publiques. Notre but est de détecter les différentes positions qui peuvent exister au sein du mouvement, ainsi qu'un profilage des personnes ayant fait appel pour cette manifestation à partir des messages publiés sur X et des biographies des auteurs de ces messages. Dans la Figure 2, nous illustrons une instance d'un message, ainsi que de la biographie de l'auteur du message.

Les données ont été collectées sur une période allant du 2 au 21 juillet 2025, soit environ 20 jours. Nous avons sélectionné pour nos expérimentations un corpus de 7 373 messages contenant le slogan principal du mouvement « Nicolas Qui Paie », filtrés à l'aide d'expressions régulières (`regex`). Ayant également pour but l'analyse des profils d'utilisateurs à partir de leurs biographies sur X, nous

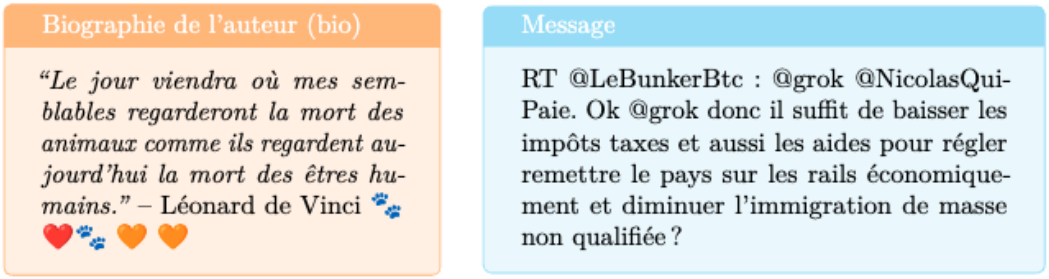

FIGURE 2 – Exemple : biographie (bulle orange) ; message publié (bulle bleue)

récupérons aussi les 7 373 biographies des auteurs (*user descriptions* en anglais) de ces messages, afin d'analyser comment se caractérisent les auteurs des messages sur le sujet. Nous effectuons un prétraitement minimal des textes, qui consiste principalement en la suppression des adresses électroniques et des hyperliens. Les données ont été obtenues depuis le prestataire des données Pulsar, en conformité avec les services de l'API de la plateforme X (ex-Twitter).

3.2 Vectorisation et *clustering*

Pour l'analyse non supervisée des messages et biographies, notre approche repose sur la vectorisation, permettant de représenter des unités textuelles de granularités variées tout en préservant leurs relations sémantiques. Nous utilisons le modèle de vectorisation BGE M3 (Chen *et al.*, 2024) qui permet de vectoriser des unités textuelles de granularités variables : phrases, émoticônes, paragraphes, voire documents entiers, tout en conservant leurs relations sémantiques. Nous exploitons également le regroupement et la visualisation reposant sur la réduction de dimensions et le *clustering* hiérarchique simultané avec l'algorithme h-NNE (Sarfraz *et al.*, 2022), dont l'originalité réside dans la production conjointe d'une représentation bidimensionnelle des données et d'une structure hiérarchique de regroupement à partir de l'espace vectoriel de grande dimension. Cette approche permet ainsi d'explorer les proximités sémantiques entre profils et messages tout en mettant en évidence des structures latentes à différentes échelles d'analyse.

Parmi les trois partitions de clusters obtenues par l'algorithme h-NNE, nous extrayons deux partitions pour chacun des ensembles *messages et biographies*, contenant respectivement 4 et 49 clusters de messages (voir la Figure 3), 6 et 63 clusters de biographies (voir la Figure 4). Ces deux partitions permettent d'une part de varier le nombre de messages par cluster, pour une meilleure comparaison des correspondances et d'autre part d'observer si les granularités différentes donnent lieu à des correspondances différentes. Les troisièmes partitions plus fines avec environ 758 (messages) et 1503 (biographies) clusters ont été écartées car trop fragmentées pour notre analyse actuelle.

3.3 Récupération et analyse des clusters messages et biographies

À la suite de la projection des données et le *clustering*, nous récupérons les deux partitions des messages et des biographies et nous extrayons les données appartenant à chacun des clusters afin d'effectuer des analyses quantitatives et qualitatives.

3.3.1 Recouvrement des clusters de messages et de biographies

La première étape à réaliser afin d'étudier la corrélation entre les messages et les biographies est le croisement des quatre partitions de *clusters* (voir la section 3.2). Cela nous permet d'observer les associations entre les clusters de messages et les clusters de biographies des auteurs. Plus précisément, nous examinons si les *clusters* produits par l'algorithme représentent des catégories distinctes des données : s'il existe une correspondance entre ces catégories et les types de biographies. Ces données nous permettent d'analyser l'hétérogénéité thématique des messages et des biographies des auteurs. Le croisement nous donne les 4 combinaisons présentées dans le Tableau 1.

3.3.2 Analyse globale quantitative : NMI et ARI

Dans le but de mesurer la correspondance entre les clusters de biographies et ceux des messages, nous utilisons deux métriques complémentaires : l'information mutuelle normalisée (NMI) et l'indice de Rand ajusté (ARI). Le NMI mesure la similarité structurelle globale entre deux partitions d'un même jeu de données (Huchet *et al.*, 2023). Dans notre cas, il permet d'évaluer dans quelle mesure la connaissance des clusters de biographies des utilisateurs renseigne sur les clusters de messages auxquels appartiennent ses messages publiés. Autrement dit, il quantifie le degré de dépendance entre la structure des groupes d'auteurs et celle des groupes de messages.

L'indice de Rand ajusté (*Adjusted Rand Index* ou ARI) mesure le degré d'accord entre deux partitions en comparant la manière dont les éléments sont regroupés par paires dans chacune de ces partitions. Il permet ainsi de vérifier si les mêmes utilisateurs tendent à être regroupés de façon similaire dans les clusters de biographies et dans ceux des messages. L'utilisation conjointe du NMI et de l'ARI permet donc d'évaluer la correspondance entre les deux partitions à la fois au niveau de leur structure globale et de la cohérence des regroupements d'éléments.

3.3.3 Spécificité (Lafon)

Afin de faciliter l'interprétation qualitative des clusters, nous utilisons la mesure de spécificité proposée par Lafon (1980). Cette mesure statistique compare la fréquence observée d'un terme dans un sous-corpus (ici, un cluster) à sa fréquence attendue dans l'ensemble du corpus, permettant d'identifier les termes significativement sur ou sous-représentés. Pour chaque cluster de messages et de biographies, nous extrayons les dix termes les plus spécifiques, qui constituent des indicateurs des thématiques ou caractéristiques dominantes du groupe. Dans le cas des biographies, qui peuvent contenir moins que dix termes, le nombre de termes spécifiques est le nombre tokens constituant la biographie.

3.3.4 Annotation des clusters par LLM

Enfin pour l'analyse des résultats à partir des termes spécifiques obtenus en 3.3.3, nous nous servons de l'annotation par un grand modèle de langue (LLM). Nous utilisons le modèle `claude-3-7-sonnet-latest` afin de générer des étiquettes pour chaque cluster. Nous fournissons au modèle les dix termes les plus spécifiques pour chaque cluster, avec un prompt. A travers nos prompts, nous effectuons deux configurations d'annotation : une première sans précision d'orientation politique, et une seconde intégrant des catégories d'orientation politique telles que gauche, droite et centre. Le choix de ces étiquettes est motivé par une analyse empirique des messages et des biographies présents dans le jeu de données.

Annotation simple Nous réalisons une annotation des *clusters* de messages et de biographies en plusieurs catégories de profils, notamment professionnels, politiques, culturels (liés à la musique ou au sport), patriotiques, ou philosophiques (citations, réflexions sur la liberté). Cette annotation est effectuée en *zéro-shot*, sans définition préalable des catégories dans le modèle. Nous assumons implicitement que le LLM est capable d'attribuer les catégories « culturelle » (musique, sport) et « philosophique » (citations, réflexions sur la liberté). Le choix des étiquettes résulte d'une appréciation

empirique fondée sur l'examen du corpus, qui demeure toutefois possiblement trop restreint pour permettre une catégorisation exhaustive. Nous présentons les étiquettes pour les biographies et les messages dans le Tableau 5, ainsi que le prompt dans le Tableau 4 en annexe.

Annotation avec des nuances politiques Nous réalisons une seconde annotation visant à intégrer des nuances politiques, en distinguant les catégories gauche, droite et centre. Cette annotation est effectuée sans fournir de définition explicite de ce qui caractérise la gauche, la droite ou le centre ; nous nous appuyons sur les connaissances intrinsèques du modèle de langage utilisé pour effectuer cette catégorisation. Nous présentons l'ensemble des étiquettes dans le Tableau 6 en annexe.

4 Analyse des résultats

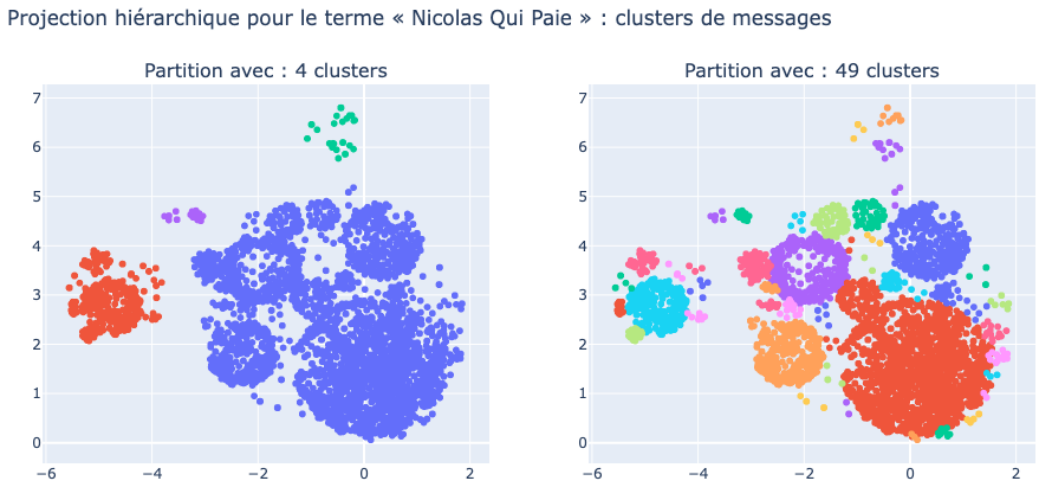

FIGURE 3 – Deux partitions de *clusters* **messages**. Partition 1 : 4 *clusters*, partition 2 : 49 *clusters*.

L'évaluation d'une analyse non supervisée en TAL constitue un défi important. Notre méthodologie d'évaluation repose sur deux approches principales. La première est quantitative, fondée sur le calcul de l'information mutuelle normalisée (NMI) et de l'indice de Rand ajusté (ARI). En parallèle, nous adoptons une approche qualitative, combinant le calcul de la spécificité (Lafon, 1980) et l'annotation par un grand modèle de langage (LLM).

4.1 Mesures de cohérence des *clusters* : *NMI* et *ARI*

À l'issue de la projection des *embeddings* et du *clustering* hiérarchique, effectués simultanément, avec l'algorithme h-NNE (Sarfraz *et al.*, 2022), nous sélectionnons deux partitions de clusters des messages avec respectivement 6 et 49 clusters et deux partitions des clusters des biographies, avec 6 et 63 clusters. Nous calculons le NMI et l'ARI pour le croisement des 2 partitions de clusters de message et des clusters de biographies des utilisateurs. L'analyse de nos expérimentations indique que les relations entre les descriptions biographiques et les messages sont très faibles, ce qui s'explique par une forte

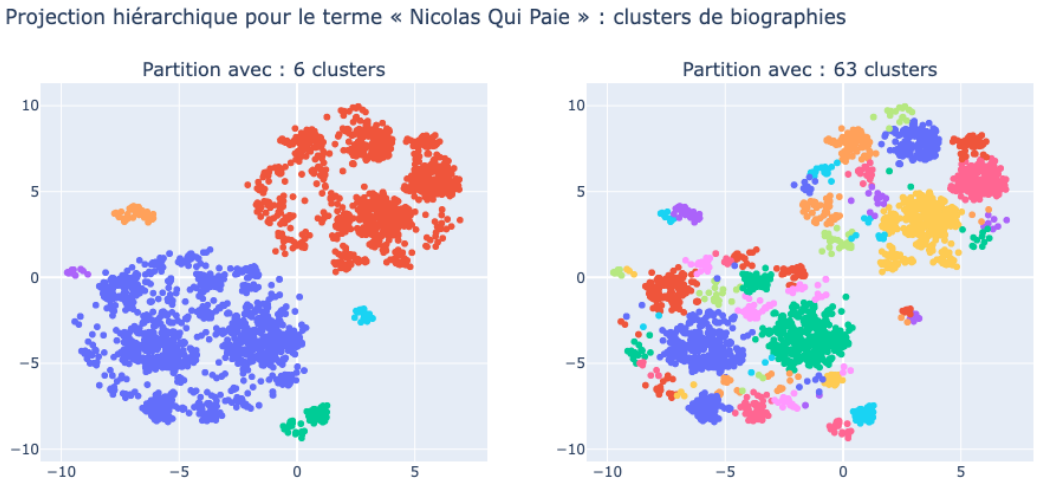

FIGURE 4 – 2 partitions de *clusters* des **biographies**. Partition 1 : 6 *clusters*, partition 2 : 63 *clusters*.

homogénéité au sein des messages étudiés dans nos expérimentations. Les scores obtenus, présentés dans le Tableau 1, montrent que la corrélation entre les clusters de messages et de biographies reste très faible, avec plusieurs niveaux de granularité. Ces résultats conduisent à la première conclusion que les messages publiés dans notre jeu de données n'apportent pas d'informations significatives sur les profils des utilisateurs.

# Bio Clusters	# Message Clusters	Score NMI	Score ARI
6	4	0.00185	0.00063
6	49	0.01187	0.00048
63	4	0.01248	0.00214
63	49	0.05824	0.00087

TABLE 1 – Les scores NMI et ARI. Croisement des deux niveaux de clusters hiérarchiques pour biographies et messages, avec l'algorithme hiérarchique de *clustering* h-NNE (Sarfraz *et al.*, 2022).

4.2 Spécificité et annotation des clusters

Pour les clusters de messages et de biographies, nous calculons les scores de spécificité (Lafon, 1980) afin d'obtenir les termes les plus spécifiques au sein de chaque cluster. Nous considérons que cette démarche permet d'extraire les termes caractéristiques de chaque clusters. Néanmoins, l'interprétation demeure difficile, en raison du nombre élevé des clusters et des termes associés. L'étiquetage des clusters des messages et des biographies par un grand modèle de langage, dans des catégories prédéterminées (voir les catégories dans les Tableaux 5 et 6), nous permet d'effectuer une analyse qualitative de notre démarche. Nous présentons nos résultats avec des annotations simples et des annotations politiquement nuancées respectivement dans les Tableaux 2 et 3. Les chiffres représentent le nombre de messages ayant été catégorisés ; lorsque plusieurs clusters sont annotés avec la même étiquette, le nombre de messages par catégorie est agrégé. Par exemple, dans le Tableau 2, pour le cas présenté en (b), les 49 clusters de messages initiaux se regroupent en 3 clusters uniques (`Culture`, `Média`, `Politique`) selon les étiquettes attribuées ; le LLM a étiqueté tous

les messages uniquement avec ces 3 catégories. La même démarche de regroupement a été effectué pour les biographies.

(a) 6 bios / 4 clusters messages

Identité (bio)	Political
identity.cultural	194
identity.philosophical	3284
identity.professional	**3895**
Total	7373

(b) 6 clusters bios / 49 clusters messages

Identité (bio)	Culture	Media	Political
identity.cultural	0	18	176
identity.philosophical	5	280	2999
identity.professional	3	363	**3529**
Total	8	661	6704

(c) 63 bios / 4 clusters messages

Identité (bio)	Political
identity.unclear	6
identity.cultural	1126
identity.philosophical	1568
identity.political	290
identity.professional	**4339**
identity.religious	44
Total	7373

(d) 63 clusters bios / 49 clusters messages

Identité (bio)	Culture	Média	Political
identity.unclear	0	1	5
identity.cultural	1	93	1032
identity.philosophical	3	150	1415
identity.political	1	20	269
identity.professional	3	394	**3942**
identity.religious	0	3	41
Total	8	661	6704

TABLE 2 – Nombre de messages par catégorie de clusters annotée par le LLM avec des étiquettes simples. La colonne gauche représente les catégories pour les biographies des auteurs. Les colonnes de droite présentent les catégories (première ligne) et le nombre de messages par catégorie de biographie correspondante. Voir l'ensemble des étiquettes dans le Tableau 5 en annexe.

En ce qui concerne les messages, les résultats montrent qu'à une petite granularité, avec 4 clusters de messages, les messages sont tous catégorisés comme politique, avec une annotation simple et en politique de droite, lorsqu'il s'agit d'une annotation politiquement nuancée (voir (a) et (c) dans les Tableaux 2 et 3). À une granularité plus fine des clusters, avec 49 clusters de messages (voir (b) et (d) dans les Tableaux 2 et 3), nous observons que les messages sont catégorisés au maximum en 3 catégories parmi les 10 catégories mentionnées par le prompt : `"political.left"`, `"political.right"`, `"political.center"`, `"media"`, `"economic"`, `"cultural"`, `"religious"`, `"technological"`, `"environmental"`, `"philosophical"`. Néanmoins, la majorité des messages sont catégorisés en `political.right` (politique de droite).

En ce qui concerne la catégorisation des biographies, à une granularité basse, 6 clusters de biographies sont annotés en trois catégories : culturelle, philosophique et professionnelle, parmi les 7 catégories d'étiquettes mentionnées dans le prompt : `professionnelle`, `politique`, `culturelle`, `religieuse`, `patriotique`, `philosophique` (voir les étiquettes dans le Tableau 5 en annexe). À une granularité plus élevée (63 clusters), les biographies sont classées dans toutes les huit catégories d'étiquettes parmi : `professionnelle`, `politique de gauche`, `politique de droite`, `politique de centre`, `culturelle`, `religieuse`, `patriotique`, `philosophique`. Les biographies sans étiquettes ont été manuellement classées dans `identity.unclear` ou `pas clair`. Nous présentons en annexe quelques exemples représentatifs de l'annotation automatique réalisée par le modèle : le Tableau 7 illustre les annotations des clusters de biographies, et les Tableaux 8, et 9 montrent l'annotation des clusters de messages.

(a) 6 bios / 4 clusters messages

Identité (bio)	Political.right
identity.cultural	194
identity.philosophical	3284
identity.political.right	75
identity.professional	**3820**
Total	7373

(b) 6 clusters bios / 49 clusters messages

Identité (bio)	Media	Political.left	Political.right
identity.cultural	18	17	159
identity.philosophical	280	181	2823
identity.political.right	10	3	62
identity.professional	353	221	**3246**
Total	661	422	6290

(c) 63 bios / 4 clusters messages

Identité (bio)	Political.right
identity.unclear	6
identity.cultural	1577
identity.patriotic	15
identity.philosophical	805
identity.political.center	48
identity.political.left	41
identity.political.right	1501
identity.professional	**3336**
identity.religious	44
Total	7373

(d) 63 bios / 49 clusters messages

Identité (bio)	Media	Political.left	Political.right
identity.unclear	1	1	4
identity.cultural	126	89	1362
identity.patriotic	0	1	14
identity.philosophical	83	52	670
identity.political.center	2	3	43
identity.political.left	6	0	35
identity.political.right	139	81	1281
identity.professional	301	193	**2842**
identity.religious	3	2	39
Total	661	422	6290

TABLE 3 – **Annotation des clusters avec orientations politiques** (gauche, droite, centre). La colonne gauche représente les catégories pour les biographies des auteurs. Les colonnes de droite présentent les catégories (première ligne) et le nombre de messages par catégorie de biographie correspondante. Voir l'ensemble des étiquettes dans le Tableau 6 en annexe.

5 Discussion

Malgré l'originalité de cette approche visant à intégrer les auto-descriptions ou biographies des auteurs afin d'étudier les correspondances avec des catégories des messages similaires, cette étude, encore dans sa phase préliminaire, présente plusieurs limitations. Pour nos travaux futurs, nous envisageons plusieurs étapes afin de consolider notre méthodologie. Premièrement, nous effectuerons une comparaison rigoureuse avec d'autres méthodologies utilisées pour des études similaires. Pour le regroupement ou le clustering des données, nous établirons des protocoles de comparaison avec différents algorithmes issus de la littérature, tels que K-Means et HDBSCAN. Deuxièmement, la pré-sélection des catégories pour l'annotation des *clusters*, fondée sur une analyse empirique, demeure limitée et nécessite un affinement important. Elle devrait être complétée et précisée en collaboration avec des experts en sciences politiques et en sociologie.

Dans ce cadre, l'annotation automatique repose uniquement sur les capacités intrinsèques des grands modèles de langue, sans définition explicite des orientations politiques, lesquelles peuvent varier selon les contextes linguistiques (anglophone, francophone). Cela peut introduire des biais politiques et culturels. Par exemple, les notions politiques telles que « gauche », « droite » et « centre » diffèrent selon les contextes nationaux. Par ailleurs, les données utilisées pour le pré-entraînement des modèles sont elles-mêmes hétérogènes dans leur représentation des concepts politiques, ce qui peut influencer les résultats d'annotation ; à titre d'exemple, les corpus en anglais britannique, américain et canadien présentent des différences notables dans leur interprétation du discours politique.

L'évaluation de la méthodologie doit également être renforcée. L'analyse quantitative, actuellement

fondée sur des mesures telles que l'information mutuelle normalisée (NMI) et l'indice de Rand ajusté (ARI), devrait être complétée par d'autres indicateurs plus robustes. De même, l'analyse qualitative, basée sur le calcul de termes spécifiques et l'annotation par un modèle de type Claude Sonnet 3.7, mérite d'être diversifiée afin d'améliorer sa fiabilité. Nous prévoyons d'intégrer des annotations manuelles dans un cadre d'accord inter-annotateur hybride afin de renforcer la fiabilité des résultats. Enfin, l'extension à des corpus plus hétérogènes et multilingues (anglais, français, espagnol, bengali), ainsi qu'à d'autres plateformes (Reddit, Telegram), permettra d'évaluer la généralisabilité de l'approche.

Nous tenons à préciser que cette recherche ne s'inscrit dans aucune adhésion politique. Notre objectif reste l'analyse objective des phénomènes sociopolitiques observables sur les réseaux sociaux.

6 Conclusion

Notre analyse préliminaire indique une forte homogénéité au sein des messages composant notre jeu de données. Malgré la présence de plusieurs catégories de biographies, les messages sont majoritairement politiques. Ce manque de diversité ne permet pas d'établir une corrélation significative entre les regroupements des *clusters* de messages et ceux des biographies.

Bien qu'une certaine hétérogénéité soit observable dans les profils des utilisateurs à partir de leurs biographies, les messages sont majoritairement classés comme politiques, souvent associés à une orientation de droite. Cette homogénéité thématique, liée à la sélection des données, limite la diversité observable du contenu et rend difficile l'établissement d'une corrélation entre messages et profils d'auteurs. Une interprétation possible est que les personnes engagées autour du slogan « *Nicolas Qui Paie* » appartiennent majoritairement à la population active et se décrivent dans leurs biographies à l'aide de termes liés à leur profession, métier ou activité.

Par ailleurs, les biographies des utilisateurs sont souvent très courtes, et les messages se concentrent sur un slogan spécifique (« *Nicolas Qui Paie* ») du mouvement politique étudié sur une période temporelle limitée, ce qui tend à renforcer l'homogénéité des messages et à réduire la diversité de l'information exploitable par notre méthodologie.

Ce projet s'inscrit dans le cadre d'une thèse CIFRE portant sur l'analyse des données sociales et politiques issues des réseaux sociaux. Notre méthodologie combine des approches classiques d'analyse textuelle telles que le *clustering* et le calcul de spécificité, avec des approches plus récentes reposant sur l'utilisation de grands modèles de langage pour l'annotation automatique. À l'intersection du TAL, de l'analyse politique et sociale, notre objectif est de développer des méthodologies reproductibles permettant de mieux comprendre la diversité des contenus publiés sur des sujets particuliers ainsi que les liens avec les comportements des utilisateurs.

Malgré certaines limitations, nous restons convaincus que la prise en compte du contexte tel que la biographie des auteurs, joue un rôle important pour l'analyse des messages publiés sur des réseaux sociaux, souvent étudiés de manière isolée. Cette combinaison permet d'explorer des données issues de réseaux sociaux en constante évolution et met en évidence à la fois les défis méthodologiques et les opportunités qu'offre ce type d'analyse pour la recherche académique comme pour les applications industrielles.

Références

BENOIT K., DE MARCHI S., LAVER C., LAVER M. & MA J. (2025). Using large language models to analyze political texts through natural language understanding. *American Journal of Political Science*.

BOCHES D. J. & COONEY M. (2023). What counts as "violence ?" semantic divergence in cultural conflicts. *Deviant Behavior*, **44**(2), 175–189.

CAMPELLO R. J., MOULAVI D. & SANDER J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, p. 160–172 : Springer.

CHALEHCHALEH R., SALEHI M., FARAHBAKHSH R. & CRESPI N. (2024). BRaG : a hybrid multi-feature framework for fake news detection on social media. *Social Network Analysis and Mining*, **14**(35), 50. DOI : 10.1007/s13278-023-01185-7, HAL : hal-04533937.

CHEN J., XIAO S., ZHANG P., LUO K., LIAN D. & LIU Z. (2024). M3-embedding : Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 2318–2335, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.137.

DEY K., SHRIVASTAVA R. & KAUSHIK S. (2017). Twitter stance detection — a subjectivity and sentiment polarity inspired two-phase approach. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, p. 365–372. DOI : 10.1109/ICDMW.2017.53.

DIMITROV D., BARAN E., FAFALIOS P., YU R., ZHU X., ZLOCH M. & DIETZE S. (2020). Tweetscov19-a knowledge base of semantically annotated tweets about the covid-19 pandemic. In *Proceedings of the 29th ACM international conference on information & knowledge management*, p. 2991–2998.

FRAISIER O. (2018). *Détection de points de vue sur les médias sociaux numériques*. Theses, Université Paul Sabatier - Toulouse III. HAL : tel-02288853.

GHOSH S., EKBAL A. & BHATTACHARYYA P. (2022). What Does Your Bio Say ? Inferring Twitter Users' Depression Status From Multimodal Profile Information Using Deep Learning. *IEEE Transactions on Computational Social Systems*, **9**(5), 1484–1494. DOI : 10.1109/TCSS.2021.3116242.

GOLBECK J., ROBLES C., EDMONDSON M. & TURNER K. (2011). Predicting Personality from Twitter. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, p. 149–156, Boston, MA, USA : IEEE. DOI : 10.1109/PASSAT/SocialCom.2011.33.

HUCHET A., GUILLAUME J.-L. & GHAMRI-DOUDANE Y. (2023). Faites du bruit pour la détection de communautés consensuelles (mais pas trop) ! In *AlgoTel 2023 - 25èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications*, Cargèse, France. HAL : hal-04068902.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, **1**(1), 127–165. DOI : 10.3406/mots.1980.1008.

LIU Z. & SONG T. (2023). Big data analysis and user behavior prediction of social networks based on artificial neural network. *Journal of computing and information technology*, **31**(3), 185–201.

MARCHAND M. (2015). *Domaines et fouille d'opinion : une étude des marqueurs multi-polaires au niveau du texte*. Theses, Université Paris Sud - Paris XI. HAL : tel-01157951.

NEMKOVA P. A., UBANI S. & ALBERT M. V. (2025). Comparing llm text annotation skills : A study on human rights violations in social media data. *arXiv preprint arXiv :2505.10260*.

SARFRAZ M. S., KOULAKIS M., SEIBOLD C. & STIEFELHAGEN R. (2022). Hierarchical Nearest Neighbor Graph Embedding for Efficient Dimensionality Reduction. Version Number : 3, DOI : 10.48550/ARXIV.2203.12997.

SHUSTER S. M., CAMPOS-CASTILLO C., MADANI N. & JOSEPH K. (2024). Who supports bernie ? analyzing identity and ideological variation on twitter during the 2020 democratic primaries. *Plos one*, **19**(4), e0294735.

TAN Z., LI D., WANG S., BEIGI A., JIANG B., BHATTACHARJEE A., KARAMI M., LI J., CHENG L. & LIU H. (2024). Large language models for data annotation and synthesis : A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 930–957.

THONET T. (2017). *Modèles thématiques pour la découverte non supervisée de points de vue sur le Web*. Theses, Université Toulouse 3 – Paul Sabatier. HAL : tel-01655278.

VAINIO J. & HOLMBERG K. (2017). Highly tweeted science articles : who tweets them ? An analysis of Twitter user profile descriptions. *Scientometrics*, **112**(1), 345–366. DOI : 10.1007/s11192-017-2368-0.

VANDEWEERDT C., EADY G., HJORTH F. & DINESEN T. (2026). Measuring Social and Political Identities In Social Media Self-Descriptions.

ZHANG Y., LI M., LONG D., ZHANG X., LIN H., YANG B., XIE P., YANG A., LIU D., LIN J. *et al.* (2025). Qwen3 embedding : Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv :2506.05176*.

ANNEXES

7 Prompt

7.1 Prompt principal

```
"""You're going to perform the task of Cluster tagging based on
specificity scores on French data obtained from a well known social
media platform. Tag the following cluster with a tag that best
represents the cluster based on the terms with the highest specificity
scores (Lafon). Infer deeper and explicit meanings when you can, but
don't force meaning where there exists none, ex. single number etc.
Do not be fooled by single terms, for example if there is 'gauche'
and 'droite' in the cluster, it doesn't necessarily indicate the
political inclination of the author. Pick tag amongst
tags = {tags}. The tag should be as specific as possible. Only
return the tag as output in a python string format."""
```

TABLE 4 – Prompt principal ; *{tags}* à remplacer par **bio_tags** ou **message_tags.**

7.1.1 Étiquettes simples

```
bio_tags =["identity.professional", "identity.political", "identity.cultural", "
    identity.religious", "identity.patriotic", "identity.philosophical"]

message_tags = ["political", "media", "economic", "cultural", "religious", "
    technological", "environmental", "philosophical"]
```

TABLE 5 – Étiquettes simples ; *bio_tags* = étiquettes pour biographies, *message_tags* = étiquettes pour messages

7.1.2 Étiquettes politiquement nuancées

```
bio_tags = ["identity.professional", "identity.political.left", "identity.
    political.right", "identity.political.center", "identity.cultural", "identity.
    religious", "identity.patriotic", "identity.philosophical"]

message_tags = ["political.left", "political.right", "political.center", "media",
    "economic", "cultural", "religious", "technological", "environmental", "
    philosophical"]
```

TABLE 6 – Étiquettes politiquement nuancées ; *bio_tags* = étiquettes pour biographies, *message_tags* = étiquettes pour messages

8 Annotation de clusters par LLM

8.1 Annotation des clusters de biographies par LLM

TABLE 7 – Exemple d'annotation des clusters de biographies à partir des scores de spécificité


```
"bio_cluster_1": {
       "identity.philosophical": {
           "responsabilité": 3.3133635520157343,
           "n'est": 2.5790786489909197,
           "rien": 2.2842791231354225,
           "sans": 2.010974004176322,
           "liberté": 1.9929532900454359,
           "La": 1.7434547074954845
       }
   },
   "bio_cluster_2": {
       "identity.political.right": {
           "TOUT": 4.3045121068389305
       }
   },
   "bio_cluster_3": {
       "identity.philosophical": {
           "sait": 3.1710765782808203,
           "Tout": 2.8575736141616885,
           "monde": 2.226293841116375
       }
   },
   "bio_cluster_4": {
       "identity.professional": {
           "perlimpinpin": 3.7482267678018513,
           "poudre": 3.5721527781785225,
           "Ingénieure": 3.44723131168931,
           "horreur": 2.9033356934432444,
           "a": 1.5376900679795276
       }
   },
   "bio_cluster_5": {
       "identity.political.right": {
           "Facho-sapiens": 4.3045121068389305
       }
   },
```

8.2 Annotation des clusters de messages par LLM

TABLE 8 – Exemple 1 : annotation des clusters de messages à partir des scores de spécificité


```
"message_cluster_23": {
        "media": {
            "Linda,": 3.428300948659747,
            "auditrice": 3.428300948659747,
            "Radio": 3.428300948659747,
            "Sud": 3.1273499034128083,
            "sent": 2.951337609129179,
            "quitter": 2.951337609129179,
            "France.": 2.68829338427957,
            "envie": 2.5836764269343893,
            "@SudRadio:": 2.2679980861798774,
            "marre": 2.1743696365850598,
            "coup": 2.077814398795242,
            "a": 1.9744876890760836,
            "gueule": 1.7737977142006176,
            "monde": 1.5648195228245156,
            "On": 1.446892966780505,
            "payer": 1.417330752989419,
            "tous": 1.407259470153186,
            "Il": 1.1710656954431113,
            "tout": 1.1030605726867058,
            "»": 1.0066974293142141,
            "Le": 0.9330527140331196,
            "«": 0.8454801700774289,
            "@NicolasQuiPaie": 0.3034501006066528,
            "RT": 0.23605655031436826
        }
    },
```

TABLE 9 – Exemple 2 : annotation des clusters de messages à partir des scores de spécificité


```
    "message_cluster_24": {
        "political.right": {
            "l'alliance": 4.278963398481795,
            "#GiletsJaunes...": 4.278963398481795,
            "@ToddVousGuette": 3.9779438049760514,
            "intéressant": 3.9779438049760514,
            "Ça": 2.204641518547222,
            "être": 2.1630417457571367,
            "va": 1.9487463469689152,
            "#NicolasQuiPaie": 1.5215399989662042
        }
    },

    ...
    },
    "message_cluster_26": {
        "political.left": {
            "gauche": 3.790607248519736,
            "intégrer.": 2.580516783092349,
            "demain": 2.4917002149656255,
            "doivent": 2.366985670001737,
            "l'extrême": 2.3106288875813785,
            "#gueux": 2.1913428043678658,
            "@monsieurnaudin:": 2.1530090309356678,
            "surtout": 2.0852609742962596,
            "Si": 1.818368468097837,
            "mouvement": 1.2879468528294944,
            "#NicolasQuiPaie": 1.1733526302810204,
            "a": 1.0233650027378414,
            "RT": 0.3694242662654902
        }
    },
```