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

Annotation collaborative de faits et d'opinions dans des données conversationnelles

Léo Rongieras[1] Luce Lefeuvre[1]
(1) SNCF - DTIPG, Saint Denis, 93210, France
ext.leo.rongieras@sncf.fr, luce.lefeuvre@sncf.fr

RÉSUMÉ

L'écoute client sur les réseaux sociaux se limite souvent à l'analyse des thématiques et des sentiments, négligeant les dynamiques argumentatives présentes dans les conversations. Cet article présente la constitution d'un corpus conversationnel extrait de Bluesky, centré sur le transport ferroviaire, afin d'étudier l'interaction entre l'expression de faits et l'expression d'opinions. À partir de données filtrées par modélisation thématique, une campagne d'annotation manuelle a été menée sur 450 conversations (1 117 messages). L'évaluation approfondie de l'accord inter-annotateurs (Alpha de Krippendorff, Gamma, F1-IoU) met en évidence les défis liés à la délimitation fine des segments (unitizing problem) dans des textes où descriptions factuelles et jugements subjectifs sont étroitement intriqués.

ABSTRACT

Collaborative annotation of facts and opinions in conversationnal data.

Customer listening on social media is often limited to topic and sentiment analysis, neglecting the argumentative dynamics present in conversations. This paper presents the creation of a conversational corpus extracted from Bluesky, centered on railway transport, to study the interaction between expressions of facts and opinions. Based on data filtered through topic modeling, a manual annotation campaign was conducted on 450 conversations (1,117 posts). An in-depth evaluation of inter-annotator agreement (Krippendorff's Alpha, Gamma, F1-IoU) highlights the challenges associated with the fine-grained delimitation of segments (unitizing problem) in texts where factual descriptions and subjective judgments are closely intertwined.

MOTS-CLÉS : Argumentation, Faits, Opinions, Réseaux sociaux, Annotation.

KEYWORDS: Argumentation, Facts, Opinions, Social media, Annotation.

1 Introduction

Pour les entreprises de services, où l'expérience vécue constitue le cœur de la proposition de valeur, comprendre finement ce que disent les clients est un levier de fidélisation et de performance important. L'attention apportée aux différents retours des clients vis à vis des services proposés est donc essentielle pour que les entreprises améliorent à la fois leur offre de service et également leur communication. Les réseaux sociaux numériques sont rapidement apparus comme une source riche de points de vue, d'avis, d'irritants ou d'éléments de satisfaction exprimés par des clients ou des non-clients. Les solutions de veille ou d'écoute sociale existantes analysent le contenu textuel de ces réseaux sous l'angle des thématiques abordées, des sentiments et des émotions exprimés par les

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 107–120.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

utilisateurs. Pertinentes pour avoir un aperçu général de ce qui se dit sur les réseaux sociaux, ces analyses ne permettent cependant pas de rendre compte des interactions qui s'opèrent au sein des échanges entre les utilisateurs, et plus particulièrement des mécanismes d'argumentation en jeu.

L'article propose ici une partie d'un travail plus large dont l'ambition est d'étudier la manière dont les personnes interagissent entre elles sur les réseaux sociaux. Plus précisément, notre travail s'intéresse à l'expression de faits et l'expression d'émotions, que nous considérons comme deux stratégies argumentatives qui peuvent être mises en oeuvre au cours d'une conversation. Après avoir dressé un bref état de l'art et présenté la problématique de l'étude, cet article présente la méthode de constitution d'un corpus de données conversationnelles issues de Bluesky, et annoté en faits et opinions. Nous présentons ensuite nos résultats en nous appuyant sur des métriques utilisées pour évaluer l'accord inter-annotateurs, et discutons ces résultats en conclusion.

2 État de l'art et problématique

L'extraction et l'analyse des structures argumentatives au sein de textes en langage naturel relèvent du domaine de la fouille d'arguments (*Argument Mining*). Au cœur de cette discipline se trouve la tâche de détection d'assertions (*Claim Detection*), qui consiste à identifier la conclusion centrale ou le point de vue défendu par un utilisateur au sein d'un texte (Stab & Gurevych, 2017; Sundriyal *et al.*, 2022).

2.1 Détection d'assertions et argumentation

Historiquement, les recherches en fouille d'arguments se sont d'abord concentrées sur des corpus de nature monologique où la structure rhétorique est explicite. Les premiers travaux ont ciblé des documents juridiques (Moens *et al.*, 2007), des articles de Wikipédia (Levy *et al.*, 2014), ou encore des essais (Stab & Gurevych, 2017) pour détecter des assertions indépendantes du contexte. L'objectif est généralement d'isoler la déclaration principale des prémisses (les preuves ou justifications) qui la soutiennent (par exemple :*"L'accès aux centres-villes doit être limité pour les voitures [Assertion] car cela réduit significativement la pollution de l'air [Prémisse]").*

Appliqués aux données des réseaux sociaux, les travaux de Habernal et Gurevych (Habernal & Gurevych, 2017), ont montré la nécessité d'adapter les schémas argumentatifs classiques aux spécificités de ces données : brièveté des messages, contenu implicite, sarcasme, syntaxe particulière, présence d'émojis, etc. Sur des plateformes de microblogging, d'autres études (Bosc *et al.*, 2016) ont mis en évidence l'importance de la contextualisation des messages pour réussir à en extraire la trame argumentative, comme intégrer la branche complète de la conversation. En effet, l'argumentation sur les réseaux sociaux se construit de manière collaborative ou conflictuelle à travers les fils de discussion. Plus récemment, des cadres d'évaluation spécifiques aux réseaux sociaux, tels que CLAIMSCAN (Sundriyal *et al.*, 2023), ont été développés pour détecter non seulement la présence d'une assertion dans un texte bruité, mais également pour en délimiter la portée exacte.

D'autres travaux s'intéressent à l'argumentation dans un contexte interactionnel. (Feger & Dietze, 2024), par exemple, proposent un jeu de données issu de Twitter et un schéma distinguant des arguments de type *inférence* et *informationnel*. Dans le but d'extraire les suites d'interaction (soutien ou attaque) dans des conversations Reddit, (Almerge *et al.*, 2025) s'appuient sur des annotations générées artificiellement par des modèles de langue. Toutefois, l'utilisation de modèles pour l'annotation

automatique présente des limites. Comme le soulignent (Sun *et al.*, 2026) dans leurs récents travaux sur l'extraction d'arguments faiblement supervisée par des LLMs (Large Language Models), les données pré-annotées automatiquement (« pseudo-labels ») s'avèrent extrêmement bruitées. Leur méthode nécessite d'ailleurs des modules complexes de raffinement des frontières et de débruitage des relations pour compenser les erreurs initiales de la machine. Ces constats soulignent le besoin persistant de corpus de référence annotés manuellement, garantissant une caractérisation fine des types d'arguments étudiés.

2.2 Caractérisation de l'argumentation : distinction entre faits et opinions

L'identification d'une assertion n'est cependant qu'une première étape dans l'analyse des arguments échangés au cours d'une conversation. Pour qualifier finement les dynamiques d'échanges et les mécanismes de persuasion mis en jeu, il est crucial de caractériser la nature de l'assertion extraite. Ainsi, une distinction entre les faits et les opinions doit être établie. Cette distinction permet d'aller au-delà de la simple analyse de sentiments. Les faits peuvent être définis comme étant des énoncés objectifs, vérifiables, basés sur des expériences ou des données ; tandis que les opinions sont caractérisées par des énoncés subjectifs, des jugements de valeur, des croyances (Alhindi *et al.*, 2020; Wiebe *et al.*, 2005).

Par ailleurs, Alhindi et al. (Alhindi *et al.*, 2020) démontrent que l'intégration des caractéristiques argumentatives joue un rôle déterminant pour classifier et distinguer efficacement les faits des opinions, offrant ainsi des indices cruciaux sur la posture de l'auteur. Appliquée au domaine de l'écoute client, cette distinction est pertinente, car la posture des auteurs peut révéler l'expression d'une expérience concrète, ou l'expression d'un jugement de valeur. Toutefois, sur une plateforme comme Bluesky, les utilisateurs tendent à entrelacer étroitement ces deux dimensions. La description factuelle d'un événement peut servir à l'expression d'éléments subjectifs, générant des messages hybrides.

C'est précisément en raison de cette hybridation complexe que nous avons fait le choix méthodologique de ne recourir à aucune pré-annotation automatique pour la construction de notre ressource. Face à des frontières aussi ténues, soumettre des textes pré-annotés par un algorithme à des annotateurs humains comporte un risque majeur de biais d'ancrage (Baledent, 2022). L'annotateur a en effet tendance à valider la proposition de la machine plutôt qu'à analyser finement les nuances rhétoriques propres au domaine.

Pour apporter des éléments de réponse à cette difficulté, notre approche repose sur la délimitation fine de segments textuels, sur des données de Bluesky, afin d'isoler le noyau factuel de son enrobage subjectif au sein d'un même message. Nous souhaitons ainsi caractériser finement ce qui correspond à l'expression d'un fait ou d'une opinion dans un contexte de données conversationnelles comme celles de Bluesky. Pour cela, nous avons mis en place une campagne d'annotation que nous décrivons dans la suite.

3 Création d'un corpus issu de Bluesky de faits et opinions

3.1 Extraction de données conversationnelles

Afin d'obtenir un ensemble brut de conversations nous avons ciblé Bluesky comme réseau social pertinent. Bluesky a un fonctionnement similaire à celui de X, basé sur un système de fils de discussion permettant de suivre les échanges de manière chronologique. Une conversation commence à partir du moment où un utilisateur poste un message, qui peut être commenté par d'autres utilisateurs ou par lui-même pour alimenter la conversation.

Nous avons ciblé le transport ferroviaire comme domaine d'étude, faisant l'hypothèse que c'est un sujet qui est largement discuté sur les réseaux sociaux, et qui peut donner lieu à des polémiques, et donc à l'expression de faits ou d'opinions. Environ 8000 conversations en français ont ainsi été extraites via l'API offerte par Bluesky en selectionnant tous les fils de discussion qui contiennent au moins une occurence du mot "SNCF", sur la période couvrant 2023 à mars 2025. Cette extraction abouti à un ensemble 51 586 messages dans 8296 conversations et 11 682 utilisateurs uniques.

Dans le but d'obtenir un corpus textuel, nous avons retiré les discussions qui contenaient des éléments multimédias (images, vidéos, etc.) ainsi que les messages ne comportant que des emojis ou des réactions sans texte. Cela réduit notre extraction à un corpus de 47 007 messages, 10578 utilisateurs, et 8158 conversations.

3.2 Sélection des données à annoter

L'utilisation du mot-clé "SNCF" a permis de capter une grande quantité de données pertinentes, mais il engendre également un bruit important (publications historiques, flux de streaming, discussions politiques générales ou métaphoriques) qui ne relève pas directement de l'expérience des utilisateurs ou du service de transport.

Afin de sélectionner plus finement les conversations constitutives du corpus d'annotation, nous avons appliqué une approche par modèles thématiques (Topic Modeling) en utilisant le framework *BERTopic* (Grootendorst, 2022). Cette méthode s'appuie sur des plongements lexicaux (embeddings) et une variation de TF-IDF (c-TF-IDF) pour regrouper les documents par similarité sémantique. L'objectif de cette étape est de filtrer les conversations où le mot "SNCF" apparaît hors du sujet des transports ou de l'utilisation des services liés à la SNCF, et donc de sélectionner un sous-ensemble de données les plus pertinentes possibles par rapport à notre objectif. Par exemple, dans le Tableau 1, le thème 77, relatif à une diffusion en direct sur Twitch ("Zelda streaming event") a été écarté. Une analyse qualitative des groupements thématiques obtenus à ainsi été réalisée pour vérifier l'adéquation des données sélectionnées à notre objectif. Cette étape permet de centrer les données sur des thématiques propres à l'experience des clients comme les retards (thème 2) ou les retours sur l'application SNCF Connect (thème 73).

À l'issue de cette analyse, nous avons conservé uniquement les thèmes en lien avec les transports (perturbations, services en gare, confort à bord, application mobile), afin que la tâche d'annotation soit centrée sur la qualité de service et le ressenti des usagers ou des non-usagers.

TABLE 1 – Premiers et derniers *thèmes* identifiés par le modèle BERTopic sur le corpus BlueSky.

ID	Label attribué	Mots-clés représentatifs	Volume
0	Historical rail developments	gare, sncf, la, de, du, fer, rail	924
1	SNCF disruptions	trafic, actu, sncf, fr, perturbations	310
2	Train delays and issues	retard, train, je, minutes, heure	245
3	Sleep and trains	je, est, ai, ça, pas, que, un	238
...	...	...	...
73	SNCF Connect feedback	connect, arghh, percé, damned, exhaustif	11
74	SNCF workers' anger	fret, démantèlement, cheminots, colère	11
76	Labor strikes and rights	exerce, exercer, droit, francetvinfo	10
77	Zelda streaming event	twitch, tv, raillus, hyrule, link	10

3.3 Séquence d'annotation

Parmi les conversations restantes, une selection aléatoire de 450 conversations a été faite pour constituer le corpus d'annotation. Ces conversations ont été distribuées à 4 annotateurs de telle sorte que chaque annotateur ait à annoter environ 400 messages. Sur l'ensemble de ces messages, 80 sont communs à chaque annotateurs qui serviront de corpus d'analyse inter-annotation. Les annotateurs ont un profil identique : même niveau de formation (deuxième année de Master), et tous effectuaient un stage de recherche dans le même organisme.

3.3.1 Protocole d'annotation

Le protocole d'annotation est structuré de manière séquentielle afin de garantir la cohérence contextuelle. Les annotateurs traitent les données conversation par conversation, en suivant l'ordre chronologique des messages. Pour chaque unité traitée, l'annotateur a accès à l'historique complet des échanges précédents afin de lever les éventuelles ambiguïtés sémantiques.

Le processus de décision suit une structure hiérarchique décrite ci-dessous :

1. **Filtrage Thématique (niveau Conversation) :** L'annotateur doit d'abord déterminer si la conversation globale traite du domaine des **transports**.
   — Si la thématique est confirmée, l'annotateur procède à l'analyse granulaire de chaque message,
   — Dans le cas contraire, la conversation est écartée.
2. **Identification et Délimitation des Affirmations (niveau Message) :** Pour chaque message de la conversation sélectionnée, l'annotateur doit identifier et délimiter la ou les **affirmations** (*claims*).

   Afin de garantir la cohérence des annotations, les règles de sélection suivantes s'appliquent :
   — **Multiplicité :** Un même message peut faire l'objet de plusieurs annotations si l'annotateur y identifie plusieurs séquences d'affirmations distinctes,
   — **Exclusivité spatiale :** Les séquences textuelles annotées ne peuvent en aucun cas se chevaucher (*no overlap*). Chaque affirmation doit être délimitée de manière stricte et indépendante.

3. **Classification de l'Affirmation (annotation conjointe) :** il est important de préciser que le protocole repose sur un modèle d'annotation unifiée : un même annotateur peut identifier et annoter les deux classes. Chaque séquence identifiée lors de l'étape précédente doit ainsi être typée selon l'une des deux catégories suivantes :
   — **Fait (Fact) :** L'affirmation est considérée comme factuelle si elle porte sur un élément objectivement vérifiable, et ce, **peu importe son degré de vérité**. Qu'une information soit factuellement exacte ou erronée (mensonge, erreur), elle relève de cette catégorie dès lors qu'il existe un moyen objectif de la vérifier (ex. : *« Le train de 8h est arrivé à 9h »*, *« La ligne B compte 15 stations »*).
   — **Opinion :** L'affirmation est classée comme opinion si elle ne peut être vérifiée de manière purement objective. Elle regroupe les points de vue subjectifs, les jugements de valeur, les ressentis ou les préférences personnelles (ex. : *« Les sièges de ce bus sont très inconfortables »*, *« Ce retard est scandaleux »*).

Bien qu'il aurait été intéressant d'établir une comparaison entre ce modèle d'annotation conjointe et un modèle par spécialisation (où un annotateur se consacrerait exclusivement à la recherche de faits et un autre aux opinions) afin d'en évaluer l'impact sur les mesures d'accord inter-annotateurs, l'approche simultanée a été privilégiée. Ce choix permet à l'annotateur de mieux appréhender la rhétorique globale du locuteur et de capturer plus efficacement la frontière, parfois ténue, entre fait et opinion au sein d'un même message.

3.3.2 Analyse inter-annotation

Une attention particulière à l'evaluation et l'analyse des accords inter-annotateurs doit être portée dans le but d'évaluer la difficulté de la tâche proposée et de déterminer la cohérence, les points de désaccord possibles et évaluer les consignes données (Artstein & Poesio, 2008; Baledent, 2022).

Avant de réaliser ces évaluations, certaines limites à notre travail méritent d'être mentionnées. Tout d'abord, l'annotation de ce type de données soulève des défis liés à la nature même des réseaux sociaux. Les annotateurs doivent faire face à un "bruit" inhérent à la nature des textes qui proviennent de ce type de source (ponctuation répétitive, majuscules non standards, présence d'émojis et de hashtags) et qui vient perturber les schémas de lecture classiques. Ensuite, contrairement à des tâches de classification où les items sont prédéfinis et connus par les annotateurs, ces derniers sont ici libres de délimiter eux-mêmes ce qui constitue l'objet de leur annotation. Cette liberté se heurte au problème de délimitation des unités (unitizing problem) (Mathet *et al.*, 2015; Li *et al.*, 2024). Face à un continuum de texte, la délimitation des frontières devient très variable : un annotateur peut adopter une approche minimaliste (en sélectionnant uniquement le cœur de l'expression subjective), tandis qu'un autre adoptera une approche maximaliste (en incluant davantage d'éléments contextuels de l'expression comme des interjections, de la ponctuation expressive ou des propositions introductives), ce qui est illustré par la Figure 1, commentée dans les résultats. Enfin, les annotateurs devaient choisir strictement entre deux classes (fait ou opinion), alors que les verbatims des réseaux sociaux peuvent être hybrides.

Finalement, les annotateurs produisent zéro, une, ou plusieurs sous-séquences de tokens sélectionnées à partir d'un message. Notre analyse inter-annotateurs doit donc prendre en compte à la fois la sélection de ces sous-séquences, à savoir la localisation ainsi que les bornes qui fixent ces sous-séquences. Les annotateurs sont libres de déterminer eux-mêmes les bornes de ce qu'ils annotent avec

la seule contrainte de ne pas pouvoir faire chevaucher les séquences.

Dans ce cadre, l'analyse inter-annotateurs doit prendre en compte les problématiques suivantes :

— Les annotateurs définissent-ils les mêmes séquences **aux mêmes endroits** ? (problème de la localisation)
— Les annotateurs attribuent-ils **les mêmes étiquettes** (type opinion ou fait) à ces séquences ? (problème de catégorisation)

Ces deux questions ont guidé nos choix dans les métriques que nous décrivons dans la suite.

3.4 Métriques d'évaluation

Pour prendre en compte les problèmes de localisation et de catégorisation mentionnées plus haut, nous avons sélectionné les métriques qui nous paraissent les plus adaptées pour évaluer les tâches réalisées par les annotateurs. Deux métriques ont prinicpalement été utilisées : l'Alpha de Krippendorff (Krippendorff, 1995; Krippendorff *et al.*, 2016) et la métrique γ (Mathet *et al.*, 2015). La première est fondée sur le concept de désaccord. La seconde propose une approche unifiée et holistique du problème de l'unitizing et de la catégorisation. A notre connaissance, la définition d'une métrique idéale dans le cas d'annotation de séquences libres reste un problème ouvert (Li *et al.*, 2024).

3.4.1 L'Alpha de Krippendorff au niveau des tokens

Afin d'appliquer cette métrique classique à une tâche de segmentation, nous avons projeté les annotations sur une séquence de *tokens* en utilisant un schéma d'étiquetage BIO (Begin, Inside, Outside). L'Alpha de Krippendorff (α) évalue ainsi l'accord global sur la classification de chaque token, en tenant compte du désaccord attendu par hasard. Bien que robuste, cette approche tend à pénaliser les légers décalages de frontières. Dans notre corpus, cette métrique atteint un score de 0.4567, reflétant la difficulté de la tâche.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

D_o représente le désaccord observé (*observed disagreement*) entre les annotateurs, et D_e représente le désaccord attendu par hasard (*expected disagreement*). Pour une métrique de distance δ^2_{ck} entre les catégories c et k :

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck} \delta^2_{ck} \tag{2}$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c n_k \delta^2_{ck} \tag{3}$$

3.4.2 La famille de métriques Gamma

Contrairement aux métriques discrètes, la métrique γ évalue conjointement l'alignement des segments (positionnement) et leur catégorisation. Elle repose sur la recherche d'un alignement optimal minimisant le désaccord entre les annotateurs. Notre évaluation distingue le γ global calculé sur les tokens (0.4884) et sa composante purement catégorielle, le γ_{cat} (0.2221), qui isole la difficulté liée au choix du type d'opinion ou de fait, indépendamment des frontières.

$$\gamma = 1 - \frac{\delta(A)}{\delta_{\text{random}}} \tag{4}$$

$\delta(A)$ est le désordre observé (observed disorder) de l'ensemble des annotations A, calculé comme la moyenne des distances d'appariement. δ_{random} est le désordre attendu, estimé de manière empirique par échantillonnage aléatoire (génération d'annotations fictives basées sur la distribution des données).

Le désordre d'un alignement $\bar{a}$ est défini par :

$$\delta(\bar{a}) = \frac{1}{|\bar{a}|} \sum_{a \in \bar{a}} d(a) \tag{5}$$

3.4.3 Appariement des segments et F1-score basé sur l'Indice de Jaccard

Pour évaluer l'accord de manière plus intuitive et quantifier la proximité des délimitations, nous avons implémenté une évaluation par paires d'annotateurs combinant l'Intersection sur Union (IoU, ou Indice de Jaccard) et le F1-score. L'indice de Jaccard, formulé comme le rapport entre l'intersection des *tokens* de deux segments et leur union, sert de métrique de base pour évaluer le degré de chevauchement. L'appariement des segments entre deux annotateurs est réalisé via l'algorithme hongrois, qui cherche à maximiser l'indice de Jaccard global des paires. Une fois les segments appariés, nous appliquons différents seuils d'indice de Jaccard pour déterminer si un appariement constitue une correspondance valide (considérée alors comme un Vrai Positif), ce qui permet ensuite de calculer le F1-score. Cette méthode met en évidence l'impact du problème de délimitation, illustré dans la Figure 3.4.3, et décrit ci-dessous :

- **F1-strict (IoU = 1) :** Lorsque l'on exige une correspondance exacte des frontières du segment, l'accord s'effondre (0.2549), prouvant l'inefficacité des métriques strictes pour ce type de données.
- **F1-span (IoU $\geq$ 0.6) :** En exigeant un chevauchement substantiel (seuil d'IoU de 60%), la métrique offre un compromis réaliste (0.4645).
- **F1-partial (IoU > 0) :** Si l'on considère comme valide tout chevauchement, même minime, entre deux annotations de même type, l'accord monte à 0.7331. Ce score élevé indique que les annotateurs repèrent généralement les mêmes zones subjectives, mais divergent sur leur étendue.

Afin de quantifier précisément la proximité des délimitations indépendamment des faux positifs ou faux négatifs purs, nous avons également calculé l'indice de Jaccard moyen sur l'ensemble des segments correctement appariés (avec un seuil de chevauchement partiel). Son score s'élève à 0.7059, ce qui confirme une forte cohésion autour du noyau des expressions, malgré des marges floues.

Illustration des trois critères d'évaluation par paires pour l'accord inter-annotateurs (F1-strict, F1-span et F1-partial). Les flèches indiquent l'appariement des segments.

3.5 Résultats de l'analyse inter-annotation

TABLE 2 – Résumé des métriques de performance évaluant l'accord inter-annotateurs.

Métrique	**Score**
F1 (partial span)	0.7331
IoU (boundary)	0.7059
γ (token)	0.4884
F1-span (0.6)	0.4645
α (token)	0.4567
κ (token)	0.4199
F1-strict	0.2549
γ_{cat}	0.2221

Le Tableau 2 présente les scores obtenus pour chaque métrique utilisée. Plusieurs réflexions en découlent. Tout d'abord, les métriques d'accord comme l'Alpha de Krippendorff ou le Gamma montrent toutes deux un accord modéré entre les annotateurs. Couplées à l'indice de Jaccard moyen nous pouvons suggérer que les annotateurs repèrent les mêmes régions du texte mais qu'il existe une difficulté importante à délimiter précisément les bornes des segments. Ce résultat peut être corrélé à la nature des textes, et à leur particularité syntaxique, idiomatique, et parfois même sémantique. Une modification du guide d'annotation en appuyant ou en clarifiant d'avantage les critères de délimitation pourrait permettre d'améliorer significativement l'accord inter-annotateurs. Notamment on observe des différences d'approche entre les annotateurs en choisissant soit une approche "atomiste" en séparant les segments en plusieurs parties, ou une approche "holistique" en regroupant les segments en une seule unité. La Figure 1 illustre cette différence d'approche où l'annotateur 3 regroupe en une seule unité deux segments catégorisés comme Opinion par l'annotateur 4. Enfin, la catégorisation binaire, imposée aux annotateurs, a pu avoir un impact sur les résultats, en augmentant les désaccords sur les segments hybrides ou mixtes, dans lesquels descriptions factuelles et émotions sont présentes.

4 Corpus final et exploitation des données

Cette campagne d'annotation aboutit à un corpus de 279 conversations et 1117 messages annotés. Le Tableau 4 (en annexe) présente les statistiques descriptives du corpus, mettant en lumière la diversité des conversations et la longueur moyenne des messages. La moitié des fils de discussion sont en réalité un seul message, conservant dans une certaine mesure la distribution présente dans notre extraction initiale (Tableau 4). Le tableau 5 (en annexe) indique que le ratio de segments annotés comme "Opinion" s'élève à 62%, avec une moyenne de 1,82 annotation par message. Par ailleurs, les segments factuels tendent à être légèrement plus longs, affichant une longueur moyenne de 15,11 mots, contre 12,74 pour les opinions. La Figure 2 montre deux conversations annotées extraites de notre corpus final.

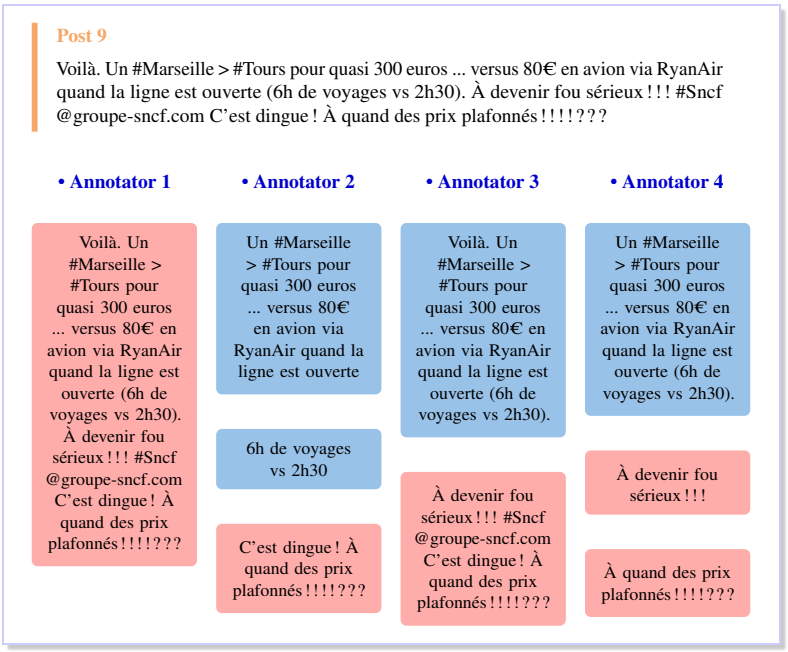

FIGURE 1 – Exemple d'annotations multiples sur un même post (problème de délimitation). Les boîtes bleues représentent les segments factuels et les boîtes rouges les segments d'opinion.

Il convient de noter que les données produites présentent certaines spécificités : le corpus est actuellement mono-domaine (transport), mono-plateforme (Bluesky) et mono-langue (français). Néanmoins, si les résultats directement issus de ce corpus sont liés à ce contexte précis, la méthode d'annotation et les métriques d'évaluation développées sont, quant à elles, indépendantes du domaine. Nous formulons donc l'hypothèse que ce cadre méthodologique est hautement transférable et peut être appliqué avec succès à d'autres types de données conversationnelles (forums, autres réseaux sociaux) ou à d'autres thématiques d'étude (santé, politique, service client).

Concernant l'exploitation de cette nouvelle ressource, ce corpus a vocation à servir de vérité terrain pour de futures expérimentations algorithmiques en Traitement Automatique des Langues. En l'absence de *baseline* dans cette étude exploratoire, ces données ouvrent la voie à l'entraînement et à l'évaluation de modèles de classification automatique. Il sera par exemple possible de procéder à l'ajustement de modèles de langue pré-entraînés ou d'évaluer les capacités des Grands Modèles de Langue actuels via des approches *few-shot* pour l'extraction et la classification conjointe d'affirmations. À plus long terme, la distinction automatique entre Fait et Opinion sur des fils de discussion constitue une brique essentielle pour alimenter des systèmes d'extraction d'arguments, des pipelines de *fact-checking*, ou pour affiner l'analyse de l'expérience utilisateur en séparant l'expression ou la description d'état de fait, du ressenti client ou d'une exposition d'opinion. D'un point de vue plus théorique, cette ressource peut s'avérer riche pour analyser la manière dont la co-construction des échanges s'effectue sur les réseaux sociaux, notamment via l'enchainement des arguments et des opinions. Le choix de l'annotation binaire "Fait vs Opinion" permet de distinguer ce qui relève de la présentation d'événements, de faits - vécus ou non, des prises de position. Ainsi, l'analyse des annotations peut dans une certaine mesure permettre de révéler les stratégies linguistiques utilisées par les utilisateurs pour exprimer leur point de vue ou positionnement par rapport à un énoncé. De cette manière, dans la lignée de travaux comme ceux de (Biber *et al.*, 2007; Biber & Conrad, 2000), la ressource créée peut permettre de caractériser les positionnements (*stance*) des utilisateurs, en opposition aux contenus annotés comme des faits et donc censés être neutres et objectifs.

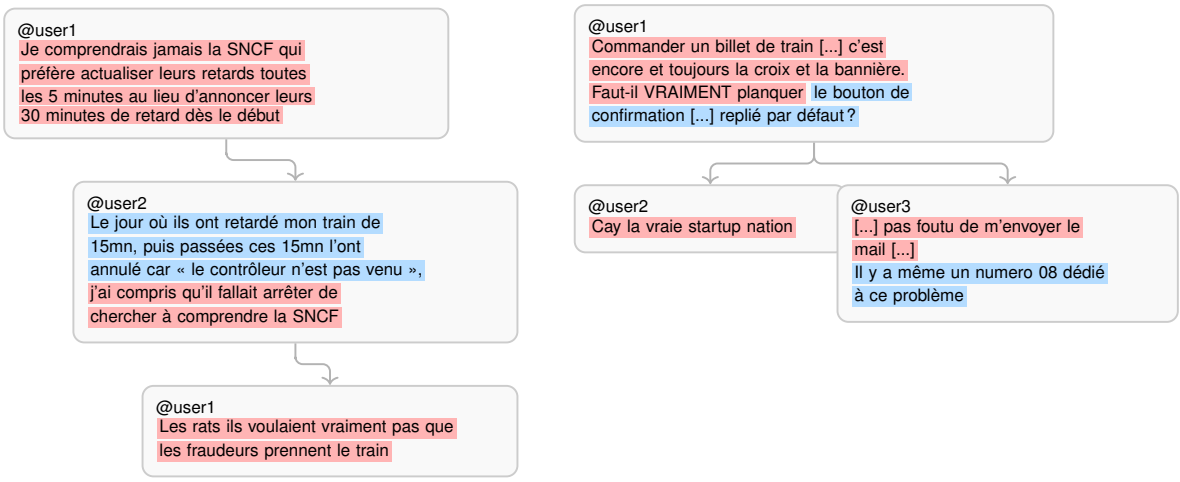

FIGURE 2 – Représentation en graphe de deux conversations annotées, centrées sur l'expérience des voyageurs. Les faits sont surlignés en bleu et les opinions sont surlignées en rouge. Cette visualisation met en évidence l'intrication des deux types de segments au sein d'un même message.

5 Conclusion

Cet article a présenté la constitution d'un corpus de conversations extrait du réseau social Bluesky, spécifiquement annoté pour distinguer les assertions factuelles des expressions d'opinions dans le contexte du transport ferroviaire. Malgré la complexité inhérente aux données de microblogging, l'analyse inter-annotateurs révèle une convergence solide sur l'identification des régions, appuyée par des métriques de chevauchement et de F1 élevées validant au moins partiellement la tâche. Le corpus final, riche de 1 117 messages et 1 620 segments annotés, offre un équilibre intéressant entre faits (38 %) et opinions (62 %), confirmant que les usagers ne se contentent pas de jugements de valeur mais étayent massivement leurs propos par des éléments d'expérience concrets et vérifiables. L'évaluation a toutefois mis en lumière les défis du problème de délimitation. La variabilité des frontières de segments illustrée par nos métriques de F1 calculées avec un seuil d'enchevêtrement à 0.6 (0.47) témoigne de la difficulté à s'accorder sur les limites exactes des expressions de faits et d'opinions dans ce type de texte. En outre, certains segments peuvent faire l'objet de désaccords sur le type, comme le montre la métrique γ_{cat} (0.2221). Cette confusion est possiblement due à notre choix de ne pas laisser la possibilité d'annoter un segment comme étant composé à la fois de faits et d'opinions. Nous avons effectivement pu remarquer que la description objective d'une situation (une panne, un retard) est en effet très souvent intriquée avec l'expression d'un jugement de valeur ou d'une émotion. Imposer un choix binaire, sans catégorie hybride, a pu contraindre les annotateurs à privilégier l'une ou l'autre de ces dimensions face à un énoncé mixte, ou à diverger sur la stratégie de découpage.

À court et moyen terme, ce corpus ouvre de nouvelles perspectives pour l'analyse des dynamiques argumentatives en ligne. Dans un premier temps, une analyse linguistique pourra être menée pour caractériser les faits des opinions, et les mettre en regard avec des travaux déjà existants. Dans un second temps, au-delà de l'analyse séquentielle ou intra-message, l'enjeu sera ensuite de croiser les dimensions langagières (la nature factuelle ou subjective de l'assertion) avec les dimensions interactives des conversations (la structure en arbre des *fils de discussion*). Il s'agira d'étudier comment la formulation d'un fait par un utilisateur déclenche des cascades d'opinions, ou à l'inverse, comment une opinion initiale est soutenue, nuancée ou réfutée par l'apport de nouveaux faits par d'autres utilisateurs.

Références

ALHINDI T., MURESAN S. & PREOTIUC-PIETRO D. (2020). Fact vs. opinion : the role of argumentation features in news classification. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6139–6149, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.540.

ALMERGE N., SANTELMO M., GÜL I., ASADI SARIJALOU A., LEBRET R., LAUGIER L. & ABERER K. (2025). Sagesse : A system for argument generation, extraction and structuring of social exchanges. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, p. 1024–1027, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3701551.3704122.

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596. DOI : 10.1162/coli.07-034-R2.

BALEDENT A. (2022). *De la complexité de l'annotation manuelle : méthodologie, biais et recommandations*. Theses, Normandie Université. HAL : tel-04011353.

BIBER D. & CONRAD S. (2000). Adverbial marking of stance in speech and writing. In S. HUNSTON & G. THOMPSON, Éds., *Evaluation in Text : Authorial Stance and the Construction of Discourse*, p. 56–73. Oxford University Press.

BIBER D., JOHANSSON S., LEECH G., CONRAD S. & FINEGAN E. (2007). *Longman Grammar of Spoken and Written English*. Longman, 6th édition.

BOSC T., CABRIO E. & VILLATA S. (2016). DART : a dataset of arguments and their relations on Twitter. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1258–1263, Portorož, Slovenia : European Language Resources Association (ELRA).

FEGER M. & DIETZE S. (2024). Taco – twitter arguments from conversations.

GROOTENDORST M. (2022). Bertopic : Neural topic modeling with a class-based tf-idf procedure.

HABERNAL I. & GUREVYCH I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, **43**(1), 125–179. DOI : 10.1162/COLI_a_00276.

KRIPPENDORFF K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology*, **25**, 47–76.

KRIPPENDORFF K., MATHET Y., BOUVRY S. & WIDLÃ¶CHER A. (2016). On the reliability of unitizing textual continua : Further developments. *Quality & Quantity : International Journal of Methodology*, **50**(6), 2347–2364. DOI : 10.1007/s11135-015-0266-1.

LEVY R., BILU Y., HERSHCOVICH D., AHARONI E. & SLONIM N. (2014). Context dependent claim detection. In J. TSUJII & J. HAJIC, Éds., *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 1489–1500, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

LI D., ROSE C., YUAN A. & ZHOU C. (2024). Estimating agreement by chance for sequence annotation. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5085–5097, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.278.

MATHET Y., WIDLÖCHER A. & MÉTIVIER J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, **41**(3), 437–479. DOI : 10.1162/COLI_a_00227.

MOENS M.-F., BOIY E., MOCHALES R. & REED C. (2007). Automatic detection of arguments in legal texts. p. 225–230. DOI : 10.1145/1276318.1276362.

STAB C. & GUREVYCH I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, **43**(3), 619–659. DOI : 10.1162/COLI_a_00295.

SUN W., LI M., DAVIS J., CABRIO E., VILLATA S. & MOENS M.-F. (2026). Weakly-supervised argument mining with boundary refinement and relation denoising. In M. AKHTAR, R. ALY, R. CAO, C. CHRISTODOULOPOULOS, O. COCARASCU, Z. GUO, A. MITTAL, M. SCHLICHTKRULL, J. THORNE & A. VLACHOS, Éds., *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, p. 1–12, Rabat, Morocco : Association for Computational Linguistics. DOI : 10.18653/v1/2026.fever-1.1.

SUNDRIYAL M., AKHTAR M. S. & CHAKRABORTY T. (2023). Overview of the claimscan-2023 : Uncovering truth in social media through claim detection and identification of claim spans.

SUNDRIYAL M., KULKARNI A., PULASTYA V., AKHTAR M. S. & CHAKRABORTY T. (2022). Empowering the fact-checkers ! automatic identification of claim spans on Twitter. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 7701–7715, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.525.

WIEBE J., WILSON T. & CARDIE C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, **39**(2), 165–210. DOI : 10.1007/s10579-005-7880-9.

ANNEXES

TABLE 3 – Statistiques du nombre d'utilisateurs par conversation

Moyenne	**Médiane**	**Min**	**Max**	**Q25**	**Q75**
5.76	1.0	1	598	1.0	4.0

TABLE 4 – Statistiques descriptives du corpus annoté

Catégorie	Métrique	Valeur
Statistiques générales	Nombre de conversations	279
	Nombre de messages uniques	1 117
Messages par conversation	Moyenne	4,00
	Médiane	1,0
	Minimum	1
	Maximum	45
Longueur du texte par message (mots)	Moyenne	25,53
	Médiane	24,0
	Minimum	1
	Maximum	64

TABLE 5 – Statistiques descriptives des annotations (Faits vs Opinions)

Catégorie	Métrique	Valeur
Volume des annotations	Total des segments annotés	1 620
	— Nombre de faits (FACT)	659
	— Nombre d'opinions (OPINION)	961
Répartition par message	Moyenne d'annotations par message	1,82
($N = 892$ messages)	Ratio moyen d'opinions	62,07 %
	Maximum d'annotations (un seul message)	6
Longueur des segments (mots)	Moyenne (FACT)	15,11
	Médiane (FACT)	12,0
	Moyenne (OPINION)	12,74
	Médiane (OPINION)	11,0