Approche hybride centrée sur l'humain pour l'évaluation explicable de la qualité de présentation des tweets organisationnels

Aya Hidri[1], Ines Saad[2, 3], Manel Ben Sassi[1], Henda Ben Ghezala[1]

[1] Laboratoire RIADI, Université de la Manouba
[2]Laboratoire MIS, Université de Picardie Jules Verne
[3]Amiens Business School

Résumé

Cet article propose un cadre hybride multicritère centré sur le concept de Human-In-The-Loop (HITL) pour l'évaluation explicable de la qualité de présentation textuelle des tweets organisationnels. L'approche combine des métriques de lisibilité classiques (Flesch Reading Ease, Kandel-Moles), une évaluation réalisée par le grand modèle de langage Llama-3.3-70B, et l'expertise humaine, agrégées via la méthode de la somme pondérée (SAW). Les résultats montrent que Llama-3.3-70B atteint une corrélation de Pearson de $r = 0{,}52$ avec les jugements de l'expert, surpassant les métriques classiques (FRE : $r = 0{,}17$; KM : $r = 0{,}21$). Ces résultats confirment l'intérêt d'une approche hybride intégrant l'intelligence artificielle et l'expertise humaine pour l'évaluation contextualisée de la qualité de présentation textuelle des tweets organisationnels.

Mots-clés

Qualité des données ; X (Twitter) ; Human-In-The-Loop ; grands modèles de langage (LLM) ; lisibilité textuelle.

Abstract

This paper proposes a hybrid multicriteria Human-In-The-Loop (HITL) approach for the explainable evaluation of the textual presentation quality of organizational tweets. The approach combines classical readability metrics (Flesch Reading Ease, Kandel-Moles), an evaluation by the large language model Llama-3.3-70B, and human expertise, aggregated via the Simple Additive Weighting (SAW) method. Our results show that Llama-3.3-70B achieves a Pearson correlation of $r = 0.52$ with expert judgments, outperforming classical metrics (FRE : $r = 0.17$; KM : $r = 0.21$). These results confirm the relevance of a hybrid approach combining artificial intelligence and human expertise for the contextualized assessment of the textual presentation quality of organizational tweets.

Keywords

Data quality ; X (Twitter) ; Human-In-The-Loop ; large language models (LLMs) ; text readability.

1 Introduction

X (anciennement Twitter) a révolutionné la communication numérique, permettant aux utilisateurs d'exprimer leurs opinions, de suivre l'actualité et de partager du contenu avec une audience mondiale. Avec plus de 440 millions d'utilisateurs actifs mensuels en 2024, cette plateforme génère quotidiennement près de 500 millions de tweets [2]. Cette quantité massive de données, souvent qualifiée de *big data*, a suscité une attention considérable dans la recherche académique [24].

Les entreprises et les organisations exploitent stratégiquement ces données pour surveiller leur réputation, évaluer la perception de leur marque et adapter leurs stratégies de communication[22]. Des entreprises telles que Nike ou Coca-Cola cherchent à s'assurer que leur positionnement est aligné avec les valeurs sociales actuelles et les attentes des consommateurs [17]. Ces entreprises doivent donc comprendre comment leurs parties prenantes réagissent aux tendances émergentes pertinentes pour leur secteur.

X intègre différents formats de contenu multimodaux (texte, images, vidéos et audio) afin d'améliorer l'engagement des utilisateurs et la visibilité des publications. Plusieurs travaux ont démontré que l'utilisation d'éléments visuels contribue à renforcer l'attractivité des messages et à stimuler l'interaction[18][14]. Toutefois, au-delà de ces éléments multimédias, la qualité textuelle demeure un facteur déterminant dans l'efficacité des communications organisationnelles sur les réseaux sociaux[6].

Cependant, la nature intrinsèque des tweets, souvent caractérisée par l'utilisation d'abréviations, d'emojis et de hashtags, pose des défis majeurs pour l'analyse sémantique. Une information mal structurée peut nuire à la compréhension du message, et limiter son utilité dans le cadre de la prise de décision organisationnelle.

Ces constats soulèvent une problématique centrale : comment évaluer de manière fiable, explicable et adaptée au contexte organisationnel la qualité de présentation textuelle des tweets, tout en tenant compte de leur brièveté et de leur style informel ?

C'est pourquoi nous proposons une approche hybride combinant l'automatisation et l'expertise humaine au sein d'un cadre multicritère, permettant de produire un score global explicable de qualité présentation textuelle.

Les questions de recherche suivantes guident cette étude :

QR1 : Dans le contexte des tweets organisationnels, dans quelle mesure les scores de lisibilité produits par Llama-

3.3-70B corrèlent-ils avec les jugements d'expert, comparativement aux métriques traditionnelles (Flesch Reading Ease, Kandel-Moles) ?
QR2 : Existe-t-il des relations significatives entre la lisibilité du texte et d'autres caractéristiques textuelles propres à la plateforme X (hashtags, emojis) ?
QR3 : Quel est l'apport d'une approche hybride intégrant l'expertise humaine dans l'évaluation de la qualité de présentation textuelle des tweets pour la prise de décision organisationnelle ?
Ces trois questions permettent d'adresser un double enjeu : d'une part, comprendre l'influence des éléments spécifiques à la plateforme X sur la lisibilité perçue ; d'autre part, évaluer la valeur ajoutée d'un cadre multicritère intégrant l'expertise humaine face aux approches automatiques. La réponse à ces questions apporte une contribution méthodologique originale à l'intersection de la qualité des données, de l'analyse des réseaux sociaux et des systèmes d'aide à la décision.
Le reste de cet article est organisé comme suit. La section 2 présente les travaux antérieurs. La section 3 introduit le cadre formel d'évaluation. La section 4 décrit les méthodes d'évaluation de la lisibilité. La section 5 présente les résultats et leur analyse. Enfin, la section 6 expose la conclusion et les perspectives de recherche futures.

2 Travaux antérieurs

L'analyse des contenus publiés sur les réseaux sociaux s'articule autour de trois axes principaux dans la littérature : l'impact des éléments visuels et textuels sur l'engagement, l'évaluation de la lisibilité des contenus textuels, et les approches automatiques fondées sur l'apprentissage automatique et les LLM. Ces trois axes constituent le socle théorique sur lequel repose notre contribution.

2.1 Éléments visuels, hashtags et emojis

Plusieurs études ont mis en évidence l'impact des éléments visuels tels que les images et les vidéos sur l'engagement des utilisateurs sur les réseaux sociaux[18][6]. Ces éléments sont traités plus rapidement que le texte, ce qui leur confère un avantage considérable en matière d'attractivité et d'impact émotionnel. Ils jouent également un rôle déterminant dans les décisions d'achat et dans l'élaboration des stratégies de communication digitale. À cet égard, des travaux empiriques ont démontré que les publications intégrant des images ou des vidéos génèrent significativement plus d'engagement positif de la part des utilisateurs [14]. Des recherches plus récentes confirment par ailleurs que la combinaison d'un texte bien rédigé et d'éléments visuels pertinents permet d'accroître la portée et la mémorabilité des messages organisationnels [1].
Au-delà des éléments visuels, d'autres travaux se sont intéressés aux caractéristiques textuelles spécifiques aux publications sur les réseaux sociaux, notamment à l'utilisation des hashtags et des emojis [25, 8]. Ces éléments contribuent à accroître la visibilité des publications, à augmenter le nombre d'impressions et à renforcer la notoriété de la marque. Les auteurs de [8] montrent que les hashtags servent principalement à augmenter la visibilité des publications et à renforcer l'identification des organisations sur les plateformes de réseaux sociaux. Par ailleurs, les auteurs de [11][15] ont mis en évidence que l'usage approprié des hashtags et des emojis permet de renforcer le ton émotionnel du message et d'améliorer la connexion avec l'audience, à condition toutefois de ne pas surcharger le texte au détriment de sa lisibilité.

2.2 Lisibilité des contenus numériques courts

Plus récemment, la lisibilité des contenus numériques a émergé comme un facteur déterminant de leur valeur perçue et de leur attractivité. Elle est définie dans la littérature comme la facilité avec laquelle un texte peut être compris par les lecteurs[19]. La lisibilité est ainsi décrite comme la « facilité de compréhension due au style d'écriture », ou encore comme le degré auquel un texte peut être appréhendé par les lecteurs[5].
La lisibilité dépend de plusieurs caractéristiques linguistiques couvrant différents niveaux du texte, notamment le vocabulaire utilisé, la longueur des mots, la complexité syntaxique, la longueur des phrases, ainsi que la cohésion textuelle [26]. Ces dimensions influencent directement la capacité des lecteurs à traiter efficacement l'information. Afin de mesurer la lisibilité de manière quantitative, plusieurs indices ont été proposés dans la littérature. Parmi les plus utilisés figurent l'indice de Flesch-Kincaid (FK) et le Gunning Fog Index (GFI) [10], qui estiment respectivement le niveau scolaire requis pour comprendre un texte en se basant sur la longueur moyenne des phrases et des mots. Ces indices visent à évaluer la difficulté de compréhension et à limiter les complexités inutiles dans la rédaction des contenus.
Cependant, la plupart des travaux sur la lisibilité se concentrent principalement sur les textes longs (contenus web ou pédagogiques), tandis que les contenus courts tels que les publications sur les réseaux sociaux restent peu étudiés [10]. Or, ces contenus présentent des spécificités importantes, notamment l'usage d'abréviations, de hashtags, d'emojis et de structures informelles, qui compliquent l'application des indices classiques [11]. Plus récemment, des travaux ont mis en évidence que certaines caractéristiques textuelles, notamment les hashtags et les mentions, sont négativement corrélées avec la lisibilité perçue des tweets [25]. Ces travaux suggèrent que la lisibilité ne constitue pas seulement un critère de qualité textuelle, mais également un levier stratégique pour les organisations dans leurs communications numériques.

2.3 Approches automatiques et grands modèles de langage

Face aux limitations des métriques classiques, des approches supervisées ont abordé l'évaluation de la lisibilité comme un problème de classification automatique, fondées sur les caractéristiques lexicales telles que la longueur des mots et la complexité lexicale[20][26]. Cependant, ces caractéristiques ne parviennent souvent pas à capturer le contexte des phrases. Certains travaux ont tenté

de pallier ces limitations en combinant des représentations n-grammes avec des méthodes d'apprentissage automatique notamment les Support Vector Machines (SVM), les Random Forest et les réseaux de neurones artificiels (ANN)[26][7]. L'application des méthodes d'apprentissage automatique à l'évaluation de la lisibilité textuelle présente un potentiel certain ; toutefois, ces approches requièrent des corpus d'entraînement annotés de grande envergure et présentent des limitations significatives en termes de généralisation interlangues et d'adaptation contextuelle. De surcroît, elles ne parviennent pas à capturer le contexte global ni la sémantique profonde du texte.

Les récentes avancées en traitement automatique du langage naturel (TALN) et en intelligence artificielle générative (IAG) ont ouvert de nouvelles perspectives pour l'évaluation automatique de la lisibilité des textes. Plusieurs travaux ont commencé à explorer le potentiel des grands modèles de langage (LLM) dans ce contexte [21][3][19][4]. Par exemple, les auteurs de [3] ont examiné l'applicabilité de ChatGPT en tant qu'outil d'évaluation de la lisibilité, en soumettant des manuels scolaires sélectionnés à une analyse textuelle par LLM. Cette étude démontre que les LLM constituent une alternative prometteuse aux formules traditionnelles, notamment pour les contextes pédagogiques spécifiques aux apprenants en langue étrangère. Dans le même contexte, les auteurs de [19] ont conduit une étude empirique analysant les capacités de génération textuelle de plusieurs LLM pré-entraînés, en adoptant la lisibilité comme métrique d'évaluation principale. Les résultats montrent que la lisibilité varie significativement selon le modèle utilisé et les contraintes imposées. Les auteurs soulignent toutefois que les métriques traditionnelles de lisibilité seules demeurent insuffisantes pour capturer l'ensemble des dimensions de la lisibilité, telles que la pertinence contextuelle et l'engagement du lecteur. Enfin, les auteurs de [4] ont démontré que les métriques traditionnelles, telles que le Flesch–Kincaid Grade Level (FKGL), présentent une faible corrélation avec les jugements humains. En revanche, les LLM montrent une meilleure corrélation avec les évaluations humaines. Ces résultats s'expliquent par la capacité des LLM à capturer des dimensions plus profondes de la lisibilité :le niveau de connaissances préalables requis pour comprendre un texte, la cohérence discursive et l'adéquation au lectorat cible. Ces aspects demeurent difficilement appréhendables par les formules syllabiques classiques.

Néanmoins, une limitation majeure subsiste : la plupart de ces méthodes ont été conçues pour des textes de longueur standard et ne sont pas directement adaptées aux contenus courts et informels des réseaux sociaux. De plus, peu de travaux proposent une approche intégrant explicitement les préférences des décideurs dans l'évaluation de la qualité textuelle des tweets organisationnels. C'est précisément cette double lacune tels que l'inadaptation des métriques classiques aux contenus courts et l'absence de prise en compte des préférences décisionnelles que notre travail cherche à combler en proposant un cadre hybride multicritère centré sur l'humain, combinant métriques classiques, évaluation par LLM (Llama-3.3-70B) et expertise humaine, offrant ainsi robustesse, explicabilité et alignement sur les besoins organisationnels. Dans ce contexte, cette étude vise à évaluer la qualité de présentation textuelle des tweets organisationnels. L'objectif est d'intégrer les préférences des décideurs afin de proposer une approche d'évaluation adaptée au contexte organisationnel. La méthodologie proposée repose sur une architecture hybride combinant l'expertise humaine via un paradigme HITL et un traitement algorithmique automatisé.

3 Évaluation de la qualité de présentation textuelle

Le critère de qualité de présentation textuelle a été consolidé par la validation d'un expert issu du domaine du marketing des médias sociaux et de la communication organisationnelle. Ce critère est évalué à travers sa dimension de lisibilité textuelle, ainsi que l'impact des hashtags et l'usage des emojis, qui constituent des éléments centraux du message dans les tweets. Il mesure la manière dont l'information est structurée et perçue par les différents publics ciblés : candidats potentiels, partenaires académiques et autres parties prenantes institutionnelles. L'approche proposée, illustrée par la Figure 1, s'articule en trois phases.

Phase 1 : Acquisition et préparation des données
Cette phase porte sur la constitution d'un corpus de 11 000 tweets issus des comptes officiels d'écoles de commerce françaises sur la plateforme X, avec extraction des éléments textuels, des hashtags, des emojis et des liens originaux.

Phase 2 : Evaluation hybride HITL
Cette phase combine une évaluation automatique de la lisibilité (FRE, Kandel-Moles et Llama-3.3-70B) et une intervention experte (HITL) consistant à sélectionner les tweets de référence via la méthode DRSA [9][23], à construire les sous-critères et à définir les poids associés.

Phase 3 : Validation et agrégation multicritère
Cette phase assure la validation par mise en correspondance des scores automatiques avec les jugements de l'expert sur les 50 tweets de référence, puis l'agrégation multicritère via la méthode SAW afin de produire un score explicable de la qualité de présentation textuelle.

Dans cette section, nous présentons le modèle mathématique multicritère fondant notre approche. La Section 3.1 établit les notations fondamentales. La Section 3.2 formalise la représentation d'un tweet organisationnel. La Section 3.3 présente la méthode de la somme pondérée (SAW).

3.1 Notations fondamentales

Soit $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ l'ensemble des m tweets à évaluer, où T_i $(i = 1, 2, \ldots, m)$ désigne un tweet individuel. Un tweet est une séquence de tokens :

$$T_i = \langle w_1, w_2, \ldots, w_k \rangle \tag{1}$$

où les tokens w_ℓ $(\ell = 1, 2, \ldots, k)$ peuvent appartenir à l'une des trois catégories textuelles : le contenu textuel, les hashtags ou les emojis.

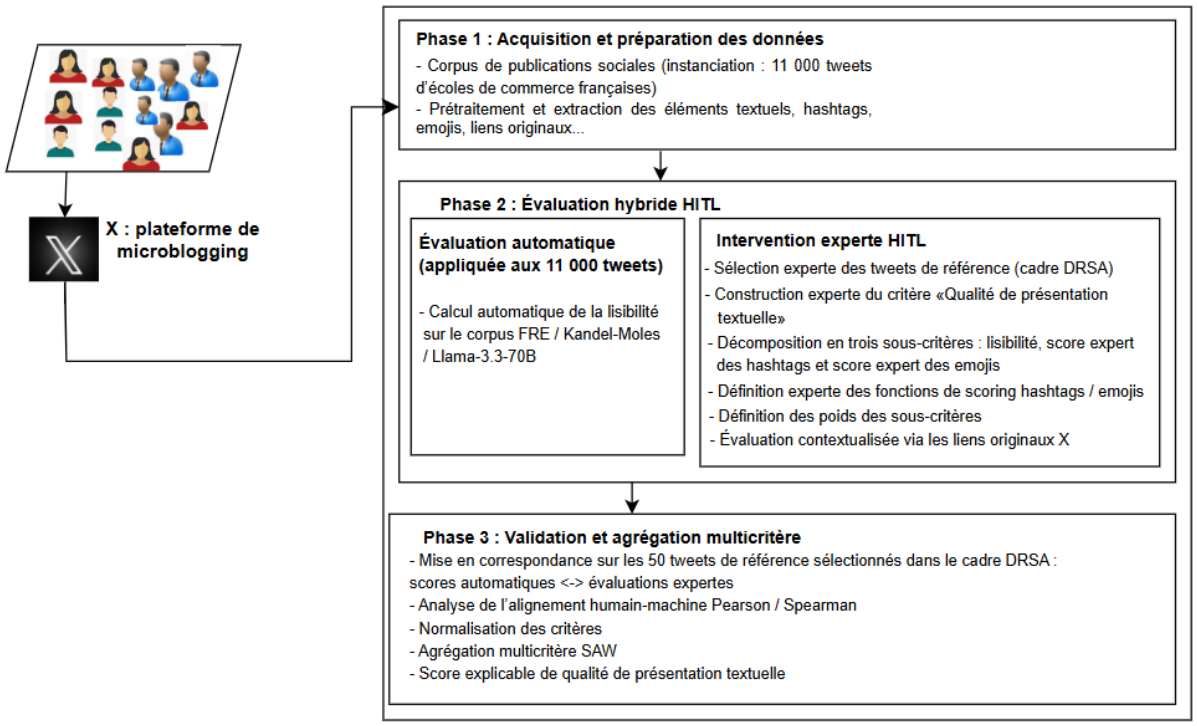

FIGURE 1 – Cadre hybride pour l'évaluation explicable de la qualité de présentation textuelle des tweets organisationnels.

Soit $\mathcal{G} = \{g_1, \ldots, g_n\}$ l'ensemble des n critères d'évaluation, où g_j $(j = 1, 2, \ldots, n)$ désigne le j-ème critère.

3.2 Représentation formelle d'un tweet

Un tweet $T_i \in \mathcal{T}$ est représenté par le triplet :

$$T_i \;=\; \big\langle\, \mathcal{C}(T_i),\; \mathcal{H}(T_i),\; \mathcal{E}(T_i) \,\big\rangle \tag{2}$$

où :

— $\mathcal{C}(T_i) \in \Sigma^*$ désigne la séquence ordonnée de tokens lexicaux porteurs du message principal, avec Σ^* l'ensemble de toutes les séquences finies de tokens ;

— $\mathcal{H}(T_i)$ désigne l'ensemble fini des hashtags (préfixés par #) assurant la catégorisation thématique et la découvrabilité du contenu ;

— $\mathcal{E}(T_i)$ désigne l'ensemble fini des emojis contribuant à l'expressivité non verbale et à la tonalité émotionnelle du message.

3.3 Agrégation SAW et score de présentation textuelle

Nous adoptons la méthode SAW [12] pour agréger les n critères en un score scalaire global. Cette méthode garantit la traçabilité de la contribution de chaque critère et l'interprétabilité directe du score par les décideurs.

Étape 1 : Construction de la matrice de décision
La matrice de décision initiale $X \in R^{m\times n}$ est définie comme suit :

$$X = (x_{ij}) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \tag{3}$$

où chaque ligne i correspond au tweet $T_i \in \mathcal{T}$, chaque colonne j au critère $g_j \in \mathcal{G}$, et $x_{ij} = g_j(T_i)$.

Étape 2 : Normalisation de la matrice de décision
Afin de rendre les n critères comparables sur une échelle commune $[0, 1]$, on suppose $x_{ij} > 0$ pour tout $i \in \{1, \ldots, m\}$ et $j \in \{1, \ldots, n\}$, ce qui garantit que les formules de normalisation ci-dessous sont bien définies. Chaque valeur de performance brute x_{ij} est normalisée selon la nature du critère g_j :

$$\tilde{x}_{ij} = \begin{cases} \dfrac{x_{ij}}{\max_i\, x_{ij}}, & \text{si } g_j \text{ est un critère bénéfice,} \\[2ex] \dfrac{\min_i\, x_{ij}}{x_{ij}}, & \text{si } g_j \text{ est un critère coût.} \end{cases} \tag{4}$$

Un critère est dit *bénéfice* lorsque des valeurs plus élevées sont préférables, et *coût* lorsque des valeurs plus faibles sont préférées. Les 3 critères retenus étant tous de type bénéfice, la matrice normalisée est :

$$\tilde{X} = (\tilde{x}_{ij}) \;\in\; [0, 1]^{m\times n}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n \tag{5}$$

Étape 3 : Définition du vecteur de pondération
Soit $W = (W_1, \ldots, W_n)^\top \in [0, 1]^n$ le vecteur de pondération, où W_j traduit l'importance relative du critère g_j, établi lors des sessions avec l'expert.

Ce vecteur vérifie la contrainte de normalisation :

$$\sum_{j=1}^{n} W_j = 1, \qquad W_j \geq 0 \quad \forall\, j \in \{1, \ldots, n\} \tag{6}$$

Étape 4 : Calcul du score global
Le score global de présentation textuelle S_i du tweet $T_i \in \mathcal{T}$ est défini par la somme pondérée des performances normalisées :

$$S_i \;=\; W^{\top} \tilde{x}_i \;=\; \sum_{j=1}^{n} W_j\, \tilde{x}_{ij}, \qquad i = 1, \ldots, m \tag{7}$$

où $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{in})^{\top} \in [0,1]^n$ est le vecteur des performances normalisées du tweet T_i, avec $S_i \in [0,1]$.

4 Méthodes d'évaluation de la lisibilité

Pour mesurer la lisibilité, deux métriques traditionnelles ont été employées : Flesch Reading Ease pour l'anglais et la formule de Kandel-Moles pour le français. Pour les tweets en anglais, l'indice Flesch Reading Ease mesure la difficulté en fonction de la longueur des phrases et de la densité syllabique[16]. Pour les tweets en français, la formule de Kandel-Moles est appliquée sur la base du nombre de mots et de syllabes[13]. Bien que les métriques classiques soient robustes, elles restent limitées aux caractéristiques de surface (longueur des mots et des phrases). Pour capturer la véritable difficulté linguistique, nous intégrons le modèle Llama-3.3-70B via l'API Groq.

4.1 Kandel–Moles pour le contenu en français

Pour évaluer la lisibilité des tweets en langue française, nous utilisons la formule de Kandel et Moles. Cette métrique est spécifiquement conçue pour la structure phonétique et syllabique du français, ce qui la rend particulièrement adaptée à l'analyse de contenus numériques en français. Contrairement aux indices Dale-Chall ou Gunning Fog, qui s'appuient sur de larges corpus de référence ou des listes de « mots difficiles » souvent peu fiables pour les contenus des réseaux sociaux en raison de l'argot et des abréviations, la formule de Kandel-Moles repose sur des mesures orthographiques et syllabiques.
La formule a été légèrement adaptée pour tenir compte du format court des tweets. Compte tenu de leur brièveté et de leur structure informelle, chaque tweet est traité comme une phrase fonctionnelle unique ($N_{\text{phrases}} = 1$). La formule est définie comme suit :

$$RE_{\text{KM}} = 209 - 1{,}15 \times \left(\frac{\#\text{Mots}}{\#\text{Phrases}}\right) - 68 \times \left(\frac{\#\text{Syllabes}}{\#\text{Mots}}\right)$$

où #Mots représente le nombre total de mots du tweet (y compris les hashtags), #Phrases désigne le nombre de phrases, et #Syllabes correspond au nombre total de syllabes. Le score résultant se situe généralement entre 0 et 100, les valeurs les plus élevées indiquant une meilleure facilité de lecture. Le décompte syllabique suit les conventions françaises standard, avec des ajustements spécifiques pour les terminaisons en « e » muet (ex. -e, -es) afin de refléter fidèlement le rythme oral et la réalité linguistique du contenu.

4.2 Flesch Reading Ease pour le contenu en anglais

Cet indice est particulièrement adapté aux textes courts dans la mesure où il ne dépend d'aucun dictionnaire de fréquences externe[11]. Il repose sur deux paramètres principaux : la longueur moyenne des phrases et la densité syllabique moyenne des mots. La formule est appliquée comme suit :

$$RE = 206{,}835 - 1{,}015 \left(\frac{\#\text{Mots}}{\#\text{Phrases}}\right) - 84{,}6 \left(\frac{\#\text{Syllabes}}{\#\text{Mots}}\right)$$

4.3 Évaluation basée sur les LLM (Llama-3.3-70B)

Bien que les métriques de lisibilité traditionnelles restent largement utilisées, elles demeurent limitées pour capturer la véritable difficulté linguistique, car elles reposent principalement sur des caractéristiques de surface telles que la longueur des mots et la structure des phrases. Pour pallier ces limites, les approches automatiques ont intégré des indicateurs linguistiques plus profonds, notamment la fréquence des mots, l'analyse morphologique et la structure syntaxique, ce qui a conduit au développement de modèles d'apprentissage automatique capables d'évaluer dynamiquement la lisibilité.
Plus récemment, l'émergence des grands modèles de langage (LLM) a introduit un nouveau paradigme dans l'évaluation de la lisibilité. Contrairement aux formules traditionnelles, les LLM adoptent une perspective holistique en capturant à la fois la cohérence sémantique et les relations discursives implicites. Plusieurs travaux ont exploré le potentiel de différents LLM open-source à cet effet, couvrant une grande variété d'architectures, de tailles de modèles et de données d'entraînement, et montrant que les modèles de grande taille offrent un meilleur alignement avec les jugements humains, notamment pour les contenus courts et informels. Dans cette étude, nous retenons le modèle *Llama-3.3-70B*. Ce modèle évalue la lisibilité de courts textes selon une échelle ordinale à cinq niveaux, dans laquelle une valeur plus élevée indique une meilleure lisibilité : Très difficile (1), Difficile (2), Assez difficile (3), Facile (4), Très facile (5).
Afin d'optimiser la qualité des évaluations produites par Llama-3.3-70B, nous avons expérimenté trois prompts distincts. Le prompt sélectionné ancre l'évaluation dans une perspective pédagogique, cohérente avec le contexte des écoles de commerce étudiées. La conversion des évaluations qualitatives en valeurs numériques permet une comparaison directe avec les métriques classiques et facilite l'intégration dans le cadre multicritère SAW.

TABLE 1 – Échelles de niveaux de lisibilité

Niveau	Métriques traditionnelles		LLM
	FRE (anglais)	KM (français)	Llama-3.3-70B
Très difficile	0–29	0–29	1
Difficile	30–49	30–49	2
Assez difficile	50–69	50–69	3
Facile	70–89	70–89	4
Très facile	90–100	90–100	5

5 Résultats et analyse

Cette section présente une analyse comparative des scores et niveaux de lisibilité des tweets organisationnels obtenus à l'aide de Llama-3.3-70B et de formules traditionnelles telles que Flesch Reading Ease (FRE) et Kandel-Moles (KM). Les observations mettent en évidence la pertinence des outils basés sur l'IA pour évaluer la lisibilité de contenus courts, informels et bilingues des tweets sur X.

Pour permettre une comparaison quantitative, les évaluations qualitatives produites par Llama-3.3-70B ont été converties en valeurs numériques selon l'échelle du Tableau 1, où une valeur plus élevée indique une meilleure lisibilité perçue, en cohérence avec les indices FRE et KM.

5.1 Description du jeu de données

Le corpus utilisé dans cette étude comprend 11 000 tweets collectés auprès des comptes officiels d'écoles de commerce françaises sur la plateforme X. Ces institutions utilisent activement X comme canal de communication stratégique pour diffuser des informations relatives à leurs programmes académiques, leurs événements institutionnels et leurs partenariats. La diversité des registres employés entre communication formelle, annonces d'événements et interactions avec les étudiants fait de ce corpus un terrain d'analyse représentatif des enjeux de qualité rédactionnelle en contexte organisationnel.

Un sous-ensemble de 50 tweets a été sélectionné à partir de ce corpus et annoté par un expert en communication numérique, occupant actuellement le poste de responsable de la communication au sein d'une école de commerce et disposant de plus de 20 ans d'expérience dans la gestion des réseaux sociaux. La taille de ce sous-ensemble a été volontairement limitée afin de permettre une analyse experte détaillée, cohérente et fondée sur une évaluation qualitative approfondie. L'expert a joué deux rôles distincts et indépendants dans cette étude. D'une part, il a défini les poids de la méthode SAW à travers un processus structuré d'élicitation, réalisé préalablement et indépendamment de la tâche d'annotation. D'autre part, il a annoté les 50 tweets en fournissant des jugements de lisibilité servant de référence de vérité terrain pour l'évaluation.

5.2 Analyse comparative des scores de lisibilité

Le Tableau 2 présente une comparaison des scores de lisibilité pour un échantillon de tweets, en confrontant les scores produits par les métriques traditionnelles (FRE et Kandel-Moles), les évaluations du modèle Llama-3.3-70B ainsi que les jugements de l'expert.

Dans l'ensemble, la classification produite par Llama-3.3-70B est cohérente avec celle de l'expert, notamment pour les tweets dont le niveau de difficulté est clairement défini. Par exemple, les tweets T_1, T_2, T_3 et T_9 sont jugés « Très facile » à la fois par Llama-3.3-70B et par l'expert. De même, les tweets T_7 et T_8 sont tous deux évalués comme « Facile » par le modèle et par l'expert. Le tweet T_5 est identifié comme « Difficile » par Llama-3.3-70B ainsi que par l'expert. Notons toutefois que le score FRE associé (8,87) le classerait dans la catégorie « Très difficile » selon les seuils présentés dans le Tableau 1. Cet écart illustre les limites de l'indice FRE pour les contenus courts et contextuels.

Des divergences subsistent néanmoins pour les tweets T_6, T_{10}, T_{28}, T_{43} et T_{50}. Le cas du Tweet 50 est particulièrement illustratif : son score KM (52,75) indique une difficulté intermédiaire selon la formule classique, tandis que l'expert l'évalue comme « Facile » et Llama-3.3-70B lui attribue un score de 3 (« Assez difficile »). Cette divergence entre les métriques automatiques d'une part, et le jugement de l'expert d'autre part, révèle que les formules traditionnelles peinent à rendre compte de la lisibilité perçue dans des contenus courts, ce qui met en évidence la nécessité de combiner l'automatisation et l'expertise humaine afin d'obtenir une évaluation plus complète et plus robuste de la qualité de présentation textuelle.

TABLE 2 – Comparaison des scores de lisibilité pour un échantillon de tweets

Tweet	FRE	KM	Llama-3.3-70B	Jugement expert
T_1	–	75,40	5	Très facile
T_2	–	72,15	5	Très facile
T_3	–	78,30	5	Très facile
T_4	–	45,20	3	Assez difficile
T_5	8,87	–	2	Difficile
T_6	10,13	–	1	Assez difficile
T_7	–	48,90	4	Facile
T_8	–	55,60	4	Facile
T_9	–	82,10	5	Très facile
T_{10}	12,16	–	1	Difficile
⋮	⋮	⋮	⋮	⋮
T_{28}	18,45	–	1	Assez difficile
T_{43}	–	41,30	3	Très difficile
⋮	⋮	⋮	⋮	⋮
T_{50}	–	52,75	3	Facile

5.3 Interprétation des résultats

La distribution des scores de lisibilité révèle des comportements distincts entre les trois approches. Les métriques traditionnelles reposent sur des caractéristiques linguistiques de surface telles que la longueur des phrases ou la densité

syllabique. En revanche, Llama-3.3-70B intègre des informations sémantiques et contextuelles, lui permettant d'évaluer la lisibilité de façon plus proche de la perception humaine. Cette distinction reflète la différence entre une mesure de la forme syntaxique du texte (approche classique) et une évaluation de son sens perçu dans un contexte donné (approche LLM).

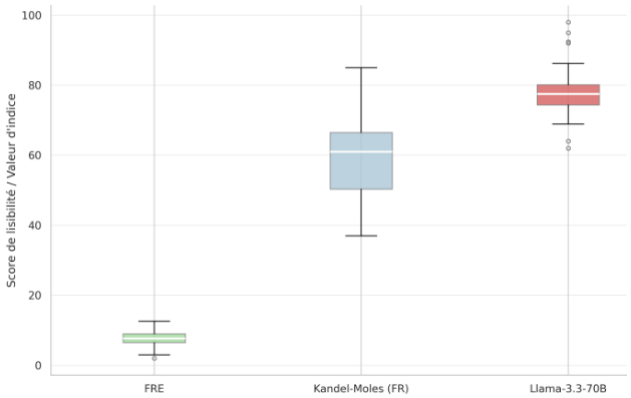

FIGURE 2 – Distribution des scores de lisibilité

5.4 Caractéristiques textuelles spécifiques à X

Les tweets contiennent des éléments textuels spécifiques qui les distinguent des textes écrits traditionnels, notamment les emojis, les hashtags, les mentions (@) et les hyperliens. Ces éléments sont largement utilisés pour transmettre des émotions, structurer l'information ou interagir avec d'autres utilisateurs. Dans le contexte de l'évaluation de la lisibilité, ces caractéristiques peuvent influencer la compréhension des tweets.

Emojis : Les emojis sont des symboles visuels utilisés pour enrichir les messages textuels en fournissant des indices contextuels ou émotionnels supplémentaires. Afin d'évaluer l'influence des emojis sur la lisibilité des tweets, ces derniers ont été regroupés en cinq catégories définies par l'expert : 0, 1, 2, 3 et $>$ 3 emojis. Le Tableau 3 présente, pour chaque catégorie, le nombre de tweets, le score moyen de lisibilité (SM), ainsi que les écarts-types (σ) correspondants.

TABLE 3 – Scores de lisibilité selon le nombre d'emojis

Nombre d'emojis	Nombre de tweets	SM	Écart-type
0 emoji	9 050	2,38	0,38
1 emoji	1 310	2,46	0,30
2 emojis	351	2,50	0,28
3 emojis	137	2,50	0,26
$>$ 3 emojis	109	2,42	0,46
Total	**11 000**	**2,40**	**0,37**

Les scores de lisibilité restent globalement stables pour toutes les catégories, avec une variation de 2,38 à 2,50 sur l'échelle 1–5. Les tweets contenant 2 ou 3 emojis présentent un score légèrement supérieur ($SM = 2{,}50$), tandis que les tweets sans emoji affichent le score le plus faible ($SM = 2{,}38$). Toutefois, ces écarts demeurent minimes et ne permettent pas de conclure à un effet significatif du nombre d'emojis sur la lisibilité perçue. Les faibles écarts-types (compris entre 0,26 et 0,46) témoignent d'une relative homogénéité des scores au sein de chaque catégorie, à l'exception de la catégorie $>$ 3 emojis**,** qui présente une dispersion légèrement plus élevée ($\sigma = 0{,}46$), imputable à la plus grande variabilité des tweets fortement chargés en emojis.

Hashtags : Les hashtags sont utilisés dans les tweets comme marqueurs de métadonnées permettant de référencer des sujets ou des thèmes et de rendre le contenu facilement découvrable sur les réseaux sociaux. Bien qu'ils permettent de catégoriser l'information et d'améliorer la visibilité du contenu, un nombre excessif de hashtags peut réduire la clarté du message et affecter négativement sa lisibilité. Dans le cadre de cette étude, les tweets ont été regroupés en fonction du nombre de hashtags qu'ils contiennent, selon une classification définie avec l'expert.

Le Tableau 4 présente, pour chaque catégorie, le nombre de tweets et l'écart de score de lisibilité par rapport aux tweets sans hashtag, calculé sur les trois méthodes.

TABLE 4 – Écarts de score de lisibilité selon les catégories de hashtags

Nombre de hashtags	Nombre de tweets	FRE	KM	Llama-3.3-70B
0	5 221	-	-	-
1	1 898	−0,47	−0,43	−0,12
2–4	2 808	−0,56	−0,51	−0,18
≥ 5	1 073	−0,76	−0,69	−0,27
Total	**11 000**			

La Figure 3 illustre la distribution des écarts de score de lisibilité pour les trois approches étudiées (FRE, Kandel-Moles et Llama-3.3-70B), calculée sur l'ensemble des **5 779** tweets contenant au moins un hashtag (**52,54** % du corpus). On observe que la majorité de la distribution se concentre dans les valeurs négatives pour les trois méthodes, ce qui suggère que l'inclusion de hashtags tend globalement à complexifier la structure du message et à réduire le score de lisibilité.

Toutefois, les écarts observés diffèrent selon la méthode : FRE et Kandel-Moles produisent des écarts plus importants que Llama-3.3-70B, qui présente une distribution plus stable. Cela s'explique par le fait que les métriques classiques pénalisent davantage les mots concaténés sans espaces, caractéristique des hashtags, tandis que Llama-3.3-70B est capable d'en inférer le sens dans le contexte global du tweet.

Bien que la présence de hashtags puisse modifier de manière significative le score de lisibilité d'un tweet, leur im-

pact sur la distribution globale de l'échantillon demeure relativement limité, notamment pour Llama-3.3-70B.
À titre d'illustration, considérons le tweet suivant :

> *#BusinessSchool #MBA #HigherEducation #Leadership — Join our MBA program and boost your career. Applications are now open !*

Pour ce type de tweet, l'inclusion des hashtags entraîne un écart moyen de $\Delta = -0{,}76$ point selon les métriques classiques, contre $\Delta = -0{,}27$ point pour Llama-3.3-70B, par rapport au texte brut. Cette comparaison illustre que les hashtags complexifient davantage la structure formelle mesurée par FRE et Kandel-Moles, tout en enrichissant la sémantique thématique perçue par Llama-3.3-70B.

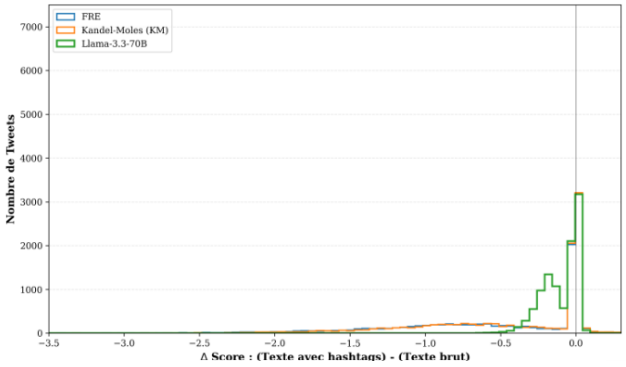

FIGURE 3 – Distribution des écarts de score de lisibilité selon le nombre de hashtags

6 Analyse de corrélation

Pour évaluer l'alignement entre les métriques automatisées et les préférences de l'expert, une matrice de corrélation complète a été calculée en utilisant à la fois le coefficient de Pearson (r) et celui de Spearman (ρ). Ces deux mesures permettent de quantifier le degré d'accord entre les évaluations automatisées et les jugements de l'expert, en couvrant à la fois les relations linéaires et les relations monotones ordinales.
Le coefficient de corrélation produit-moment de Pearson (r) mesure la relation linéaire entre deux variables continues :

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{8}$$

où x_i et y_i représentent les valeurs individuelles de chaque méthode, $\bar{x}$ et $\bar{y}$ leurs moyennes respectives, et N le nombre total d'observations.
Compte tenu de la nature ordinale des niveaux de lisibilité, le coefficient de corrélation de rang de Spearman (ρ) a également été employé :

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{9}$$

où d_i est la différence entre les rangs de chaque observation i, et N le nombre total d'observations.

La complémentarité de ces deux coefficients garantit une évaluation robuste : le coefficient de Pearson capture la cohérence des écarts numériques, tandis que le coefficient de Spearman valide la stabilité de l'ordre de classement, dimension cruciale pour les systèmes d'aide à la décision ordinaux.

TABLE 5 – Corrélations entre les jugements de l'expert et les caractéristiques textuelles

Caractéristique / Modèle	**Pearson** r	**Spearman** ρ
Flesch Reading Ease	0,17	0,22
Kandel-Moles	0,21	0,23
Nombre de hashtags	−0,23	−0,05
Nombre d'emojis	0,04	0,09
Efficacité des emojis	0,11	0,13
Mots par phrase	−0,16	−0,27
Hapax legomena (%)	0,36	0,36
Fautes d'orthographe (%)	0,27	0,21
Parenthèses (%)	−0,30	−0,27
Abréviations (%)	−0,03	−0,11
Références croisées (%)	−0,15	−0,21
Difficulté grammaticale	−0,09	−0,14
Polysémie (%)	−0,04	−0,13
Llama-3.3-70B	**0,52**	**0,50**

Les résultats présentés dans le Tableau 5 confirment que la lisibilité des tweets organisationnels constitue un construit multidimensionnel. Le modèle *Llama-3.3-70B* se distingue nettement ($r = 0{,}52$, $\rho = 0{,}50$), en atteignant un niveau d'alignement avec les jugements de l'expert supérieur à celui obtenu par les métriques classiques. La faible corrélation de FRE ($r = 0{,}17$) souligne son inadéquation structurelle pour les contenus courts, où la longueur des phrases est contrainte par la limite de caractères de la plateforme. La corrélation modérée de KM ($r = 0{,}21$) reflète sa meilleure adaptation aux spécificités phonétiques du français, mais confirme néanmoins ses limites pour les contenus informels et multimodaux des réseaux sociaux.
Parmi les indicateurs textuels complémentaires, les *hapax legomena* ($r = 0{,}36$) constituent le prédicteur de complexité le plus fort après *Llama-3.3-70B*, confirmant que la diversité lexicale est un signal pertinent de la difficulté perçue. À l'inverse, la présence de parenthèses ($r = -0{,}30$) est négativement corrélée avec les jugements de l'expert, confirmant que ces éléments perturbent la fluidité de lecture dans le format court du tweet.
Par ailleurs, l'analyse des caractéristiques spécifiques à la plateforme X révèle des résultats nuancés. Le nombre de hashtags ($r = -0{,}23$) est négativement corrélé avec les jugements de l'expert, ce qui confirme que leur usage excessif nuit à la lisibilité perçue dans le format court du tweet. En revanche, les emojis présentent un effet inverse. Leur présence montre une corrélation légèrement positive ($r = 0{,}04$, $\rho = 0{,}09$) avec les jugements de l'expert, suggé-

rant qu'un usage modéré des emojis contribue positivement à la lisibilité perçue.

7 Évaluation globale de la qualité de présentation textuelle par SAW

Cette section présente l'application numérique de la méthode SAW à un échantillon représentatif de tweets issus du corpus. Le calcul se déroule en quatre étapes : construction de la matrice de décision, normalisation, agrégation pondérée et classement final.

7.1 Définition des critères et des poids

Les trois critères retenus et leurs poids W_j ont été définis lors de la session de validation avec l'expert, conformément à la contrainte de normalisation $\sum_{j=1}^{3} W_j = 1$.
Le Tableau 6 présente ces critères et les poids associés.

TABLE 6 – Critères d'évaluation et poids définis par l'expert (HITL)

	Critère	**Type**	**Poids** W_j
g_1	Lisibilité	Bénéfice	0,60
g_2	Présence des hashtags	Bénéfice	0,20
g_3	Usage des emojis	Bénéfice	0,20

7.2 Normalisation de la matrice de décision

Une normalisation des scores bruts est nécessaire afin de les rendre comparables sur une échelle commune $[0, 1]$. Chaque performance brute x_{ij} est normalisée selon la nature du critère g_j, conformément à l'Équation (4). Dans le cas présent, tous les critères retenus étant de type bénéfice, une valeur plus élevée indiquant une meilleure qualité, la normalisation s'effectue par rapport au maximum observé sur chaque critère.
À titre d'illustration, les deux cas extrêmes du critère g_1 sont présentés ci-dessous, couvrant respectivement le tweet le plus lisible et le moins lisible de l'échantillon. Le maximum observé sur ce critère est $x_1^* = \max_i\{x_{i1}\} = 5{,}0$ (score Llama-3.3-70B maximal, niveau Très facile).
Pour le tweet T_9, qui présente la lisibilité maximale ($x_{9,1} = 5{,}0$, niveau Très facile) :

$$\tilde{x}_{9,1} = \frac{x_{9,1}}{x_1^*} = \frac{5{,}0}{5{,}0} = 1{,}000 \tag{10}$$

Pour le tweet T_6, dont le score Llama-3.3-70B est le plus faible de l'échantillon ($x_{6,1} = 1{,}0$, niveau Très difficile) :

$$\tilde{x}_{6,1} = \frac{x_{6,1}}{x_1^*} = \frac{1{,}0}{5{,}0} = 0{,}200 \tag{11}$$

Ce calcul est appliqué de manière identique à l'ensemble des tweets et pour chacun des trois critères g_1, g_2 et g_3. La matrice normalisée complète $\tilde{X} = (\tilde{x}_{ij}) \in [0, 1]^{m\times 3}$, avec m tweets en lignes et $n = 3$ critères en colonnes, est présentée dans le Tableau 7.

7.3 Agrégation pondérée et classement

Le score global S_i de chaque tweet T_i est calculé conformément à l'Équation (7). Le Tableau 7 présente les scores normalisés $\tilde{x}_{ij}$ et le score global S_i pour l'échantillon de tweets évalués.

TABLE 7 – Scores SAW appliqués à l'échantillon de tweets

Tweet	$\tilde{x}_{i1}$	$\tilde{x}_{i2}$	$\tilde{x}_{i3}$	S_i
T_1	1,000	0,600	1,000	0,920
T_2	1,000	1,000	0,556	0,911
T_3	1,000	1,000	0,556	0,911
T_4	0,600	1,000	0,556	0,671
T_5	0,400	1,000	0,778	0,596
T_6	0,200	1,000	0,556	0,431
T_7	0,800	1,000	0,556	0,791
T_8	0,800	1,000	0,556	0,791
T_9	1,000	1,000	1,000	**1,000**
T_{10}	0,200	1,000	0,556	0,431
⋮	⋮	⋮	⋮	⋮
T_{28}	0,200	1,000	0,556	0,431
T_{43}	0,600	1,000	0,556	0,671
⋮	⋮	⋮	⋮	⋮
T_{50}	0,600	1,000	0,556	0,671

Le tweet T_9 obtient le score maximal ($S_9 = 1{,}000$), reflétant une excellence sur les trois critères : lisibilité maximale évaluée par Llama-3.3-70B ($\tilde{x}_{9,1} = 1{,}000$, niveau Très facile), le score des hashtags optimale ($\tilde{x}_{9,2} = 1{,}000$) et le score des emojis maximale ($\tilde{x}_{9,3} = 1{,}000$), en cohérence avec le jugement expert *Très facile*.
À l'inverse, les tweets T_6, T_{10} et T_{28} présentent les scores les plus faibles ($S_i = 0{,}431$), en raison d'une lisibilité très limitée ($\tilde{x}_{i1} = 0{,}200$, score Llama-3.3-70B = 1, niveau Très difficile).
Ces résultats confirment l'effet de compensation partielle inhérent à la méthode SAW : une faiblesse sur le critère dominant g_1 (lisibilité, $W_1 = 0{,}60$) ne peut être que partiellement compensée par les critères secondaires g_2 et g_3. La décomposition additive du score S_i offre ainsi une explicabilité directement opérationnelle, permettant à l'expert de formuler des recommandations ciblées en vue d'optimiser la communication institutionnelle sur la plateforme X.

8 Conclusion

Dans cet article, nous avons présenté une approche hybride centrée sur l'humain pour l'évaluation explicable de la qualité de présentation textuelle des tweets organisationnels. En combinant l'expertise humaine avec les capacités d'analyse du modèle Llama-3.3-70B au sein d'un cadre multi-critère SAW, nous avons formalisé trois critères pertinents lisibilité, la présence des hashtags et usage des emojis et

produit des scores de qualité à la fois robustes, explicables et contextualisés.

Les résultats expérimentaux montrent que le modèle Llama-3.3-70B surpasse les métriques classiques en capturant efficacement les nuances sémantiques et contextuelles des messages courts, avec une corrélation de Pearson $r = 0{,}52$ avec les jugements de l'expert. Cette capacité à aligner les scores automatiques sur l'évaluation humaine témoigne de la pertinence d'un cadre hybride qui tire parti à la fois de la cohérence des modèles de langue et du jugement expert. Par ailleurs, l'analyse des éléments spécifiques à la plateforme X (hashtags et emojis) a confirmé leur rôle non négligeable sur la lisibilité et la complexité perçue des messages : leur prise en compte explicite dans le modèle multicritère améliore la finesse de l'évaluation par rapport à des approches purement lexicales. Cette approche offre ainsi un compromis entre automatisation et compréhension fine du contenu, répondant directement aux besoins des décideurs pour la prise de décision stratégique en matière de communication numérique.

Enfin, cette recherche ouvre la voie à plusieurs perspectives. Tout d'abord, l'intégration de la qualité multimodale (image, vidéo) et la validation sur d'autres types d'organisations (collectivités, ONG, institutions publiques) élargiront la portée et la généralité de l'approche. Ensuite, l'extension vers l'évaluation en temps réel des flux de données sur les réseaux sociaux ouvrirait la voie à des systèmes de veille et d'aide à la décision opérationnels. En combinant le raisonnement humain et l'intelligence artificielle, cette approche constitue une avancée significative vers des systèmes de support décisionnel plus explicables et fiables dans l'analyse des données des réseaux sociaux à des fins organisationnelles.

Références

[1] I. Fotovat Ahmadi, A. Waltenrath, and C. Janze. Congruency and users' sharing on social media platforms : A novel approach for analyzing content. *Journal of Advertising*, 51(4) :489–507, 2022.

[2] A. Albladi, M. Islam, and C. Seals. Sentiment analysis of twitter data using nlp models : A comprehensive review. *IEEE Access*, 8 :122199–122219, 2020.

[3] B. Basaran. Enhancing readability assessment for language learners : A comparative study of ai and traditional metrics in german textbooks. *European Journal of Educational Research*, 15(1) :101–119, 2025.

[4] I. Cachola, D. Khashabi, and M. Dredze. Evaluating the evaluators : Are readability metrics good measures of readability ? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[5] E. Dale and J. S. Chall. The concept of readability. *Elementary English*, 26(1) :19–26, 1949.

[6] S. W. Davis, C. Horváth, A. Gretry, and N. Belei. Say what ? how the interplay of tweet readability and brand hedonism affects consumer engagement. *Journal of Business Research*, 100 :150–164, 2019.

[7] A. Gaydhani, V. Doma, S. Kendre, et al. Detecting hate speech and offensive language on twitter using machine learning : An n-gram and tf-idf based approach. *arXiv preprint arXiv :1809.08651*, 2018.

[8] D. C. Gkikas, K. Tzafilkou, P. K. Theodoridis, A. Garmpis, and M. C. Gkikas. How do text characteristics impact user engagement in social media posts : modeling content readability, length, and number of hashtags in facebook. *International Journal of Information Management Data Insights*, 2(1) :100067, 2022.

[9] S. Greco, B. Matarazzo, and R. Slowinski. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129(1) :1–47, 2001.

[10] A. Ismail, K. S. Kuppusamy, A. Kumar, and P. K. Ojha. Connect the dots : Accessibility, readability and site ranking – an investigation with reference to top ranked websites of government of india. *Journal of King Saud University - Computer and Information Sciences*, 31(4) :528–540, 2019.

[11] P. Jacob and A. L. Uitdenbogerd. Readability of twitter tweets for second language learners. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 19–27, 2019.

[12] I. Kaliszewski and D. Podkopaev. Simple additive weighting – a metamodel for multiple criteria decision analysis methods. *Expert Systems with Applications*, 2016.

[13] L. Kandel and A. Moles. Application de l'indice de flesch à la langue française. *Cahiers d'Études de Radio-Télévision*, 19 :253–274, 1958.

[14] F. Khan, Z. Hafeez, A. Ijaz, and R. Shaheen. Predicting factors to get maximum social media engagement of customers for brand building. In *Proceedings of the Al Yamamah University Engineering Forum (YUENG)*, Riyadh, Saudi Arabia, 2019.

[15] M. L. Khan, M. Wasim, and A. Kaur. The role of emojis and hashtags in social media engagement : A sentiment and communication perspective. *Social Network Analysis and Mining*, 14(1) :1–15, 2024.

[16] J. P. Kincaid, R. P. Fishburne Jr., and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Research Branch, Millington, TN, USA, 1975.

[17] G. S. Kumar, N. Y. Reddy, R. G. Mudhiraj, and S. Ch. Prasad. Deep learning-based social media trend analyzer to predict trends over time. *Journal of Science Engineering Technology and Management Science*, 2(7S) :69–79, 2025.

[18] R. Kumar, R. Sinha, S. Saha, and A. Jatowt. Extracting the full story : A multimodal approach and dataset to crisis summarization in tweets. *IEEE Transactions on Computational Social Systems*, 2024.

[19] F. Marulli, L. Campanile, M. S. De Biase, S. Marrone, L. Verde, and M. Bifulco. Understanding readability of large language models output : An empirical analysis. *arXiv preprint*, 2024.

[20] W. Pan, X. Li, X. Chen, and R. Xu. Textual form features for text readability assessment. *Natural Language Processing*, 31(3) :800–841, 2025.

[21] G. M. Pascoal, M. van den Bosch, O. Viberg, and J. Wong. Improving text readability to support student comprehension and learning : An llm-powered approach. In *Two Decades of TEL : From Lessons Learnt to Challenges Ahead*, pages 291–305, September 2025.

[22] A. T. Rimadewi, Y. Azis, D. Sari, and I. Soemaryani. Social media reporting : How to do it right for strategic decision making. *Journalism and Media*, 6(4), 2025.

[23] I. Saad. Explainable AI and Multicriteria Decision Support : Real World Applications and Perspectives. In *Proceedings of EURO 2025*, pages 22–25, Leeds, United Kingdom, June 2025.

[24] C. Salvatore, S. Biffignandi, and A. Bianchi. Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156 :601–630, 2021.

[25] I. Temnikova, S. Vieweg, and C. Castillo. The case for readability of crisis communications in social media. In *Proceedings of the 24th International Conference on World Wide Web Companion (WWW '15 Companion)*, pages 1245–1250, 2015.

[26] L. Zhang, J. Abhani, J. B., A. Yadav, M. S. Ab Yajid, F. Mowafaq, and S. Vats. Automatic text readability assessment for educational content based on graph representation learning. *Scientific Reports*, 2026.