Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Analyse de la perception de l'offre INTERCITÉS de jour : Classification multi-étiquettes des émotions dans les tweets

Chang Liu[1] Hélène Flamein[2] Luce Lefeuvre[2] Fanny Hanen[1]
(1) SNCF Voyageurs INTERCITÉS Direction Marketing, 2 rue Traversière, 75012 PARIS
(2) SNCF DTIPG, 1-3 avenue François Mitterrand, 93210 Saint-Denis, France
ext.chang.liu@sncf.fr, helene.flamein2@gmail.com, luce.lefeuvre@sncf.fr, fanny.hanen@sncf.fr

RÉSUMÉ

La Direction Marketing de SNCF Voyageurs INTERCITÉS souhaite améliorer l'expérience des voyageurs en procédant à l'analyse automatique de la perception de son offre à travers les ressentis partagés sur les réseaux sociaux. L'un des axes de notre recherche se focalise sur la détection des émotions en multi-étiquettes qui traduisent cette perception. Pour accomplir cette tâche, nous ajustons tout d'abord un modèle de langue pré-entraîné à l'aide d'un corpus préalablement annoté en émotions, puis nous le spécialisons sur notre corpus, axé sur le contexte ferroviaire d'INTERCITÉS. Notre approche obtient un F1-Micro score de 0,55, un F1-Macro score de 0,44 et une exactitude de 0,826.

ABSTRACT

Analysis of the perception of the INTERCITÉS day train service : Multi-Label classification of emotions in tweets

The Marketing Department of SNCF Voyageurs INTERCITÉS aims to improve the passenger experience by conducting an automatic analysis of the perception of its service through feelings shared on social media. Our research focuses on the detection of multi-label emotions that reflect this perception. To accomplish this task, we first adjust a pre-trained language model using a corpus previously annotated with emotions. Then we specialize our model on our corpus, specific to the INTERCITÉS railway context. Our approach achieves a F1-Micro score of 0.55, a F1-Macro score of 0,44, and an accuracy of 0.826.

MOTS-CLÉS : Perception, Détection des émotions, Classification multi-étiquettes, CamemBERT, Mesures d'évaluation.

KEYWORDS: Perception, Emotion Detection, Multi-label Classification, CamemBERT, Evaluation Measures.

1 Introduction

L'amélioration constante de l'expérience client demeure une préoccupation centrale pour les entreprises opérant dans le secteur des services, et plus particulièrement dans le domaine du transport ferroviaire. Dans cette ère numérique, où les clients comme les non-clients expriment leurs opinions et ressentis de manière instantanée et publique sur les plateformes en ligne, l'analyse des données issues des réseaux sociaux offre de nouvelles perspectives pour mieux comprendre les attentes, les

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 323–337.

préoccupations et les satisfactions des voyageurs. En nous basant sur un corpus de tweets concernant l'offre INTERCITÉS de jour, notre recherche s'attache à analyser les émotions des utilisateurs telles qu'elles transparaissent dans les tweets. L'analyse des émotions des tweets est une tâche complexe, non seulement en raison de leur nature textuelle, mais aussi parce qu'un seul tweet peut véhiculer des émotions complexes, définies comme une combinaison d'émotions simples ou de processus cognitifs plus nuancés, tels que l'amour ou la culpabilité. L'annotation des émotions en multi-étiquettes semble nécessaire et permet de saisir la nuance des messages humains. Dans un tweet, un auteur peut en effet exprimer des avis divergents sur différentes thématiques. Par exemple, dans le tweet suivant issu de notre corpus INTERCITÉS, «@Intercites Bah t'es bien t'es coincé tu ne peux plus bosser», l'auteur exprime à la fois de l'«angoisse» et de la «colère». Nous avons ainsi exploré les stratégies permettant de surmonter la complexité de la détection de plusieurs émotions au sein d'un même tweet. De plus, nous nous sommes interrogées sur les moyens de gérer un corpus d'étude non annoté. En réponse à ces défis, un modèle de classification multi-étiquettes a été entraîné sur un corpus préalablement annoté en émotions selon une typologie adaptée à nos données et aux objectifs du projet.

2 État de l'art : détection des émotions dans les tweets

Dans le domaine de l'analyse des émotions, plusieurs auteurs se sont intéressés à leur classification, notamment en psychologie. Ainsi, Ekman (1992) identifie six émotions fondamentales : la colère, le dégoût, la peur, la joie, la tristesse et la surprise. Pour classifier les émotions, Plutchik (1980), s'inspirant du modèle d'Ekman, propose une visualisation en forme de roue qui comprend quatre ensembles bipolaires : joie et tristesse ; colère et peur ; confiance et dégoût ; surprise et anticipation. En TAL, les méthodes élémentaires de détection des émotions reposent sur l'utilisation de lexiques pour identifier des émotions liées à des états psychologiques via des mots-clés, une approche soulignée dans l'étude de Rabeya *et al.* (2017). Deux lexiques sont principalement utilisées dans ces approches : WordNet-Affect (Strapparava *et al.*, 2004) et NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013) également appelé EmoLex.

Les données issues des réseaux socio-numériques, et plus particulièrement de Twitter/X, ont donné lieu à de nombreux travaux en *sentiment analysis*, *emotion recognition* ou *opinion mining*, notamment parce qu'ils sont un lieu de polémique. D'un point de vue linguistique, les tweets peuvent inclure des textes linéaires simples, des émoticônes, des liens URL vers des sites externes, ainsi que des éléments spécifiques aux réseaux sociaux tels que les hashtags (marqués par un #) pour organiser l'information et les pseudos (précédés de @) qui identifient les utilisateurs (Paveau, 2017). D'après Baziotis *et al.* (2018), l'utilisation de ces formes de communication spécifiques éloigne les tweets des structures linguistiques traditionnelles et rend leur traitement complexe.

Ainsi, diverses approches sont étudiées pour détecter les émotions dans les tweets. Lora *et al.* (2020) évaluent plusieurs techniques d'apprentissage automatique, comme les modèles bayésiens naïfs, les SVM (*Support Vector Machine*), et la Régression Logistique, ainsi que diverses architectures de réseaux de neurones. Les méthodes d'apprentissage automatique affichent un score F1 supérieur ; toutefois, l'approche utilisant les CNN (*Convolutional Neural Networks*) avec des plongements lexicaux pré-entraînés surpasse les autres approches en termes d'exactitude. Plusieurs travaux suggèrent l'utilisation de réseaux de neurones récurrents, en particulier les LSTM (*Long Short-Term Memory*), pour effectuer la détection des émotions dans les tweets (Kabir & Madria, 2021 ; Javed & Muralidhara, 2022). Aslam *et al.* (2022) combinent deux réseaux neuronaux récurrents différents, les LSTM et les

GRU (*Gated Recurrent Units*), et leur modèle atteint une exactitude de 0.91. Les auteurs observent que diminuer la taille du corpus d'entraînement entraîne une baisse de performance du modèle. Par ailleurs, l'utilisation des modèles basés sur les transformers (Vaswani *et al.*, 2017) est largement répandue (Yu *et al.*, 2018; Camacho-Collados *et al.*, 2022), et plus particulièrement le modèle de langage BERT (Devlin *et al.*, 2018). Notamment, Chowdhury & Pal (2023) appliquent des LSTM après avoir ajusté le modèle BERT pour la classification des émotions ; cette approche combinatoire permet d'atteindre un score F1 de 0,71.

Quant à la classification multi-étiquettes des émotions, Kim & Klinger (2018) présentent une analyse des émotions dans les textes littéraires, ayant pour objectif d'identifier les émotions et de les relier aux personnages, à leur origine (causes) et à leurs cibles. Ils construisent pour cela des modèles LSTM bidirectionnels avec une couche de CRF (*Conditional Random Fields*). L'étude montre que l'intégration des CRF améliore l'étiquetage des séquences lorsque les données montrent des dépendances entre les éléments. Enfin, dans sa thèse, Etienne (2023) introduit un modèle capable de réaliser des prédictions simultanées du caractère émotionnel d'un texte, des modes d'expression présents dans les textes, et des types d'émotions véhiculées, tout en effectuant une classification multi-étiquettes des catégories émotionnelles. L'auteure démontre que l'approche prenant en compte plusieurs aspects émotionnels peut améliorer la robustesse du modèle.

Beaucoup plus récemment, dans le domaine des LLMs (*Larges Language Models*) génératifs, Wang *et al.* (2023) développent un test psychométrique nommé SECEU (*Situational Evaluation of Complex Emotional Understanding*) et l'appliquent sur différents LLMs pour évaluer la compréhension émotionnelle des LLMs et des humains. Le LLM génératif le plus efficace pour comprendre les émotions s'est avéré être GPT-4. Liu *et al.* (2024) décrivent quant à eux EmoLLMs, formés par l'affinage de différents LLMs en utilisant un ensemble de données nommé AAID (*Affective Analysis Instruction Dataset*) construit par le SemEval-2018 (Mohammad *et al.*, 2018). Dans les tâches SemEval-2018, ces modèles surpassent les LLMs génératifs open source.

Certaines études se heurtent au manque de données annotées pour détecter les émotions dans les textes. Pour répondre à cette problématique, différentes solutions sont proposées. Par exemple, Baziotis *et al.* (2018) utilisent un corpus de 550 millions de tweets en anglais pour entraîner les plongements lexicaux et un deuxième corpus de 61 854 tweets constitué par SemEval-2017 (Rosenthal *et al.*, 2019) dans le but de réaliser un apprentissage par transfert. Le modèle est ensuite affiné à l'aide des données dans le corpus de SemEval-2018. Cette étude montre que l'utilisation de plusieurs corpus permet de créer un ensemble de données diversifié et représentatif des différentes expressions des émotions. Cependant, cela augmente le coût en temps pour normaliser et harmoniser les données. Guibon *et al.* (2021), pour leur part, explorent l'utilisation de deux corpus distincts pour la classification des émotions dans des dialogues : le premier est utilisé pour la phase d'entraînement, et le second pour évaluer la performance du modèle. L'utilisation de différents corpus permet de tester la généralisation du modèle à diverses sources de données et également d'offrir une évaluation robuste de sa performance.

Compte tenu des travaux existants, plusieurs paramètres doivent être considérés afin de capter les émotions des utilisateurs à propos des services INTERCITÉS : le genre textuel ainsi que les émotions significatives présentes dans notre corpus, et l'architecture du système de prédiction. Ainsi, nous envisageons d'explorer l'utilisation de modèles basés sur les transformers, en particulier l'architecture BERT (Luo & Wang, 2019; Huang *et al.*, 2019; Camacho-Collados *et al.*, 2022). La combinaison de divers modèles nous parait aussi pertinente comme approche, car cela permet d'avoir une architecture plus complexe et de traiter plus finement les données afin d'avoir de meilleures performances, notamment pour la classification des émotions (Aslam *et al.*, 2022; Chowdhury & Pal, 2023).

3 Jeux de données et typologie des émotions

3.1 Présentation du corpus d'étude

Le corpus principal de notre étude englobe l'ensemble des tweets en français diffusés au cours de l'année 2022 et intégrant les termes « intercités » ou « intercité ». Notre étude se concentre exclusivement sur les tweets des utilisateurs relatifs aux lignes INTERCITÉS de jour. L'objectif est de cerner précisément les publications provenant de voyageurs actuels ou potentiels. Dans cette optique, nous avons essayé d'exclure les tweets évoquant l'offre INTERCITÉS de nuit et ceux venant de comptes affiliés à la SNCF ou à des médias, ainsi que ceux se référant à d'autres lignes de transport. Nous avons également éliminé les hyperliens tout en préservant les emojis, ces derniers étant convertis en format textuel cf. (Ex. 1).

(Ex. 1) *1/3 :stop_sign : Service dégradé :stop_sign : Des nouvelles de la ligne POLT ! L'INTERCITÉS de Toulouse est arrivé à Austerlitz avec 2h10 de retard. INTERCITÉS 3665 retardé d'une heure. Pourquoi ? Réponse d'un contrôleur » @Intercites @SNCF*

Après nettoyage, le corpus d'étude comporte 11 025 tweets non-annotés pour un total de 276 716 tokens, hors mentions et hashtags. Il recense 24 718 mentions et 2 272 hashtags. Les tailles minimale, moyenne et maximale des tweets sont respectivement de 1, 24 et 72 tokens.

3.2 Constitution du corpus de référence

Pour analyser les émotions présentes dans notre corpus, une typologie adaptée aux données et à nos objectifs a été choisie. Nous travaillons notamment sur les émotions simples, parce que nous considérons qu'elles répondent aux besoins d'identifier la perception qu'ont les clients et non-clients de l'offre INTERCITÉS, et parce qu'elles facilitent la conception et l'annotation du corpus. Nous nous sommes ainsi inspirées du modèle d'émotions élaboré par Plutchik (1980) et l'avons ajusté. Les émotions sur lesquelles nous travaillons sont les suivantes : joie, tristesse, angoisse, colère, dégoût, neutre. Nous avons décidé de regrouper l'anticipation et la peur du modèle de Plutchik sous l'étiquette «angoisse». En effet, la peur est peu représentée dans les données ferroviaires. L'analyse linguistique de quelques tweets nous a amenées à considérer que l'angoisse est une émotion fréquemment ressentie par les voyageurs en cas de manque d'information et/ou de situation perturbée. L'intensité de cette émotion se situe entre celle de la peur et celle de l'anticipation, avec une nuance plus pondérée. Deux annotateurs spécialistes en linguistique ont annoté un échantillon de 249 tweets en multi-étiquettes selon la typologie mentionnée plus haut, et le taux d'accord global pour l'ensemble des étiquetages, mesuré par le coefficient Kappa de Cohen (Cohen, 1960), est de 0,69. Cette valeur suggère un accord substantiel selon Landis & Koch (1977) mais imparfait entre les annotateurs, indiquant que certaines ambiguïtés subsistent dans le processus d'annotation. Néanmoins, pour les cas les plus difficiles, les annotateurs sont parvenus à un consensus sur l'annotation de l'échantillon de tweets. Cette annotation permet de confirmer l'hypothèse selon laquelle un tweet peut être marqué de plusieurs émotions. Les statistiques de l'annotation sont présentées dans les Tables 1 et 2.

Cet échantillon est considéré comme notre corpus de référence, et sert à valider le comportement du modèle dédié à la détection des émotions. Il est à noter que, dans ce corpus de référence, la répartition des 301 émotions identifiées reste déséquilibrée : les émotions de tristesse (4%) et de dégoût (6%) sont nettement moins représentées que les autres émotions. En outre, bien que les tweets annotés en

uni-étiquette restent largement majoritaires (77%), la part de tweets annotés en multi-étiquettes (23%) reste significative.

Émotions	Nombre	Pourcentage
Colère	82	27%
Joie	74	25%
Neutre	58	19%
Angoisse	57	19%
Tristesse	19	6%
Dégoût	11	4%
Total	301	100%

TABLE 1 – Répartition des émotions dans le corpus de référence annoté en uni-étiquettes et multi-étiquettes

Type d'annotation	Nombre de tweets	Pourcentage
Uni-étiquette	192	77%
Multi-étiquette	57	23%
Total	249	100%

TABLE 2 – Répartition des étiquettes uniques ou multiples dans le corpus de référence

3.3 Constitution du corpus d'entraînement

Tout entraînement de modèle pour des tâches de traitement de données langagières nécessite un nombre important de données d'exemples. Comme mentionné précédemment, le corpus INTERCITÉS n'est pas annoté en émotions et nous ne disposons pas de données équivalentes annotées. Pour franchir ce premier obstacle, notre stratégie consiste à entraîner le modèle sur un corpus annoté pré-existant, et à appliquer ensuite les connaissances acquises par le modèle pour annoter notre corpus d'étude en émotions. En effet, selon l'état de l'art, la fusion de données provenant de différents corpus est une solution viable en cas de manque de données (Baziotis *et al.*, 2018; Guibon *et al.*, 2021). Après l'examen de plusieurs corpus, nous décidons d'utiliser le corpus DEFT 2015 [1] (Hamon *et al.*, 2015), qui regroupe des tweets axés sur le thème du changement climatique. Dans ce corpus, chaque tweet est annoté pour refléter une émotion, une opinion ou un sentiment, en utilisant un système de 19 étiquettes définies par Fraisse & Paroubek (2014) [2]. Il existe une disparité notable entre la typologie des étiquettes d'annotation utilisées dans ce corpus et celle que nous avons établie pour notre projet. En conséquence, nous avons converti les étiquettes pour aligner les annotations du corpus DEFT avec notre typologie (cf. section 3.2). Ce processus a donné naissance à notre corpus d'entraînement, que nous appelons CoarseDEFT.

La distribution des émotions dans le corpus d'entraînement est déséquilibrée (Table 3) et diffère de celle constatée dans le corpus de référence. L'émotion de joie est dominante dans le corpus d'entraînement, tandis que les émotions de tristesse et de dégoût sont sous-représentées. Les sous-représentations de certaines étiquettes font que le modèle a moins d'exemples pour apprendre à les identifier correctement.

3.4 Synthèse des données utilisées

La première phase de notre travail nous a amenées à constituer quatre corpus (cf. Table 4), utilisés à différentes phases dans les expériences que nous présentons dans la section suivante. Nous avons

1. https ://deft.lisn.upsaclay.fr/2015/
2. https ://deft.lisn.upsaclay.fr/2015/guideAnnotation.fr.php ?lang=fr

Étiquettes DEFT 2015	Étiquettes CoarseDEFT	Proportions
Amour, Apaisement, Plaisir, Satisfaction, Valorisation	Joie	52,1%
Colère, Ennui, Insatisfaction, Dévalorisation, Mépris	Colère	25,7%
Accord, Désaccord	Neutre	11,7%
Peur	Angoisse	8,5%
Tristesse, Déplaisir	Tristesse	1,6%
Dérangement	Dégoût	0,4%

TABLE 3 – Adaptation de la typologie des émotions du corpus DEFT 2015 vers la nouvelle typologie utilisée dans le corpus CoarseDEFT

divisé le corpus CoarseDEFT en différentes parties distinctes. L'entraînement initial de notre modèle est réalisé à partir d'une partie de corpus CoarseDEFT (70%). L'évaluation de sa performance, tout au long des expériences, s'effectue sur deux corpus : un échantillon du corpus CoarseDEFT (15%) et notre corpus de référence INTERCITÉS (cf. section 3.2). La seconde phase de notre travail a consisté à enrichir notre corpus d'entraînement pour améliorer les performances de notre modèle (cf. section 4.2). Pour cela, nous avons construit de manière itérative un autre corpus de 863 486 tweets, extraits de Twitter/X avec le mot-clé «SNCF».

Corpus	Nombre de tweets	Sources de tweets
Train (avant l'enrichissement)	3 002	CoarseDEFT
Val (avant l'enrichissement)	659	CoarseDEFT
Test (avant l'enrichissement)	659	CoarseDEFT
Corpus de référence	249	Corpus INTERCITÉS
Corpus d'enrichissement 1	10 776	Corpus INTERCITÉS
Corpus d'enrichissement 2	863 486	Corpus SNCF

TABLE 4 – Corpus utilisés pour l'apprentissage

4 Méthodologie

La démarche mise en place vise à spécialiser notre modèle de prédiction en fonction de notre corpus d'étude, permettant ainsi une classification multi-étiquettes efficace et précise sur les données ferroviaires. Notre objectif est donc d'entraîner un classifieur multi-étiquettes sur le corpus d'entraînement (CoarseDEFT) annoté selon différentes émotions, puis de le spécialiser sur nos données. Pour cela, nous avons mis en place une méthode itérative, qui enrichit progressivement le corpus d'entraînement avec davantage de données (cf. schéma 1). Le modèle s'entraîne sur le corpus d'entraînement et après chaque phase d'apprentissage, nous évaluons sa performance sur notre corpus de référence (cf. section 3.2). Plus précisément, nous effectuons une analyse manuelle des prédictions pour observer le comportement du modèle. Si l'analyse indique une pertinence insuffisante, nous intégrons de nouvelles données issues de nos corpus d'enrichissement (cf.Table 4) au corpus d'entraînement en fonction de cette analyse. Les itérations se poursuivent ensuite jusqu'à atteindre une performance satisfaisante.

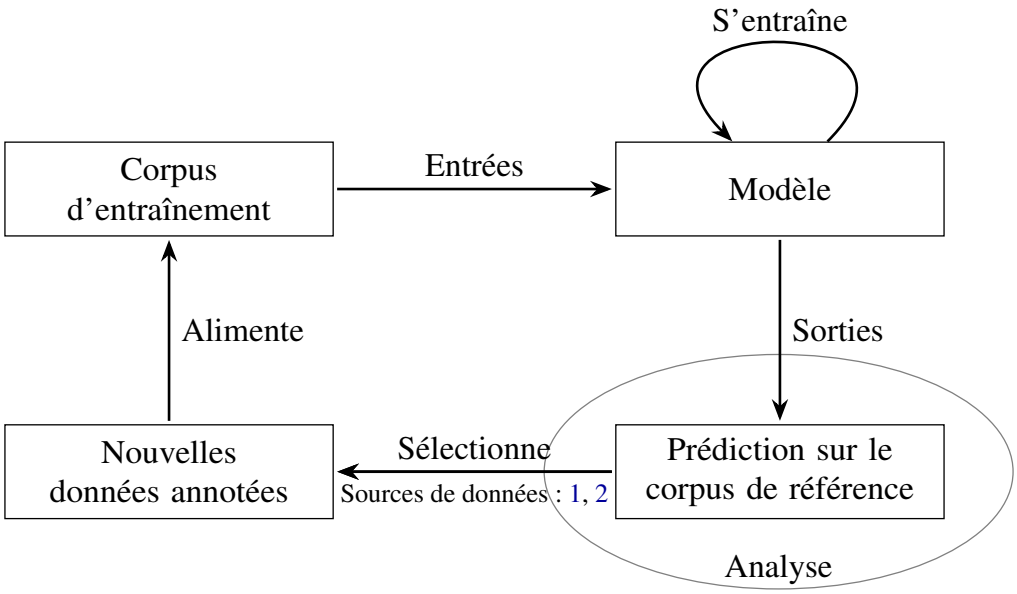

FIGURE 1 – Chaîne de traitement de la spécialisation du modèle

4.1 Configuration du modèle dédié à la classification des émotions en multi-étiquettes

Notre modèle repose sur l'affinage des deux dernières couches du modèle de langue CamemBERT (Martin *et al.*, 2020), architecture basée sur les transformers largement utilisés pour les tâches de classification de textes en français (Lincker *et al.*, 2023; Etienne, 2023). Cette représentation est enrichie par l'ajout de six couches d'encodeurs transformers, chacune avec 8 têtes d'attention. Notre structure des transformers s'inspire de l'approche explorée par Blivet *et al.* (2023) dans le cadre de leur participation au défi DEFT 2023 [3] qui traite de la classification multi-étiquettes.

Dans un modèle de classification multi-étiquettes, chaque label est traité comme une entité distincte. Si l'on considère un ensemble de N labels, le modèle génère N scores de probabilité $p_1, p_2, \ldots, p_N$ indépendants, chacun correspondant à une étiquette spécifique. Ces scores sont obtenus par la fonction d'activation *sigmoid* appliquée à la sortie du modèle pour chaque étiquette. Un seuil de décision θ est appliqué à chaque probabilité prédite pour déterminer la présence ou l'absence d'une étiquette. Si la probabilité prédite pour une étiquette donnée i, p_i est supérieure ou égale au seuil θ , alors l'étiquette i est assignée à l'instance ; sinon, elle est rejetée. La valeur de θ est généralement fixée à 0,5 dans de nombreux cas par défaut, mais cette valeur peut être ajustée pour répondre à des critères de performance spécifiques. Le choix de θ implique un équilibre entre la sensibilité (taux de vrais positifs) et la spécificité (taux de vrais négatifs) du modèle. Dans le cadre de notre étude, nous définissons le seuil à 0,35 au début de nos expériences, puisqu'initialement le modèle n'est pas assez sensible pour générer des probabilités plus élevées permettant des prédictions fiables. Cependant, à mesure que la performance du modèle s'améliore, nous augmentons progressivement ce seuil jusqu'à atteindre 0,5, comme le montre la Table 6.

3. https ://deft2023.univ-avignon.fr/

4.2 Spécialisation du modèle par enrichissement du corpus d'entraînement

Les prédictions sur le corpus de référence sont analysées pour évaluer la sensibilité du modèle. Les premières analyses indiquent un faible rappel, signifiant que le modèle ne généralise pas correctement sur les données INTERCITÉS. Bien qu'il identifie mieux les émotions fréquentes (joie, colère, neutre), le modèle génère de nombreux faux positifs, suggérant une insuffisance de données pour les émotions sous-représentées comme la tristesse et le dégoût.

Pour augmenter la performance du modèle sur le corpus d'étude, nous avons adopté les stratégies suivantes pour enrichir le corpus d'entraînement :

1. Adapter le contenu du corpus d'entraînement pour le rendre plus similaire à celui du corpus d'étude, en utilisant des données extraites du corpus INTERCITÉS. Cela permet de mieux aligner notre modèle aux spécificités des tweets d'INTERCITÉS.
2. Améliorer l'équilibre des émotions représentées dans le corpus d'entraînement, en ciblant spécifiquement les émotions sous-représentées telles que la tristesse et le dégoût. Pour cela, nous avons extrait de nouveaux tweets en utilisant le mot-clé « SNCF » et sélectionné ceux qui contiennent des termes associés à la tristesse et au dégoût, identifiés grâce au dictionnaire EmoLex (Mohammad & Turney, 2013).

L'enrichissement s'est fait selon plusieurs itérations dont le détail est présenté avec les informations précises concernant la répartition des émotions se trouve dans l'annexe A.

5 Évaluation et résultats

Afin de valider la performance des modèles, nous prenons en compte deux mesures d'évaluation : le F1-Micro, le F1-Macro et l'exactitude, fréquemment utilisées pour les tâches de classification.

Méthodes	**Exactitude**	**F1-Micro**	**F1-Macro**
SVM	**0,77**	0,18	0,07
Random Forest	**0,77**	0,20	0,10
Régression logistique	0.74	0,25	0,17
Naive Bayes	0.65	**0,36**	**0,29**
Arbre de décision	0,71	0.28	0,20

TABLE 5 – Évaluation de méthodes de référence pour la définition d'une *baseline*

Avant de mettre en oeuvre notre méthodologie, nous construisons tout d'abord une *baseline*. Pour cela, nous évaluons les performances des méthodes classiques de classification automatique sur notre corpus (cf. Table 5). Ces approches manifestent des difficultés à équilibrer l'exactitude avec les scores F1-Micro et F1-Macro. Naive Bayes se distingue par ses performances supérieures en F1-Micro et en F1-Macro. L'écart global entre F1-Micro et F1-Macro indique que les classes majoritaires peuvent influencer le résultat du F1-Micro. Les scores obtenus des approches soulignent la complexité de la tâche de classification multi-étiquettes et confirment qu'il existe une marge significative pour l'amélioration des modèles.

Ensuite, nous avons réalisé quatre cycles d'enrichissement de notre corpus d'entraînement, suivis chacun par une phase de ré-entraînement du modèle. Après chaque cycle, nous avons évalué la performance du modèle à l'aide des métriques sélectionnées. L'évaluation s'effectue d'abord sur le corpus de référence composé de 249 tweets (cf. Figure 1). Les résultats obtenus après chaque session d'entraînement sont présentés dans la Table 6.

Corpus	**Taille du corpus d'entraînement**	**Exactitude**	**F1-Micro**	**F1-Macro**	**Seuil**
CoarseDEFT	3661 tweets	0,77	0,39	0,22	0,35
CoarseDEFT-V2	4361 tweets	0,814	0,51	0,36	0,45
CoarseDEFT-V3	5561 tweets	0,816	0,50	0,43	0,5
CoarseDEFT-V4	6201 tweets	**0.826**	**0.55**	**0,44**	0,5

TABLE 6 – Évolution des performances du modèle au cours des différentes itérations de l'entraînement

Les résultats de l'évaluation montrent d'abord que notre approche permet d'atteindre la *baseline* définie dès la première itération en exactitude et en F1 Micro et de la dépasser ensuite au fur et à mesure que le corpus est alimenté par les données provenant des deux méthodes que nous avons mentionnées dans la partie précédente. Nous observons des améliorations significatives de F1-Micro lors de l'ajout des données aux corpus CoarseDEFT-V2 et CoarseDEFT-V4, lesquelles concernent des tweets relatifs à INTERCITÉS. Par ailleurs, l'intégration de données correspondant aux émotions sous-représentées dans le corpus CoarseDEFT-V3 contribue à une amélioration du score F1-Macro. Les scores F1 pour chaque étiquette d'émotion, obtenus lors des différentes itérations, sont présentés dans l'annexe B. L'augmentation du score d'exactitude prouve d'une part l'amélioration de la spécialisation du modèle et confirme que les nouvelles données ajoutées n'introduisent pas de bruit dans les prédictions. L'augmentation des scores F1 Micro et Macro semble valider nos hypothèses : l'adaptation du corpus aux données ferroviaires et l'équilibrage de la distribution des émotions améliorent la précision du modèle.

Compte tenu de la complexité de l'annotation des émotions, confirmée par le taux d'accord inter-annotateurs obtenu lors de l'annotation des émotions dans le corpus de référence (cf. section 3.2), il est pertinent de se demander si les les faibles scores F1 Micro et Macro observés dans la Table 6 peuvent être attribué à des divergences dans l'interprétation de la typologie des émotions parmi les annotateurs.

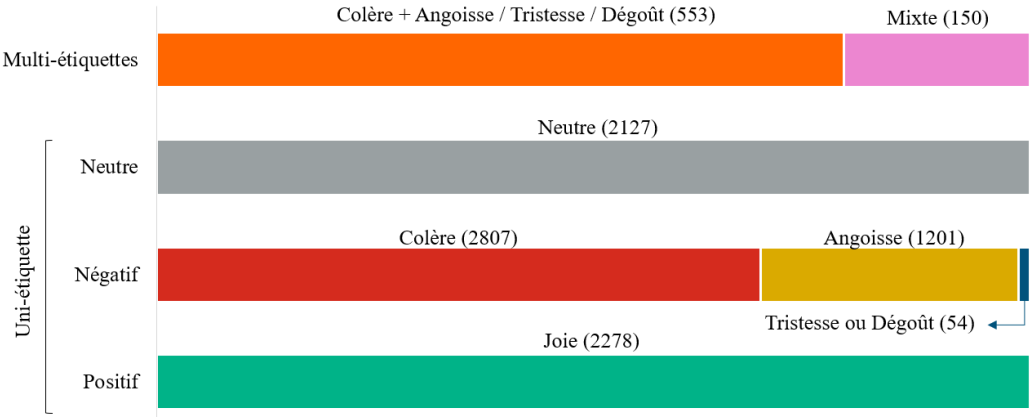

FIGURE 2 – Répartition des émotions dans le corpus INTERCITÉS

Bien que le modèle nécessite des améliorations, il affiche déjà une performance satisfaisante, en particulier en termes de spécificité. Appliqué à l'ensemble des tweets de notre corpus d'étude, il nous permet d'extraire des statistiques qui donnent une première indication de la répartition des émotions. D'après la Figure 2, les émotions les plus exprimées sont la colère, la joie, le neutre et l'angoisse. Les émotions négatives dominent dans les tweets. La sur-représentation des émotions négatives colère et angoisse peut certainement être relativisée et attribuée aux particularités de la plate-forme Twitter/X, sur laquelle les utilisateurs sont plus enclins à partager des commentaires et des messages négatifs.

6 Conclusion et discussion

Pour répondre aux enjeux commerciaux et industriels de notre projet, nous implémentons une chaîne de traitement qui entraîne un modèle de classification multi-étiquettes capable de détecter les émotions exprimées dans les tweets. Une partie importante de cette chaîne de traitement implique la constitution d'un corpus d'entraînement, annoté selon une typologie adaptée à notre étude. En l'occurrence, le corpus DEFT 2015 est repris, adapté à une nouvelle typologie d'émotions et enrichi en tweets émotionnellement marqués et traitant du domaine ferroviaire. Cette démarche permet au modèle de transférer ses acquis à notre corpus d'étude, lequel n'est pas annoté. Notre modèle, soit une combinaison du modèle CamemBERT et de transformers, dépasse d'emblée la baseline établie, et atteint un F1-Micro score de 0,55, un F1-Macro score de 0,44 et une exactitude de 0,826. Ce modèle se révèle être précis et efficace pour la classification des émotions dans les situations où une seule étiquette doit être prédite.

Dans le cadre de cette étude, nous avons fait face à plusieurs défis. Premièrement, l'annotation des émotions dans les tweets est une tâche subjective, qui complique la mise en place de critères d'annotation uniformes. Deuxièmement, notre corpus d'entraînement montre une répartition inégale entre les tweets uni-étiquettes, qui dominent, et les tweets multi-étiquettes. De surcroît, la distribution des différentes émotions reste déséquilibrée au sein du corpus, même après l'ajout de nouvelles données. Ces déséquilibres exercent une influence négative sur les performances du modèle. Cependant, il est essentiel de déterminer si ces inégalités sont propres et inhérentes à notre corpus INTERCITÉS. Au lieu d'essayer d'équilibrer artificiellement le corpus d'entraînement, il serait plus avisé d'adapter le modèle pour qu'il s'accommode de cette distribution inégale des étiquettes. Enfin, nous devons reconsidérer la nature des données utilisées lors de l'entraînement du modèle. Les données contenues dans le corpus d'entraînement (CoarseDEFT et ses versions successives), qui servent à l'apprentissage du modèle, se distinguent de celles du corpus de référence utilisé pour les prédictions et les évaluations. Le modèle présente des limitations dans le transfert des connaissances du corpus d'entraînement vers le corpus INTERCITÉS.

Malgré une sensibilité parfois limitée pour identifier l'intégralité des vrais positifs, les résultats permettent de dégager des tendances représentatives de la perception qu'ont les clients et les non-clients de l'offre INTERCITÉS et répondent aux besoins spécifiques de la la Direction Marketing. Pour aller au-delà de la classification multi-étiquettes des émotions, l'Analyse de Sentiments Basée sur les Aspects (ABSA) (Pontiki *et al.*, 2016) est une piste de recherche envisagée afin de relier les émotions aux thématiques évoquées dans les tweets. Cela permettra d'analyser plus précisément la manière dont sont perçus les différents aspects spécifiques de l'offre INTERCITÉS.

Références

ASLAM N., RUSTAM F., LEE E., WASHINGTON P. B. & ASHRAF I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model. *Ieee Access*, **10**, 39313–39324. DOI : 10.1109/access.2022.3165621.

BAZIOTIS C., ATHANASIOU N., CHRONOPOULOU A., KOLOVOU A., PARASKEVOPOULOS G., ELLINAS N., NARAYANAN S. & POTAMIANOS A. (2018). Ntua-slp at semeval-2018 task 1 : Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv :1804.06658*. DOI : 10.18653/v1/s18-1037.

BLIVET A., DEGRUTÈRE S., GENDRON B., RENAULT A., SIOUFFI C., GAUDRAY-BOUJU V., CERISARA C., FLAMEIN H., GUIBON G., LABEAU M. *et al.* (2023). Participation de l'équipe ttgv à deft 2023 : Réponse automatique à des qcm issus d'examens en pharmacie. *Actes de CORIA-TALN 2023. Actes du Défi Fouille de Textes@ TALN2023*, p. 23–38.

CAMACHO-COLLADOS J., REZAEE K., RIAHI T., USHIO A., LOUREIRO D., ANTYPAS D., BOISSON J., ESPINOSA-ANKE L., LIU F., MARTÍNEZ-CÁMARA E. *et al.* (2022). Tweetnlp : Cutting-edge natural language processing for social media. *arXiv preprint arXiv :2206.14774*. DOI : 10.18653/v1/2022.emnlp-demos.5.

CHOWDHURY M. S. M. & PAL B. (2023). Bert-based emotion classification approach with analysis of covid-19 pandemic tweets. In *Applied Informatics for Industry 4.0*, p. 109–121. Chapman and Hall/CRC. DOI : 10.1201/9781003256069-10.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46. DOI : 10.1177/001316446002000104.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv (Cornell University)*.

EKMAN P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**(3-4), 169–200. DOI : 10.1080/02699939208411068.

ETIENNE A. (2023). *Analyse automatique des émotions dans les textes : contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants*. Thèse de doctorat, Université de Nanterre-Paris X.

FRAISSE A. & PAROUBEK P. (2014). Toward a unifying model for opinion, sentiment and emotion information extraction. In *The 9th International Conference on Language Resources and Evaluation*, p. 3881–3886 : European Language Resources Association (ELRA).

GUIBON G., LABEAU M., FLAMEIN H., LEFEUVRE L. & CLAVEL C. (2021). Meta-learning for classifying previously unseen data source into previously unseen emotional categories. In *1st Workshop on Meta Learning and Its Applications to Natural Language Processing, ACL 2021*. DOI : 10.18653/v1/2021.metanlp-1.9.

HAMON T., FRAISSE A., PAROUBEK P., ZWEIGENBAUM P. & GROUIN C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (deft). In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*.

HUANG C., TRABELSI A. & ZAÏANE O. R. (2019). Ana at semeval-2019 task 3 : Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv :1904.00132*.

JAVED N. & MURALIDHARA B. (2022). Emotions during covid-19 : Lstm models for emotion detection in tweets. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications : ICMISC 2021*, p. 133–148 : Springer. DOI : 10.1007/978-981-16-6407-6_13.

KABIR M. Y. & MADRIA S. (2021). Emocov : Machine learning for emotion detection, analysis and visualization using covid-19 tweets. *Online Social Networks and Media*, **23**, 100135–100147. DOI : 10.1016/j.osnem.2021.100135.

KIM E. & KLINGER R. (2018). Who feels what and why ? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1345–1359 : Association for Computational Linguistics.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

LINCKER E., GUINAUDEAU C., PONS O., DUPIRE J., BARBET I., HUDELOT C., MOUSSEAU V. & HURON C. (2023). Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires. In *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 121–130 : ATALA.

LIU Z., YANG K., ZHANG T., XIE Q., YU Z. & ANANIADOU S. (2024). Emollms : A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv :2401.08508*. DOI : 10.48550/arxiv.2401.08508.

LORA S. K., SAKIB N., ANTORA S. A. & JAHAN N. (2020). A comparative study to detect emotions from tweets analyzing machine learning and deep learning techniques. volume 12, p. 6–12.

LUO L. & WANG Y. (2019). Emotionx-hsu : Adopting pre-trained bert for emotion classification. *arXiv preprint arXiv :1907.09669*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*. DOI : 10.18653/v1/2020.acl-main.645.

MOHAMMAD S., BRAVO-MARQUEZ F., SALAMEH M. & KIRITCHENKO S. (2018). Semeval-2018 task 1 : Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, p. 1–17. DOI : 10.18653/v1/s18-1001.

MOHAMMAD S. M. & TURNEY P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, **29**(3), 436–465.

PAVEAU M.-A. (2017). *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*. Hermann.

PLUTCHIK R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, p. 3–33. Elsevier. DOI : 10.1016/b978-0-12-558701-3.50007-7.

PONTIKI M., GALANIS D., PAPAGEORGIOU H., ANDROUTSOPOULOS I., MANANDHAR S., AL-SMADI M., AL-AYYOUB M., ZHAO Y., QIN B., DE CLERCQ O. *et al.* (2016). Semeval-2016 task 5 : Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, p. 19–30 : Association for Computational Linguistics. DOI : 10.18653/v1/s16-1002.

RABEYA T., FERDOUS S., ALI H. S. & CHAKRABORTY N. R. (2017). A survey on emotion detection : A lexicon based backtracking approach for detecting emotion from bengali text. In *2017 20th international conference of computer and information technology (ICCIT)*, p. 1–7 : IEEE. DOI : 10.1109/iccitechn.2017.8281855.

ROSENTHAL S., FARRA N. & NAKOV P. (2019). Semeval-2017 task 4 : Sentiment analysis in twitter. *arXiv preprint arXiv :1912.00741*. DOI : 10.18653/v1/s17-2088.

STRAPPARAVA C., VALITUTTI A. *et al.* (2004). Wordnet affect : an affective extension of wordnet. In *Lrec*, volume 4, p. 1083–1086 : Lisbon, Portugal.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention Is All You Need. arXiv :1706.03762 [cs].

WANG X., LI X., YIN Z., WU Y. & LIU J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*. DOI : 10.1177/18344909231213958.

YU J., MARUJO L., JIANG J., KARUTURI P. & BRENDEL W. (2018). Improving multi-label emotion classification via sentiment classification with dual attention transfer network. DOI : 10.18653/v1/d18-1137.

A Enrichissement du corpus d'entraînement

Corpus	Données	uni-étiquette	multi-étiquttes	Neutre	Joie	Tristesse	Colère	Angoisse	Dégoût
CoarseDEFT	**Train** : 3002 tweets	3002	0	352	1564	47	774	255	10
	Val : 659 tweets	659	0	73	347	9	170	57	3
CoarseDEFT-V2	**Train** : 3002 + 700 tweets d'Intercités	3484	218	484	1685	64	1042	404	21
	Val : 659 tweets	659	0	73	347	9	170	57	3
CoarseDEFT-V3	**Train** : 3002 + 700 tweets d'Intercité + 500 tweets «tristesse» + 500 tweets «dégoût»	3977	725	516	1781	190	1127	344	115
	Val : 659 + 100 tweets «tristesse» + 100 tweets «dégoût»	749	110	93	349	19	214	66	90
CoarseDEFT-V4	**Train** : 3002 + 1050 tweets d'Intercité + 800 tweets «tristesse» + 500 tweets «dégoût»	4553	799	599	1786	343	1275	435	115
	Val : 659 + 100 tweets «trsitesse» + 100 tweets «dégoût»	749	110	93	349	19	214	66	90

TABLE 7 – Évolution des données du modèle au cours des différentes itérations de l'entraînement

B Analyse des résultats par émotion

Corpus	F1 score par émotion						F1-Micro	F1-Macro
	Neutre	Joie	Tristesse	Colère	Angoisse	Dégoût		
CoarseDEFT	0.31	0.47	0	0.55	0	0	0.39	0.22
CoarseDEFT-V2	0.43	0.65	0	0.68	0.40	0	0.51	0.36
CoarseDEFT-V3	0.5	0.54	0	0.62	0.45	0.50	0.50	0.43
CoarseDEFT-V4	0.52	0.6	0	0.6	0.61	0.27	0.55	0.44

TABLE 8 – Évaluation de la classification par étiquette