ADAPTATION DE DOMAINE POUR L'ANALYSE DE SENTIMENTS : DES REVUES FORMELLES AUX TWEETS INFORMELS

Julien F. Issert[1], *Karl E. Gérard*[1] *et Théo B. Demany*[1]

[1]Université du Mans, Le Mans, France
{julien.issert, karl.gerard, theo.demany}.etu@univ-lemans.fr

ABSTRACT

Ce projet explore les défis de l'adaptation de domaine dans l'analyse de sentiments, en effectuant une transition de critiques de produits structurées (Amazon) vers du contenu informel issu d'un réseau social (Twitter). Nous définissons un système de référence utilisant une vectorisation TF-IDF et une régression logistique entraînée sur le jeu de données *Amazon Fine Food Reviews* (Domaine Source). La robustesse du modèle est évaluée sur le jeu de données *tweet_emotions* (Domaine Cible) pour diagnostiquer l'impact du « Domain Shift ». Nos résultats montrent une chute significative de la performance, l'exactitude passant de 78,74 % à 44,98 % en mode zero-shot. Ce rapport détaille le pipeline de prétraitement, la stratégie de regroupement des classes (mapping) et l'établissement d'une référence par échantillonnage aléatoire sur plusieurs budgets (1 % à 100 %) avec intervalles de confiance, servant de base aux futures stratégies d'apprentissage actif.

Index Terms— Analyse de sentiments, Adaptation de domaine, Domain Shift, NLP, Apprentissage Actif.

1. INTRODUCTION

L'objectif de cette étude est de concevoir un système de classification capable de s'adapter à des changements de distribution de données, en passant de critiques de produits structurées (Amazon) à des messages informels (Twitter). Alors que les modèles supervisés excellent lorsque les données d'entraînement et de test sont similaires, ils échouent souvent face à de nouveaux contextes, un phénomène appelé « Domain Shift ». Cette première phase établit les performances de base et quantifie la dégradation causée par l'écart entre les domaines avant l'implémentation de l'apprentissage actif.

2. TRAVAUX ANTÉRIEURS ET ÉTAT DE L'ART

L'analyse de sentiments cross-domaine est un défi classique. Les travaux de Wu et al. [4] soulignent que le décalage lexical entre domaines est le principal frein à la généralisation. Pour réduire le coût d'annotation induit par ce décalage, Settles [5] définit l'apprentissage actif comme une solution clé, notamment via des stratégies d'incertitude ou de diversité. Des études comme celles de Li et al. [6] démontrent que les stratégies mixtes sont souvent supérieures pour l'adaptation de domaine.

3. PROTOCOLE EXPÉRIMENTAL ET DATASETS

Conformément aux directives de projet, deux sources de données distinctes simulent l'adaptation de domaine

- **Domaine Source (Amazon) :** Le jeu de données *Amazon Fine Food Reviews* [7], comprenant des critiques longues et formelles.
- **Domaine Cible (Twitter) :** Le jeu de données *Emotion Detection from Text* [8], caractérisé par des messages courts, de l'argot et des emojis.

3.1. Regroupement des Labels (Mapping)

Le dataset Twitter comporte 13 émotions fines. Pour assurer la compatibilité avec le modèle Amazon, nous avons opéré un regroupement vers trois classes cibles :

- **POSITIVE :** *love*, *happiness*, *relief*, *enthusiasm*.
- **NEGATIVE :** *anger*, *sadness*, *worry*, *hate*.
- **NEUTRAL :** *neutral*.
- **Exclusions :** Les catégories ambiguës comme *surprise* ont été exclues pour garantir la cohérence sémantique lors du test.

3.2. Pipeline de Prétraitement (Preprocessing)

Un nettoyage rigoureux est appliqué aux deux domaines pour assurer l'uniformité:

1. **Regex :** Suppression des URLs, mentions (@) et hashtags (#). Les points d'exclamation sont conservés.
2. **Normalisation :** Conversion systématique en minuscules.
3. **Lemmatisation :** Utilisation du WordNetLemmatizer (NLTK) pour regrouper les variantes morphologiques.
4. **Filtrage :** Élimination des mots de moins de deux lettres.

4. REPRODUCTIBILITÉ ET PARTITIONNEMENT

4.1. Partitionnement des données

Le jeu de données est divisé de manière aléatoire et stratifiée pour conserver la distribution des classes. Les volumes sont répartis comme suit :

Tableau 1. Répartition des sous-ensembles de données.

eSous -ensemble	Domaine	Rôle Expérimentale	Taille
$D_{S,Train}$	Amazon	Entraînement de la Baseline	32 000
$D_{S,Test}$	Amazon	Validation Interne	8 000
$D_{T,Pool}$	Twitter	Pool d'Apprentissage Actif	~30 250
$D_{T,Test}$	Twitter	Evaluation Cross-Domain Finale	7 560

Tableau 2. Représentation des classes dans les sous-ensembles de données.

eSous -ensemble	POS(%)	NEG(%)	NEU(%)
$D_{S,Train}$	78,21	14,25	7,54
$D_{S,Test}$	78,21	14,25	7,54
$D_{T,Pool}$	34,68	42,48	22,84
$D_{T,Test}$	34,67	42,48	22,85

4.2. Protocole de sélection et de mise à jour

Le protocole expérimental suit une approche stricte pour isoler l'effet du changement de domaine :

- **Domaine source :** Les données Amazon sont écartées lors de l'adaptation sur Twitter afin d'évaluer la capacité de transfert pur vers le domaine cible .
- **Sélection :** Nous adoptons une stratégie **non-incrémentale** ; chaque budget (ratio) fait l'objet d'un nouvel échantillonnage indépendant dans le pool $D_{T,Pool}$.
- **Mise à jour :** Pour garantir l'indépendance des mesures, le modèle subit un réentraînement complet (réinitialisation des poids) à chaque étape, excluant tout effet de mémoire des itérations précédentes .

5. ARCHITECTURE DU MODÈLE BASELINE

Le système de référence est construit sur un pipeline Scikit-Learn optimisé pour le traitement de données textuelles à haute dimensionnalité:

- **Vectorisation :** Nous utilisons une méthode TF-IDF avec des bi-grammes $(ngram\_range = (1, 2))$. Le vocabulaire est contraint à 15 000 variables afin de garantir une efficacité computationnelle maximale.
- **Classification :** L'algorithme retenu est une Régression Logistique. Le paramètre *class_weight='balanced'* est appliqué pour ajuster automatiquement les poids et compenser le déséquilibre entre les classes positives et neutres.
- **Complexité et Performance technique :** * Paramètres : Le modèle totalise 45 003 paramètres (45 000 coefficients et 3 ordonnées à l'origine/intercepts).
    - **Temps de calcul :** L'entraînement complet de la baseline sur les 32 000 échantillons Amazon s'effectue en 13,52 secondes.
    - **Inférence :** Le temps de prédiction sur le jeu de test (8 000 échantillons) est de 1,01 seconde.

6. ANALYSE DU DOMAIN SHIFT

6.1. Performances Internes (Amazon → Amazon)

Le modèle affiche une **Accuracy de 78,74 %** (voir figure 1.A en annexe). La classe positive est excellente (F1: 0,89), tandis que la classe neutre reste difficile à isoler (F1: 0,32).

6.2. Évaluation Cross-Domain (Amazon → Twitter)

En mode Zero-Shot, l'exactitude chute brutalement à **44,98 %** selon l'analyse cross-domain (voir figure 1.B en annexe).

- **Diagnostic :** Nous identifions un décalage lexical majeur et une différence structurelle (messages courts vs longs).
- **Biais :** Le modèle sur-prédit le positif (Recall: 0,72) car il ne reconnaît pas les marqueurs négatifs spécifiques aux réseaux sociaux.

7. STRATÉGIE ALÉATOIRE ET FINE-TUNING

Nous avons évalué l'impact de l'injection progressive de données du domaine cible (Twitter) sur les performances du modèle. Pour chaque palier budgétaire, l'expérience est répétée sur cinq itérations indépendantes afin de rapporter l'exactitude moyenne et de mesurer la variabilité des résultats via l'intervalle de confiance (CI).

Tableau 3 : Performance de la sélection aléatoire par budget.

Budget (%)	Accuracy Moyenne (%)	Écart-Type (±)	Gain vs Baseline (pts)
0 % (Zero-shot)	44,9 8 %	-	-
1 %	46,57 %	1,12	+ 1.59
5 %	50,72 %	0,85	+ 5,74

10 %	52.63 %	0,45	+	7,65
20%	53,38 %	0,32	+	8.40
30%	54,13 %	0,28	+	9.15
50%	55,21 %	0,21	+	10.23
75%	55,76 %	0,18	+	10.78
100%	56,02 %	0,12	+	11.04

7.1. Analyse de la courbe d'apprentissage

La figure 2 (voir Annexe) illustre l'évolution de la performance sur 8 paliers budgétaires. L'écart-type, représenté par la zone d'ombre, démontre une stabilité croissante du modèle à mesure que le budget augmente. On observe une progression logarithmique : un gain majeur (+7,65 pts) est réalisé avec seulement 10 % des données, suivi d'une phase de saturation au-delà de 50 %

8. CONCLUSION ET PERSPECTIVES

Cette première phase démontre qu'un modèle entraîné sur des données formelles ne généralise pas sur des données sociales sans adaptation. Bien que l'échantillonnage aléatoire améliore les résultats, il atteint un plafond de verre à **56 %**. La suite consistera à implémenter des stratégies d'**Apprentissage Actif** (incertitude et diversité) pour sélectionner les échantillons les plus informatifs et optimiser le coût d'annotation

9. RÉFÉRENCES

[1] Directives du projet, "Active and Interactive Learning Project Guideline," 2024.

[2] Grille d'évaluation, "Note 2024 - Sheet1," 2024.

[3] Modèle ICASSP 2024, "Author Guidelines for Proceedings Manuscripts," 2024.

[4] F. Wu et al., "Active Sentiment Domain Adaptation," *Proc. of the 55th ACL*, 2017.

[5] B. Settles, "Active Learning Literature Survey," *Technical Report*, 2009.

[6] S. Li et al., "Active Learning for Cross-domain Sentiment Classification," *Proc. of the 23rd IJCAI*, 2013.

[7] Kaggle, "Amazon Fine Food Reviews Dataset," 2017.

[8] Kaggle, "Emotion Detection from Text Dataset," 2020.

10. ANNEXE

Fig. 1.a.: Performance de référence du modèle de classification de sentiments sur les données Amazon.

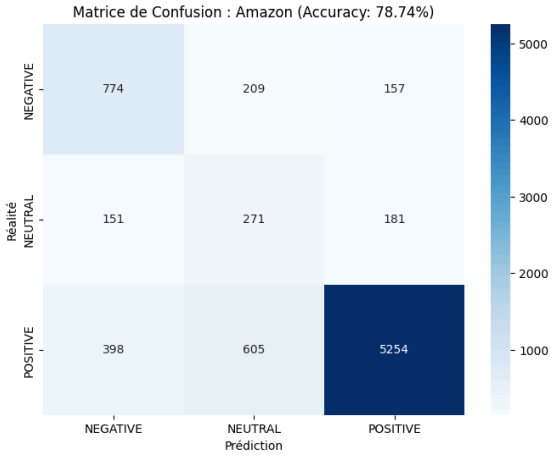

Fig. 1.b.: Généralisation du modèle Amazon sur des données hors-distribution sur les données Twitter.

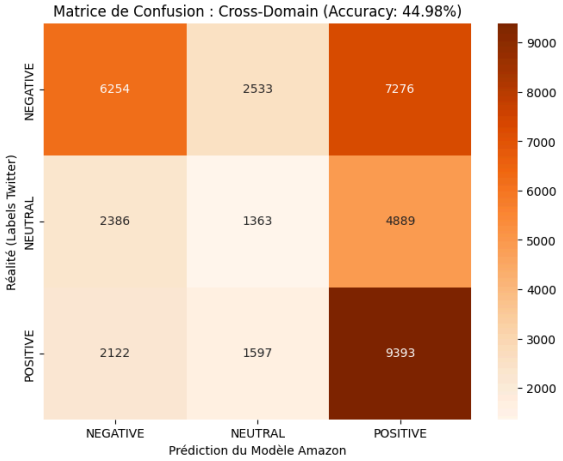

Fig. 2.: Courbe de performance du fine-tuning aléatoire (moyenne et variabilité) par rapport à la baseline Zero-Shot.