

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

Karl E. Gérard¹, Julien F. Issert¹ et Théo B. Demany¹

Code Source: <https://github.com/KarlGerard/Active-Learning-Sentiment-Adaptation>

¹Université 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:

- **Regex** : Suppression des URLs, mentions (@) et hashtags (#). Les points d'exclamation sont conservés.
- **Normalisation** : Conversion systématique en minuscules.
- **Lemmatisation** : Utilisation du WordNetLemmatizer (NLTK) pour regrouper les variantes morphologiques.
- **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 cumulative** : On adopte une stratégie de sélection cumulative (nested subsets). Pour garantir la cohérence de la courbe d'apprentissage, les jeux de données sont emboîtés : l'échantillon de budget $n+1$ contient intégralement l'échantillon n .
- **Processus de mise à jour continu (Warm Start)** : On adopte une stratégie d'entraînement **non-incrémentale**. À chaque palier budgétaire, le modèle n'est pas entraîné uniquement sur les nouvelles données, mais sur le **volume total cumulé** d'échantillons Twitter sélectionnés.
- **Adaptation des poids** : On active ensuite le paramètre `warm_start=True`. Cela permet au classifieur de conserver les connaissances linguistiques acquises sur le domaine source (Amazon) comme point de départ constant, tout en adaptant progressivement ses frontières de décision aux spécificités du domaine informel (Twitter) à mesure que de nouvelles données sont injectées.

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

Cette étape permet de confronter les résultats obtenus sur le domaine source aux performances réelles en milieu inconnu, mettant ainsi en lumière la chute d'efficacité liée au passage vers Twitter.

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. Evaluation 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) | 44,98 % | 0.0047 | - |
| 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. STRATÉGIES D'APPRENTISSAGE ACTIF (AL)

Pour optimiser l'adaptation au domaine cible (Twitter) et maximiser le **Retour sur Investissement (ROI)** de l'annotation, on a implémenté trois familles de stratégies de sélection. Le protocole suit une approche cumulative avec un *Fine-tuning (Warm Start)* pour assurer la continuité de l'apprentissage.

8.1. Stratégies basées sur l'incertitude

Deux métriques d'indécision du modèle ont été introduite pour identifier les échantillons situés près de la frontière de décision :

- **Entropy Sampling** : On mesure l'imprévisibilité de la distribution de probabilité des classes.
- **Margin Sampling** : On sélectionne les échantillons où la différence entre les deux classes les plus probables est minimale.

8.2. Stratégies basées sur la diversité

Pour garantir une couverture exhaustive de l'espace latent de Twitter, trois méthodes ont été déployées

- **Diversity (K-Means)** : On sélectionne les centroïdes les plus représentatifs du pool cible.
- **Density** : On priorise les échantillons situés dans les régions les plus denses du domaine cible.

- **Max_Dist** : On maximise l'exploration en sélectionnant les points les plus éloignés du jeu d'entraînement actuel.

8.3. Stratégie combinée: Hybridation Incertitude-Marge

On a conçu une méthode hybride visant à réconcilier l'incertitude et la robustesse. Pour chaque tweet x , on calcule son score d'entropie S_e et son score de marge S_m . Comme ces métriques ont des échelles différentes, on applique une **normalisation min-max** sur l'intégralité du pool : $S'_i = \frac{S_i - \min(S)}{\max(S) - \min(S)}$. Le score final de priorité est une somme pondérée $S_{total} = 0.5 \times S'_e + 0.5 \times S'_m$. Cette approche permet de sélectionner des échantillons qui font consensus sur leur difficulté pour le modèle.

9. ANALYSE ET COMPARAISON DES PERFORMANCES

Les résultats quantitatifs montrent une évolution de l'exactitude par rapport à la référence **Zero-Shot (45,23 %)** dès l'injection des premiers échantillons du domaine cible. Les courbes de performance individuelles (voir **Fig. 4 à 9** en Annexe) illustrent des dynamiques de convergence variées selon la pertinence des échantillons sélectionnés.

Tableau 4 : Comparaison des performances par stratégie (Budget 10 %)

| Stratégie | Accuracy Moyenne (%) | Gain /Random (pts) |
|-------------------|----------------------|--------------------|
| Random (Baseline) | 51,83 % | - |
| Entropy | 53,35 % | + 1.52 |
| Margin | 53,48 % | + 1.65 |
| Diversity | 52,37 % | + 0.54 |
| Combined | 52,57 % | + 0.74 |
| Max_Dist | 51,37 % | - 0.46 |
| Density | 45,27 % | - 6.56 |

L'analyse de la **Figure 3** (en Annexe) révèle que les stratégies d'incertitude (**Margin** et **Entropy**) surperforment la baseline aléatoire sur l'ensemble des budgets, atteignant un pic de **57,66 %**. À l'inverse, la stratégie **Density** présente une contre-performance notable, restant sous le niveau de la baseline jusqu'à 20 % de budget.

10. INVESTIGATION DE LA NATURE DES ÉCHANTILLONS

L'investigation qualitative au palier de 10 % de budget permet de diagnostiquer les préférences de sélection des algorithmes.

Tableau 5 : Nature des échantillons sélectionnés à 10 % de budget

| Stratégie | Longueur Moy. (car.) | Taux Exclamation (!) | Taux OOV (%) |
|----------------------|-------------------------|----------------------|--------------|
| Random (Baseline) | 59,04 | 0,48 | 58,52 % |
| Entropy | 65,24 | 0,51 | 58,88 % |
| Margin | 63,54 | 0,52 | 58,19 % |
| Diversity | 36,96 | 0,46 | 62,22 % |
| Combined | 63,20 | 0,48 | 58,27 % |
| Max_Dist | 51,42 | 0,38 | 62,43 % |
| Density | 21,64 | 0,42 | 70,84 % |

10.1. Interprétations et hypothèses de recherche

L'analyse croisée des métriques textuelles et des performances permet de dégager trois enseignements majeurs sur le comportement des modèles en contexte d'adaptation de domaine :

- **Corrélation entre dimension textuelle et pouvoir discriminant** : On observe que les méthodes basées sur l'incertitude (Entropy, Margin, Combined) privilégient des tweets significativement plus longs (> 63 car.) que la moyenne aléatoire. Un message plus long offre une densité d'information supérieure, permettant au classifieur d'ajuster ses frontières de décision dans l'espace latent de Twitter. La stratégie Margin, qui affiche la meilleure performance à 10 % (53,48 %), est également celle qui sélectionne les échantillons les plus riches en ponctuation expressive, capturant ainsi des marqueurs de sentiment forts.
- **Limites de l'approche par densité dans les domaines bruités** : La stratégie Density présente une contre-performance (45,27 % contre 51,83 % pour le hasard). Elle sélectionne des tweets extrêmement courts (21,64 car.) associés à un taux d'OOV de 70,84 %. On peut émettre l'hypothèse que l'algorithme se laisse piéger par le "bruit" structurel de Twitter (hashtags isolés, emojis récurrents), qui, bien que fréquents, sont dépourvus de substance sémantique pour une vectorisation TF-IDF.
- **Saturation lexicale et barrière inter-domaine** : Le taux global de mots inconnus (OOV) de 59,16 % observé à 100 % du budget confirme la profondeur du *Domain Shift*. Plus de la moitié du vocabulaire Twitter est absente du corpus Amazon, ce qui explique la convergence de toutes les stratégies vers un plafond de performance situé entre 56 % et 57 %.

10.2. Analyse qualitative des échantillons

Afin de valider les mécanismes de sélection propres à chaque famille d'algorithmes, nous avons extrait systématiquement les cinq premiers tweets sélectionnés par

chaque stratégie au palier critique de 10 % du budget. Cette approche permet d'observer la correspondance directe entre les critères mathématiques de sélection et la réalité sémantique des données cibles. L'intégralité de ces extractions est consultable dans le Tableau 6 (Annexe 12.1).

- **Succès de l'Incertitude (Margin/Entropy)** : Ces algorithmes ciblent des structures textuelles complexes possédant un fort pouvoir discriminant. On observe la sélection de phrases articulées autour de concessions ou de doutes, comme le tweet de la stratégie Margin : « *that would be soooooo much and geeky [...] but work to am* » (Index 18980) ou l'aveu d'indécision de Entropy : « *im soooooooooooooo confused* » (Index 247). Ces "cas frontières" obligent le modèle à affiner ses connaissances sémantiques au-delà des mots-clés triviaux.
- **Échec de la Représentativité par Densité** : La stratégie Density s'enferme dans une redondance de "bruit sémantique". Elle sélectionne exclusivement des messages génériques et extrêmement courts tels que « *happy mothersday* » (Index 35605) ou « *happy birthday* » (Index 21992). Bien que statistiquement fréquents, ces échantillons n'apportent aucune information structurelle nouvelle, expliquant la stagnation de la performance.
- **Exploration et Diversité** : Les méthodes de diversité pure capturent des contextes variés. Si Diversity se focalise sur des interactions sociales courtes comme « *good morning* » (Index 26451) ou « *thankyou* » (Index 20595), Max_Dist explore des zones sémantiques plus larges et descriptives telles que « *sunniest day in age and im in bed* » (Index 11424), assurant une couverture exhaustive de l'espace latent.
- **Performance de la méthode Combinée** : En fusionnant incertitude et représentativité, Combined sélectionne des tweets longs et informatifs traitant de sujets profonds (« *well that propels people to change direction [...]* », Index 30858), maximisant ainsi le ROI d'étiquetage par échantillon.

11. CONCLUSION

Cette étude démontre que l'adaptation de domaine entre des critiques de produits formelles (Amazon) et des messages informels (Twitter) se heurte à un décalage lexical majeur (OOV > 59 %). Cependant, l'Apprentissage Actif, et plus particulièrement la stratégie **Margin**, permet d'optimiser le transfert de connaissances en surpassant la sélection aléatoire dès les premiers paliers budgétaires. Un gain de **+8,25 points** d'exactitude par rapport au mode *zero-shot* est ainsi réalisé avec seulement 10 % des données cibles annotées.

11. 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.

12. ANNEXE

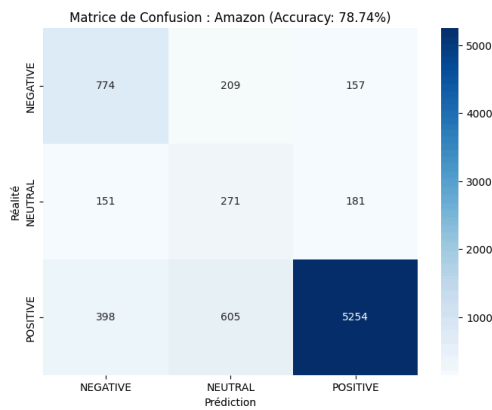


FIG. 1.A.: Performance de référence du modèle de classification de sentiments sur les données Amazon.

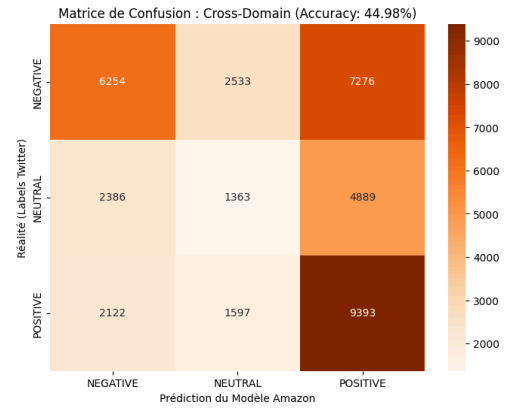


FIG. 1.B.: Généralisation du modèle Amazon sur des données hors-distribution sur les données Twitter.

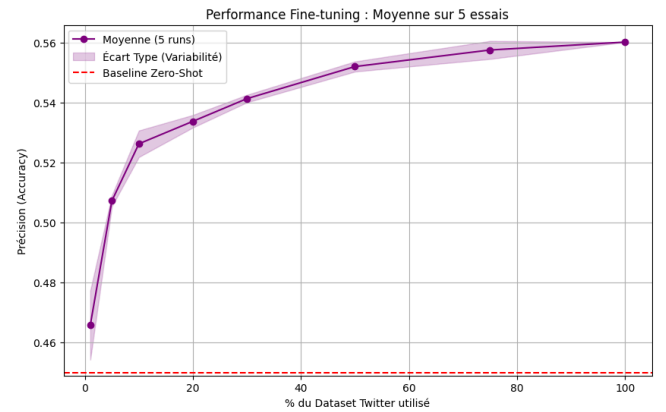


FIG. 2.: Courbe de performance du fine-tuning aléatoire (moyenne et variabilité) par rapport à la baseline Zero-Shot.

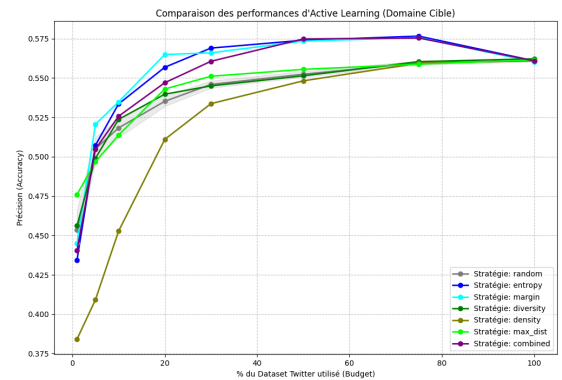


FIG. 3.: Comparaison globale des performances d'Active Learning.

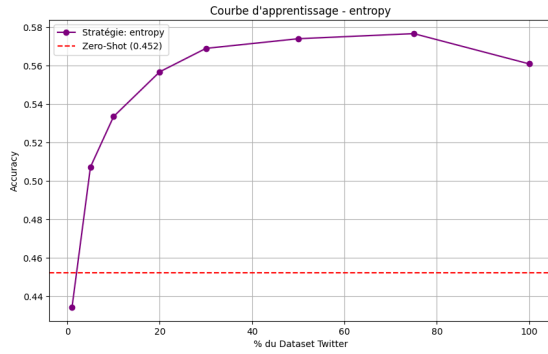


FIG. 4.: Dynamique d'apprentissage de la stratégie basée sur l'incertitude par Entropie de Shannon.

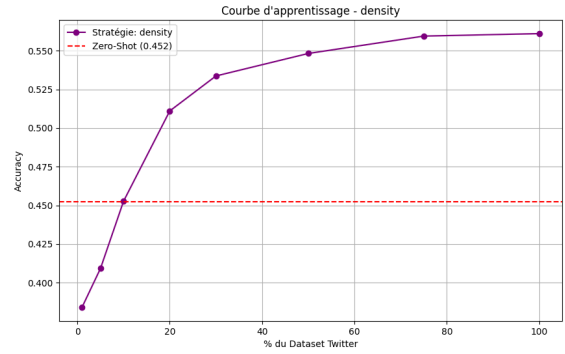


FIG. 7.: Dynamique d'apprentissage de la stratégie de représentativité basée sur la densité locale.

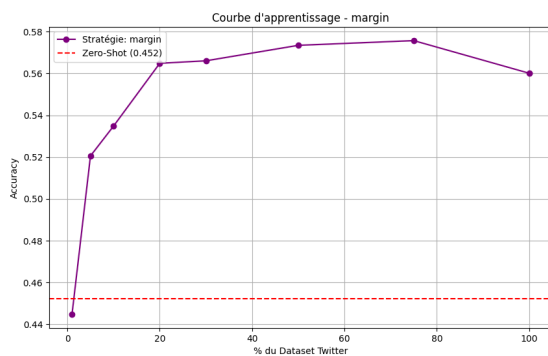


FIG. 5.: Dynamique d'apprentissage de la stratégie basée sur l'incertitude par Marge de confiance.

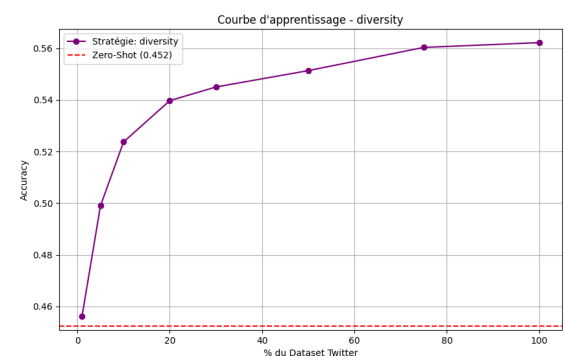


FIG. 8.: Dynamique d'apprentissage de la stratégie de diversité par partitionnement de l'espace cible via K-Means.

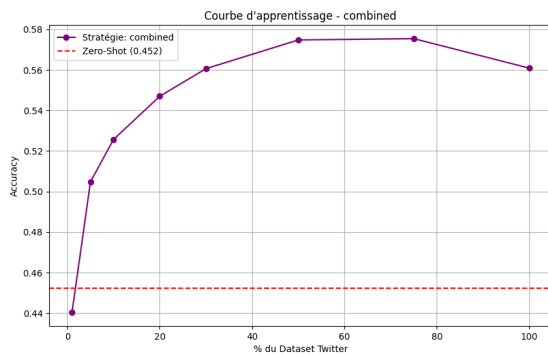


FIG. 6.: Dynamique d'apprentissage de la stratégie hybride par hybridation normalisée Incertitude-Marge.

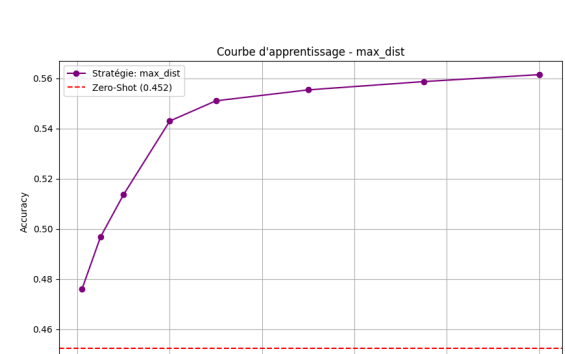


FIG. 9.: Dynamique d'apprentissage de la stratégie de diversité par exploration maximale de l'espace latent.

Tableau 6.1 : Échantillons de référence (Sélection aléatoire sans critère de pertinence) (aléatoire)

| Index | Contenu du Tweet sélectionné |
|-------|--|
| 3098 | im thirsty in the middle of the night and get to the fridge only to find my brand new bottle of crancherry juice gone fml moment!! |
| 34693 | catt totally love your default picture! you seem like such fun mom! |

8527 bwaha! it so far away though but it look fluffy!
 and jareds hair shinnny

32053 haha thats always good jam

22273 hypnotyst hmmm should beware

Tableau 6.2 : Échantillons sélectionnés par la stratégie d'incertitude : **Margin**.

| Index | Contenu du Tweet sélectionné |
|-------|---|
| 24295 | welcome back to school dont study too hard take time off and smell rose too |
| 33424 | tell brad said hi! drink and sing for me and morgan |
| 18980 | that would be soooooo much and geeky to the ultimate level! but work to am |
| 12621 | aw that wee lassie made me cry tear streamin doon ma face lol thats wee shame |
| 25597 | browsing find everything about university amp interior design |

Tableau 6.3 : Échantillons sélectionnés par la stratégie d'incertitude : **Entropy**.

| Index | Contenu du Tweet sélectionné |
|-------|---|
| 16827 | thankful for last minute doc appointment baby girl ha temp of sitting at the doc waitin |
| 36608 | cant open my eye properly maybe if sleep for lil while longer itll fix itself |
| 247 | im soooooooooooooo confused |
| 38731 | im not confused |
| 30851 | updated blog show you should be watching the unusuals go check it out |

Tableau 6.4 : Échantillons sélectionnés par la stratégie de représentativité : **Density**.

| Index | Contenu du Tweet sélectionné |
|-------|------------------------------------|
| 35605 | happy mothersday |
| 34722 | (Contenu vide après prétraitement) |
| 23596 | yummmvery nicee |
| 21992 | happy birthday |
| 36890 | happy mother day mom! |

Tableau 6.5 : Échantillons sélectionnés par la stratégie de diversité : **Diversity (K-Means)**.

| Index | Contenu du Tweet sélectionné |
|-------|------------------------------|
| 29055 | fuck you is all have to say |
| 31990 | good and you jeje |
| 20595 | thankyou |
| 26451 | good morning |
| 5086 | slept so late |

Tableau 6.6 : Échantillons sélectionnés par la stratégie de diversité : **Max_Dist**.

| Index | Contenu du Tweet sélectionné |
|-------|--|
| 11424 | sunniest day in age and im in bed |
| 38219 | not addicted just sociable |
| 9478 | looking forward to your mandarin album hope that you will come singapore again |
| 29423 | scream just played on my ipod first thing that come to mind bear machineeeee!!!! lol |
| 4195 | lont wanna get out of bed wanna go back to sleep have to open the store at work though |

Tableau 6.7 : Échantillons sélectionnés par la stratégie hybride : **Combined**.

| Index | Contenu du Tweet sélectionné |
|-------|--|
| 30858 | well that propels people to change direction no point wasting your day on something you lost the passion for |
| 10213 | majorspoilerscom ha problem cannot get the site working |
| 37822 | yay! thanks coveri sad ang alist please |
| 13535 | the following week would be better ill be gone next week |
| 6994 | saw the sun but then blink and it wa gone |