Rapport d'Évaluation des modèles pour analyser l'Analyse de sentiments

Auteurs: Sara Bevilacqua, Ziad Lahrouni, Sabrina Tamda, Mohammed Kerraz

Github : github.com/Zlahrouni/twitter-sentiment-analysis

Introduction et modélisation

Le modèle d'analyse de sentiment présenté ici traite des tweets en français pour déterminer s'ils expriment une émotion positive ou négative. Il utilise **TF-IDF**, qui attribue un poids aux mots en fonction de leur fréquence dans un tweet et dans l'ensemble du corpus, rendant l'analyse plus précise que le Bag of Words, qui ne prend en compte que la présence des mots sans différencier leur importance. Deux modèles de régression logistique sont entraînés séparément : l'un pour détecter la positivité, l'autre pour détecter la négativité. L'option **class_weight='balanced'** est utilisée pour corriger un éventuel déséquilibre dans les labels, assurant ainsi de meilleures performances, même sur des tweets minoritaires.

Concernant les données, nous avons utilisé une **base de tweets en français** issue de Kaggle. 95% des ces données sont utilisées pour l'entraînement (75% de cette base) et de l'évaluation (le jeu de test fait donc 25% de cette base). Les 5% restants seront exploités par un cron job afin d'alimenter et d'améliorer continuellement le modèle avec de nouvelles données au fil du temps.

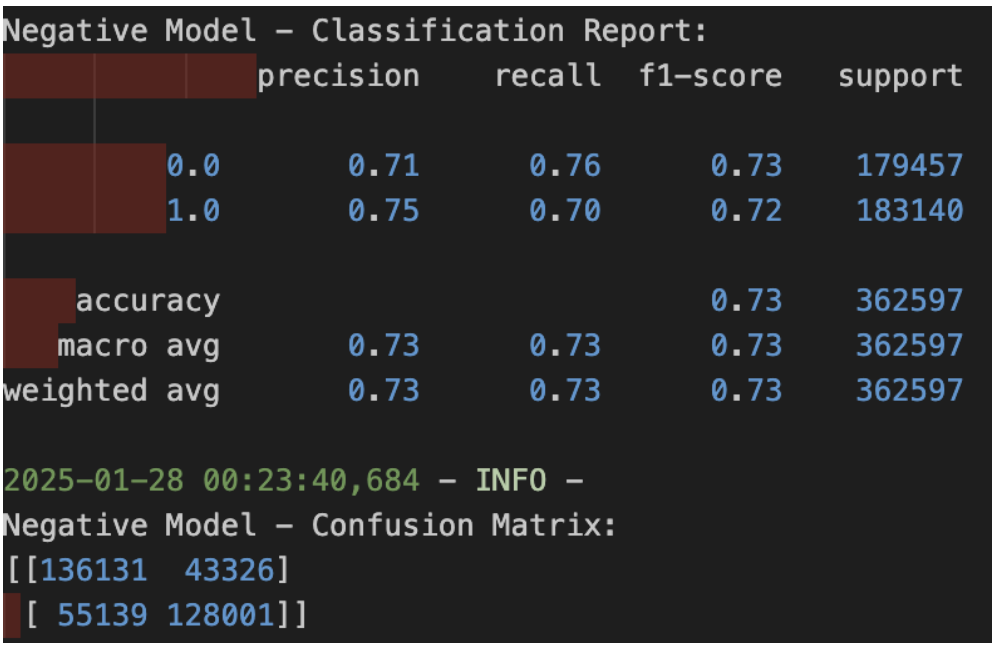
Une fois l'entraînement terminé, le modèle est sauvegardé sur le disque dur avec le vectorizer (TF-IDF). Cette étape de sérialisation est nécessaire pour une utilisation future des modèles (et du pré-processing) pour des prédictions. Il est ainsi possible de les charger dans un endpoint par exemple.

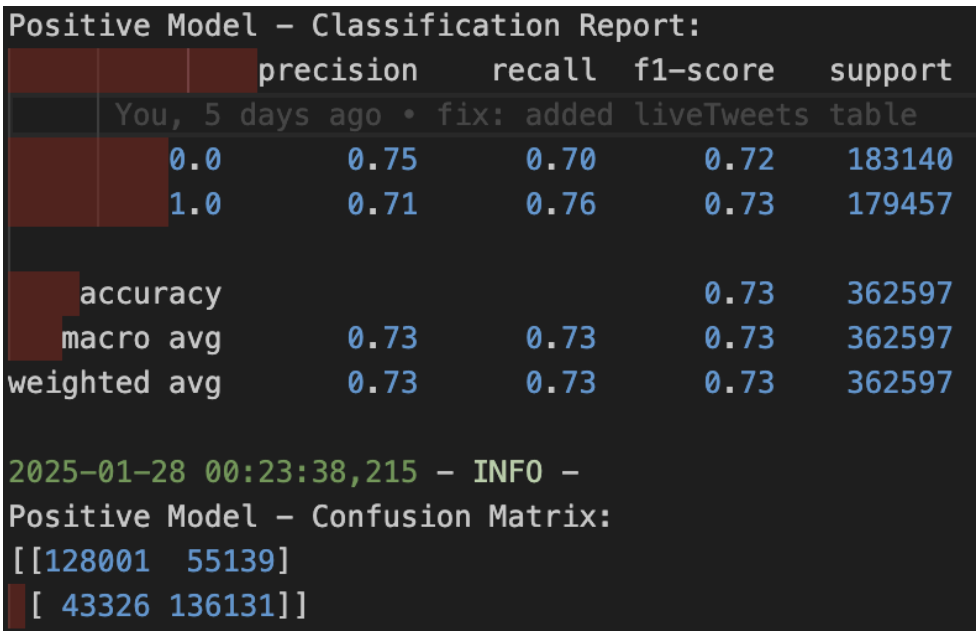
Présentation des résultats deux modèles

Afin d'évaluer les deux modèles proposés dans ce projet, plusieurs métriques de classification ont été calculées ainsi qu'une matrice de confusion.

Les deux modèles sont entraînés sur les mêmes données en input, et ont exactement les mêmes résultats. Cela est logique étant donné les caractéristiques des labels à prédire dans chacune des deux classifications. Dans la première, il faut prédire si un texte en input est positif ou non, dans la seconde il faut prédire si un texte est négatif ou non. Ce sont deux classifications binaires, avec deux booléens. Si un texte est "positif" il aura le label 1 pour la classification "positif" et le label 0 pour la classification "négatif", et inversement. Ainsi, les deux classifications vont prédire la même chose sur les mêmes inputs, mais avec un affichage différent.

En théorie, la seule différence visible entre les deux modèles pourrait être liée à l'aléatoire présent durant l'entraînement. Cela n'est pas observé en pratique cependant. Les résultats étant donc exactement symétriques, un seul modèle sera considéré dans les analyses qui vont suivre (prédiction de sentiment positif).





Analyses des matrices de confusion

La matrice de confusion du modèle de sentiment positif (positive model) présente les résultats suivants :

Matrice de Confusion

	Prédit Négatifs	Prédit Positifs
Vrai Négatifs	**128,001** Vrais Négatifs (TN)	**55,139** Faux Positifs (FP)
Vrai Positifs	**43,326** Faux Négatifs (FN)	**136,131** Vrais Positifs (TP)

Les cellules vertes représentent les prédictions correctes, les rouges les erreurs

128001 (Vrais négatifs - TN) : Nombre de commentaires négatifs correctement classés comme négatifs.

- **43326 (Faux négatifs - FN)** : Nombre de commentaires positifs incorrectement classés comme négatifs.
- **55139 (Faux positifs - FP)** : Nombre de commentaires négatifs incorrectement classés comme positifs.

- **136131 (Vrais positifs - TP)** : Nombre de commentaires positifs correctement classés comme positifs.

À partir de la matrice de confusion, on peut calculer les taux d'erreur en comparant le nombre de prédictions incorrectes avec le total des cas réels pour chaque classe. Le taux de faux positifs (FPR) correspond à la proportion de commentaires négatifs qui ont été incorrectement classés comme positifs par rapport à l'ensemble des commentaires réellement négatifs. De même, le taux de faux négatifs (FNR) représente la proportion de commentaires positifs qui ont été mal identifiés comme négatifs par rapport à l'ensemble des commentaires réellement positifs.

Le modèle présente un taux de faux positifs (FPR) de 28.83%, indiquant que 28.83% des commentaires prédits comme positifs étaient finalement négatifs. Par ailleurs, son taux de faux négatifs (FNR) est de 25.29%, ce qui signifie que 25.29% des commentaires prédits comme négatifs étaient en fait positifs. Cela montre que le modèle est légèrement meilleur pour détecter les commentaires négatifs**.**

Analyses des indices de précision, rappel et F1-score

Les scores globaux fournis sont :

Classe	Précision	Rappel	F1-score	Support
0	0.75	0.70	0.72	183140
1	0.71	0.76	0.73	179457

La précision mesure la proportion des prédictions positives qui sont réellement positives. Pour la classe 0, la précision est de 0.75, ce qui signifie que 75% des commentaires classés comme négatifs sont effectivement négatifs. Pour la classe 1, la précision est de 0.71, indiquant que 71% des commentaires classés comme positifs sont effectivement positifs. Cela montre que les deux classes ont des performances relativement proches en termes de précision, mais que le modèle a une **précision légèrement meilleure pour les commentaires négatifs**. Cependant, cette différence doit être mise **en perspective avec les résultats du rappel, qui montrent une tendance inverse.**

Le rappel indique la proportion des commentaires réels d'une classe qui ont été correctement détectés par le modèle. Pour la classe 0, le rappel est de 0.70, ce qui signifie que 70% des commentaires négatifs ont été correctement identifiés comme négatifs. En revanche, pour la classe 1, le rappel est de 0.76, indiquant que 76% des commentaires positifs ont été correctement identifiés comme positifs.

Il est nécessaire de trouver un compromis entre la précision et le rappel. Si on veut améliorer la précision, on risque de pénaliser le rappel, et inversement. C'est le F1-score qui joue ce rôle. C'est une moyenne harmonique entre la précision et le rappel, qui permet de mesurer l'équilibre entre ces deux métriques. Pour la classe 0, il est de 0.72, tandis que pour la classe 1, il est de 0.73. Cela confirme que **le modèle est globalement équilibré entre les deux classes, bien qu'il soit légèrement plus performant pour la détection des commentaires positifs**, notamment grâce à son rappel plus élevé. Toutefois, la différence est faible, ce qui indique que le modèle maintient une performance relativement homogène sur les deux types de commentaires.

Analyses des performances: forces, faiblesses, biais

L'analyse des performances du modèle est particulièrement importante dans le cadre d'un réseau social, où la détection des commentaires négatifs peut permettre de modérer le contenu et de prévenir des interactions toxiques. Une mauvaise classification pourrait soit laisser passer des commentaires inappropriés, soit censurer injustement du contenu légitime, ce qui a des implications directes sur l'expérience utilisateur et la gestion de la communauté.

Forces :

✅ **Précision élevée pour les commentaires négatifs (Précision 75%)**
Lorsqu'un commentaire est classé comme négatif, il y a une forte probabilité (75%) que cette classification soit correcte, ce qui réduit le risque de modération injustifiée sur du contenu neutre ou acceptable.

✅ **Bonne détection des commentaires positifs (Rappel 76%)**
Le modèle parvient à identifier efficacement les commentaires positifs, ce qui est bénéfique pour encourager les interactions constructives et valoriser les échanges sains sur la plateforme.

✅ **Équilibre général entre précision et rappel**
Avec un F1-score global d'environ 0.73 (accuracy), le modèle maintient une

performance homogène entre la détection des commentaires positifs et négatifs, ce qui limite les déséquilibres trop marqués.

Faiblesses :

🆘 Taux de faux positifs trop élevé (28.83%) : risque de laisser passer du contenu inapproprié

Le modèle classe presque 30% des commentaires négatifs comme positifs, ce qui signifie qu'il ne détecte pas une part importante de propos potentiellement toxiques ou inappropriés. Dans un contexte de modération, cela peut laisser passer des commentaires nuisibles qui auraient dû être signalés ou supprimés, réduisant ainsi l'efficacité du contrôle du contenu et la prévention des interactions négatives.

🆘 Taux de faux négatifs très élevé (25.29%) : risque de sur-censure

Un commentaire positif sur quatre est mal classé comme négatif, cela peut entraîner une censure excessive de discussions légitimes et nuire à l'expérience utilisateur en supprimant ou limitant l'accès à des contenus qui ne présentent pas de risque.

🆘 Difficulté à différencier la nuance dans les commentaires négatifs
Certains commentaires négatifs peuvent être constructifs (ex. critique argumentée), tandis que d'autres peuvent être toxiques. Le modèle ne fait pas de distinction, ce qui peut entraîner une classification erronée et une modération excessive ou insuffisante.

Biais :

⚠️ Impact des données d'entraînement
Un des biais majeurs du modèle provient de la nature des données d'entraînement, qui contiennent un grand nombre de commentaires **neutres**. Actuellement, le modèle fonctionne sur une classification binaire (positif vs. négatif), attribuant un label positif ou négatif à chaque commentaire, même lorsqu'il est neutre.

Recommandation pour améliorer le modèle

Afin d'optimiser la détection des commentaires négatifs et d'assurer une modération efficace tout en minimisant les erreurs, plusieurs améliorations peuvent être envisagées.

=>Introduction d'un label neutre pour une meilleure classification (et multi-class classification plus largement)

Actuellement, le modèle repose sur une classification binaire (positif vs. négatif), ce qui pose un problème pour les commentaires neutres ou ambigus. Certains messages ne sont ni positifs ni négatifs, mais simplement factuels, interrogatifs ou descriptifs, rendant leur classification forcée et imprécise. Cela peut entraîner une sur-censure si des commentaires neutres sont classés comme négatifs, ou à l'inverse, une sous-modération si des commentaires négatifs sont classés comme neutres.

La solution serait d'opter pour une classification **multi-classes (positif / neutre / négatif)**. L'ajout d'un label neutre permettrait de mieux différencier les critiques constructives des propos réellement toxiques, réduisant ainsi le taux de faux positifs et faux négatifs. Cette approche améliorerait la précision du modèle et permettrait d'adapter la modération en fonction de la nature du commentaire.

=> Ajustement du seuil de classification

Actuellement, le modèle est trop permissif avec les commentaires négatifs, ce qui entraîne **un taux élevé de faux positifs** et laisse passer des contenus inappropriés. Une solution serait d'**ajuster le seuil de classification**, en abaissant la tolérance pour les commentaires négatifs. Une analyse de la **courbe ROC-AUC** permettrait d'identifier un seuil optimal qui maximise la précision des négatifs tout en conservant un bon rappel.

=> Amélioration du prétraitement des données

Un **meilleur prétraitement des données** pourrait réduire les erreurs de classification. Des techniques comme la suppression des mots vides et la lemmatisation aideraient le modèle à mieux comprendre le sens des phrases, tandis que la détection des sarcasmes et des emojis permettrait de mieux interpréter le ton des messages. De plus, l'identification de

mots-clés à risque renforcerait la détection des contenus offensants ou toxiques.

Enfin, les **données n'ont pas été ramenées à la même échelle** (via une standardisation ou une normalisation), une optimisation de ce prétraitement pourrait améliorer sensiblement les performances du modèle en réduisant les biais liés aux variations de texte

=> Utilisation de modèles plus avancés

La régression logistique, utilisée actuellement, est un modèle simple qui peut être amélioré avec des techniques plus performantes, notamment pour mieux capturer la complexité du langage et réduire les erreurs de classification. Une première approche consisterait à utiliser des **Réseaux de Neurones Récurrents (RNN) ou des LSTM**, qui permettent de traiter les phrases dans leur entièreté et de mieux comprendre le contexte des mots au sein d'un commentaire. Une autre alternative plus avancée serait d'exploiter des modèles basés sur **les Transformers** comme **BERT, RoBERTa ou GPT**, capables d'analyser les relations complexes entre les mots et d'améliorer significativement la détection des commentaires toxiques.

Enfin, une approche **hybride combinant l'IA et des règles linguistiques** pourrait être particulièrement efficace. En intégrant des règles définies par des experts du langage, il serait possible de cibler plus précisément certains types de propos problématiques, tout en bénéficiant de la puissance des modèles d'apprentissage automatique pour affiner la classification des commentaires. Cette combinaison permettrait d'obtenir un modèle plus robuste, interprétable et performant dans un contexte de modération automatique.