Université Abdelmalek Essaadi
Faculté des Science et Techniques
d'Al-Hoceima

Analyse des sentiments des tweets

I. Analyse des sentiments des tweets

1.definition de dataset :

Le dataset des sentiments également connu sous le nom de Twitter Sentiment Analysis Dataset, est un ensemble de données largement utilisé dans le domaine de l'analyse des sentiments. Il comprend des tweets collectés à partir de Twitter, annotés avec l'étiquette du sentiment correspondant, généralement positif et négatif.

Lien de dataset : https://www.kaggle.com/datasets/kamyab20/sentiment140mv

2. Les performance de chaque modele

Dataset des sentiments	**Accuracy**	**F1-score**	**Precision**	**Recall**
Model 1 : RNN_LSTM	85%	85%	85%	85%
Model 2 : MNB	78%	85%	82%	78%
Model 3 : CNB	85%	85%	85%	85%
Model 4 : RL	87%	87%	88%	88%
Model 5 : AdaBoost	80%	81%	81%	80%
Model 6 : NusSVC	87%	88%	88%	88%

3.Interpretation des résultats

Le modele MultinomialNB est moins performant en comparaison avec les autres modelés ; bien que le MNB soit simple et rapide à entrainer, il suppose l'indépendance conditionnelle entre les caractéristiques ce qui n'est pas entièrement applicable aux données textuelles, aussi il est plus performant sur les datasets ou les classes sont déséquilibrés parce qu'il est basé sur le fréquence brute dans les calculs des probabilités.

Concernant le modele ComplementNB est parmi les algorithmes les plus performant sur ce dataset des sentiments en notant sa rapidité et cela revient au fait que CNB est performant sur les jeux donnés déséquilibré comme dans notre cas où il fonctionne en ajustant la manière dans les probabilités sont calculées pour tenir compte des données manquants par rapport à d'autre classes (basé sur la fréquence complément).

Pour Nusvc et la régression logistique, ces modèles ont réussi à capturer efficacement les relations entre les caractéristiques des données et des classes des sentiments leurs efficacité peut être attribue à leur capacité à modéliser des frontières de décision complexes tout en maintenant une interprétabilité raisonnable ce qui en fait des choix attrayants pour les applications de classifications.

Pour Adaboost, les performances sont moyennes en comparaissant avec d'autre modèles, il a démontré une capacité à améliorer progressivement la précision de la classification en agrégeant plusieurs classificateurs faibles, cependant son efficacité peut être limité dans des cas ou les données sont déséquilibrées ou lorsque les caractéristiques discriminantes sont difficiles à extraire ce qui peut entrainer une performance moins satisfont par rapport à d'autre modèles

Les résultats obtenus avec le modèle RNN_LSTM sont remarquables. Avec une précision, un rappel et un F1-score de 85%, ce modèle s'est révélé robuste et efficace pour prédire les émotions associées à des textes. Cette performance élevée peut être attribuée à la capacité des réseaux de neurones récurrents à prendre en compte les dépendances temporelles dans les données textuelles. Contrairement aux approches traditionnelles, les LSTM capturent les relations séquentielles entre les mots, ce qui leur permet de mieux comprendre le contexte et le sens des phrases. De plus, les LSTM sont capables de traiter des séquences de longueur variable, ce qui les rend adaptés à la nature flexible du langage naturel.

Bien que le modèle RNN_LSTM ait produit des résultats solides, il moins performant que les modèles tels que NusSVC et RL dans certains cas en raison de plusieurs facteurs ; la performance des modèles peut être influencée par la taille et la qualité des données d'entraînement.

I. Analyse des émotions des tweets :

1.definition de dataset :

Le dataset "Emotions" est une collection de messages en anglais provenant de Twitter, minutieusement annotés avec six émotions fondamentales : la colère, la peur, la joie, l'amour, la tristesse et la surprise. Ce dataset constitue une ressource précieuse pour comprendre et analyser le spectre diversifié des émotions exprimées dans les textes courts sur les réseaux sociaux. Chaque entrée dans ce dataset se compose d'un segment de texte représentant un message Twitter et d'une étiquette correspondante indiquant l'émotion prédominante transmise. Les émotions sont classées en six catégories : la tristesse (0), la joie (1), l'amour (2), la colère (3), la peur (4) et la surprise (5).

Lien de dataset : https://www.kaggle.com/datasets/nelgiriyewithana/emotions

2.les performances de chaque modele

Dataset des émotions	**Accuracy**	**F1-score**	**Precision**	**Recall**
Model 1 : RNN_LSTM	93%	93%	93%	93%
Model 2 : MNB	76%	76%	80%	76%

Model 3 : RL	89%	89%	89%	89%
Model 4 : CNB	88%	88%	88%	88%
Model 5 : CNN	93%	93%	93%	93%
Model 6 : AdaBoost	0.36%	0.21%	0.24%	_

3.interpretation :

En se basant sur les résultats obtenus dans les trois modèles, on peut donner les mêmes interprétations obtenues déjà dans le dataset précèdent, sauf qu'on peut ajouter que CNB et régression logistique peut s'adapter avec des datasets complexe et volumineux, avec une rapidité dans la génération des résultats en comparaison avec des algorithmes qui sont performant mais il prend beaucoup de temps à s'exécuter.

Les performances comparables et élevé du CNN et du RNN-LSTM sur le dataset des émotions suggèrent que ces deux architectures sont également efficaces pour traiter ce type de données. Bien que le CNN soit généralement considéré comme plus adapté pour l'extraction de caractéristiques locales à partir de données structurées, et que le RNN-LSTM soit privilégié pour modéliser les dépendances séquentielles à long terme, les résultats montrent que les deux approches sont capables de capturer efficacement les informations pertinentes pour la classification des émotions. Cette observation met en lumière la flexibilité et la robustesse des réseaux de neurones profonds, qui peuvent s'adapter à différents types de données et produire des performances comparables dans des contextes varié.

Notant que l'application de nuSVC sur ce dataset n'aboutit à aucun résultat malgré l'optimisation des paramètres ce qui peut être explique par l'incapacité de ce modele à traiter des datasets complexes et volumineux.

Malgré l'augmentation du nombre d'estimateurs dans le modèle Adaboost et l'utilisation de la validation croisée, les performances restent en deçà des attentes. Cette situation peut être attribuée à plusieurs facteurs ;tel que la complexité inhérente des données peut rendre l'application de modèles tels qu'Adaboost mal adaptée, le déséquilibre entre les classes .

II. Préparation des données :

1.stemming et lemmatisation :

Le stemming et lemmatisation sont des techniques utilisée en traitement automatique de langage naturel pour réduire les mots à leurs formes de base ou racine, il consiste à supprimer les affixes des mots pour extraire leurs racines, il utilise des règles heuristiques simples pour tronquer les mots ce qui rend le processus simple et rapide, Cependant le stemming peut introduire des biais et des erreurs par exemple il peut fusionner incorrectement des mots distincts comme expérience et expérimentation en la même racine, ces erreurs peuvent affecter la précision d'un modele de classification comme dans le cas de la dataset des émotions (multicalss) lorsque on a essayé d'appliquer le stemming pour le nettoyage des données les performances de modele a diminué tandis que lorsque on a appliqué la lemmatisation les performance de modèles ont augmentés , et

cela revient au fait que La lemmatisation vise à ramener les mots à leur forme canonique ou lemmes, en utilisant des ressources lexicales et des analyses morphologiques plus complexes.

Mais l'application de stemming et de lemmatisation sur la première dataset a donné les mêmes résultats ce qui veut dire que le choix entre le stemming et la lemmatisation dépendra des besoins spécifiques, y compris la taille du jeu de données, la complexité linguistique et les contraintes de temps de traitement.

2.Bag of words (BOW):

Bag of Words (BoW) est l'une des techniques les plus couramment utilisées dans le domaine du traitement automatique du langage naturel (NLTK). Il repose sur un principe simple : représenter un document sous forme d'un vecteur contenant la fréquence d'apparition de chaque mot dans un corpus de documents. Cette approche ignore l'ordre et la structure des mots dans le texte, se concentrant uniquement sur leur occurrence.

Parmi les techniques les plus courant de BOW et que nous avons utilisé dans les différents modelés qu'on a créé pour les deux dataset ; CountVectorizer : Il s'agit de la méthode la plus simple pour créer des vecteurs BoW, Elle compte simplement le nombre d'occurrences de chaque mot dans chaque document, et TF-IDF (Term Frequency-Inverse Document Frequency) ; Cette méthode attribue des poids aux mots en fonction de leur fréquence dans le document et de leur importance dans l'ensemble du corpus, le choix de l'approche convenable dépend de la nature des données, objectif de la tâche, taille de corpus et complexité de modèles. Il est souvent utile d'expérimenter plusieurs approches et de comparer leurs performances pour trouver celle qui fonctionne le mieux pour votre cas d'utilisation spécifique.

On a essayé d'appliquer TFIDvectorier et CountVectorizer sur quelques modèles comme les tableaux ci-dessous montrent :

Count vectorizer	Dataset des sentiments				Dataset des émotions			
Métriques	Accuracy	Précession	F1_score	Recall	Accuracy	Precision	F1_score	Recall
RL	0.89	0.90	0.91	0.93	O.89	O.89	O.89	O.89
CNB	0.85	0.89	0.88	0.87	O.89	O.89	O.89	O.89
MNB	0.84	0.85	0.88	0.91	0.86	0.86	0.86	0.86

TFID vectorizer	Dataset des sentiments				Dataset des émotions			
Métriques	Accuracy	Précession	F1_score	Recall	Accuracy	Precision	F1_score	Recall
RL	0.87	0.87	0.90	0.94	O.89	O.89	O.89	O.89
CNB	0.84	0.86	0.88	0.89	0.88	0.88	0.88	0.88
MNB	0.78	0.75	0.85	0.98	0.76	0.80	0.73	0.76

En général les résultats obtenus montrent clairement que l'approche utilise dépend de dataset, modelés appliqués, et objectif de modele.

3. division des données (train et test) :

Dataset des sentiments (positive et négative)

Modele de régression logistique :

Test_size	Accuracy	Recall	Précision	F1_score
0.1	0.87	0.94	0.86	0.90
0.2	0.87	0.94	0.87	0.90
0.3	0.86	0.94	0.86	0.90
0.4	0.87	0.94	0.86	0.89

Modele de complément NB :

Test_size	Accuracy	Recall	Précision	F1_score
0.1	0.87	0.94	0.86	0.90
0.2	0.85	0.87	0.89	0.88
0.3	0.86	0.94	0.86	0.90
0.4	0.87	0.94	0.86	0.89

Dataset des émotions (multicalss)

Modele de régression logistique :

Test_size	Accuracy	Recall	Précision	F1_score
0.1	0.88	0.88	0.88	0.88
0.2	0.89	0.89	0.89	0.89
0.3	0.89	0.89	0.89	0.89
0.4	0.89	0.89	0.89	0.89

Modele de complément NB :

Test_size	Accuracy	Recall	Précision	F1_score
0.1	0.85	0.86	0.90	0.88
0.2	0.88	0.88	0.88	0.88
0.3	0.85	0.86	0.90	0.88
0.4	0.88	0.88	0.88	0.88

À la lumière des résultats observés, il est clair que la taille de l'ensemble de test joue un rôle crucial dans l'évaluation des performances des modèles. Pour le dataset des sentiments, où seules deux classes sont en jeu (positif et négatif), les performances des modèles semblent être moins sensibles aux variations de la taille de l'ensemble de test. Cependant, pour le dataset des émotions, qui implique une classification multi-classe, les performances semblent être plus sensibles à ces variations. Ces observations soulignent l'importance de choisir soigneusement la taille de l'ensemble de test, en particulier dans les scénarios où la tâche de classification est plus complexe. De plus, ces résultats mettent en évidence la nécessité d'une évaluation minutieuse des performances des modèles dans divers contextes, en tenant compte de la nature spécifique des données et des objectifs de l'analyse. En fin de compte, la taille de l'ensemble de test est un facteur critique pour obtenir une estimation précise de la performance du modèle. Une taille d'ensemble de test insuffisante peut conduire à une estimation biaisée de la performance du modèle, tandis qu'une taille d'ensemble de test plus grande permet une estimation plus fiable de la performance réelle du modèle.