UEMF
الجامعة الأورومتوسطية بفاس
EUROMED UNIVERSITY OF FES
UNIVERSITÉ EUROMED DE FÈS

EIDIA
École d'Ingénierie Digitale et d'Intelligence Artificielle

TEXT MINING BASÉE SUR TF-IDF SVM

RÉALISÉ PAR :

EDDOUKS OUMAYMA

Introduction

- L'article "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification"présente une nouvelle approche de classification des actualités à l'aide de TF-IDF et SVM. Les auteurs expliquent la motivation derrière le développement d'une telle approche, qui est d'aider les utilisateurs à recevoir rapidement des informations d'actualité pertinentes sans avoir à lire toutes les actualités. Le document donne également un aperçu des études antérieures sur la classification des nouvelles et compare l'approche proposée avec d'autres méthodes. Enfin, les auteurs évaluent l'efficacité de leur approche à l'aide de mesures de performance standard et fournissent une discussion et une conclusion.

Objectif

- L'objectif de cet article est de proposer une nouvelle approche pour la classification des nouvelles en utilisant TF-IDF et SVM . L'approche proposée est évaluée à l'aide de mesures de performance standard et comparée à d'autres méthodes. Le but ultime est de fournir une méthode efficace et efficiente pour la classification des nouvelles qui peut être utilisée dans diverses applications.

Dataset

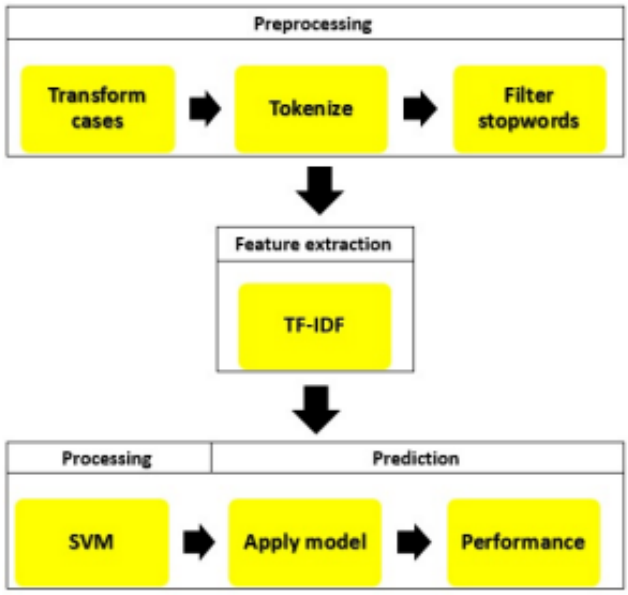
L'approche proposée pour la classification des nouvelles est évaluée à l'aide de deux ensembles de données :BBC et 20Newsgroup.

- les données BBC se composent de 2225 textes d'actualités de 2004 à 2005, couvrant cinq sujets : affaires, divertissement, politique, sports et technologie. L'ensemble de données a été collecté par les auteurs en grattant des articles de presse sur le site Web de la BBC. Les textes ont été prétraités et utilisés pour former et tester l'approche proposée pour la classification des nouvelles. BBC est utile pour évaluer l'efficacité de l'approche proposée dans la classification des articles de presse de différents domaines.
- L'ensemble de données de 20 groupes de discussion comprend 1 997 articles de presse recueillis sur Internet dans 20 classes. il a été créé à l'origine pour la recherche sur la classification de texte et est largement utilisé comme ensemble de données de référence pour évaluer les algorithmes de classification de texte. 20Newsgroup comprend des textes sur une variété de sujets, tels que l'infographie, les objets proposés à la vente, le baseball, le christianisme et des textes politiques sur les armes à feu. L'ensemble de données est utile pour évaluer l'efficacité de l'approche proposée dans la classification des articles de presse de différents domaines et la comparer avec d'autres méthodes.

L'APPROCHE PROPOSÉE

- L'approche proposée comprend trois étapes :

Le prétraitement du texte , la sélection des features basée sur TF-IDF et la clasification du texte à l'aide du modèle SVM.
La figure 1 décrit l'architecture générale de la méthode proposée



Étapes de l'algorithme

BBC

CHARGEMENT DES DONNÉES

la fonction pd.read_csv() de la bibliothèque pandas est utile pour lire le fichier CSV "bbc.csv" et stocker les données dans un DataFrame appelé df

VISUALISATION DU GRAPHIQUE DE DÉCOMPTE DES CATÉGORIES

La visualisation du graphique se fait en utilisant la fonction **sns.countplot()** de la bibliothèque **seaborn**.Le graphique résultant est un histogramme qui montre combien de fois chaque catégorie apparaît dans les données. Cela peut aider à comprendre la répartition des données par catégorie et à obtenir des informations sur la distribution des données.

PRÉTRAITEMENT DES DONNÉES

1. Transformation en minuscules : La première étape consiste à convertir tout le texte de la colonne 'text' en minuscules. Cela permet d'uniformiser le texte et d'éviter les différences de casse qui pourraient affecter les analyses ultérieures.
2. Suppression des mots vides (stop words) : Dans cette étape, les mots vides en anglais sont téléchargés à partir de la bibliothèque nltk, puis une liste de ces mots est créée. Ensuite, chaque texte de la colonne 'text' est parcouru et les mots qui font partie de la liste des mots vides sont supprimés.

L'EXTRACTION DES FEATURES(TF-IDF)

1. Division des données en ensembles d'entraînement et de test :
2. la vectorisation du texte en utilisant TfidfVectorizer. La méthode fit_transform de TfidfVectorizer est utilisée pour ajuster le modèle aux données d'entraînement et transformer les données d'entraînement et de test en représentation vectorielle.

CLASSIFICATION (SVM)

1. Apprentissage et prédiction avec un modèle SVM : Un modèle SVM (Support Vector Machine) linéaire est initialisé avec les paramètres spécifiés. Le modèle est entraîné sur l'ensemble d'entraînement à l'aide de la méthode fit. Ensuite, les prédictions sont effectuées sur l'ensemble de test à l'aide de la méthode predict.
2. Calcul des mesures de précision : L'exactitude (accuracy_score) du modèle SVM est calculée en comparant les prédictions avec les vraies étiquettes de l'ensemble de test.

RÉSULTAT

SVM ACCURACY SCORE -> 98.74326750448833

3. Rapport de classification : Le rapport de classification (classification_report) est calculé pour fournir des mesures telles que la précision, le rappel et le F-mesure pour chaque classe.

	precision	recall	f1-score	support
business	0.99	0.98	0.99	126
entertainment	0.97	0.99	0.98	95
politics	0.99	0.97	0.98	94
sport	1.00	1.00	1.00	147
tech	0.98	0.99	0.98	95
accuracy			0.99	557
macro avg	0.99	0.99	0.99	557
weighted avg	0.99	0.99	0.99	557

20 NEWSGROUP

En appliquant les mêmes étapes de prétraitement et de classification utilisées précédemment pour l'ensemble de données BBC, nous avons obtenu les résultats suivants pour le jeu de données 20 Newsgroups.

RÉSULTAT

SVM ACCURACY SCORE -> 96.96000000000001

	precision	recall	f1-score	support
comp.graphics	0.94	0.98	0.96	245
misc.forsale	0.94	0.97	0.95	257
rec.sport.baseball	1.00	0.97	0.98	250
soc.religion.christian	0.99	0.95	0.97	242
talk.politics.guns	0.99	0.97	0.98	256
accuracy			0.97	1250
macro avg	0.97	0.97	0.97	1250
weighted avg	0.97	0.97	0.97	1250

Conclusion

En conclusion ,il semble que les résultats que nous avons obtenus sont supérieurs aux résultats rapportés dans l'article, qui étaient de 97,84 % et 94,93 % pour les ensembles de données de la BBC et de 20Newsgroup, respectivement. Cependant, il est important de noter qu'il peut y avoir plusieurs raisons à la différence de résultats, telles que des différences dans les étapes de prétraitement, l'extraction de caractéristiques ou le réglage des hyperparamètres. Il est également possible que la différence soit due au hasard, car les performances des algorithmes d'apprentissage automatique peuvent varier en fonction de l'ensemble de données spécifique et de l'initialisation aléatoire de l'algorithme.