



# **Review of Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for NewsClassification**

15.06.2023

El Mouaquit Nizar

EIDIA

UEMF

## la problématique et les objectifs du papier

La problématique abordée dans le papier concerne la classification des actualités, c'est-à-dire la tâche consistant à attribuer des articles d'actualités à des catégories spécifiques en fonction de leur contenu. La classification précise des actualités est un défi important dans le domaine de l'analyse textuelle, car elle permet d'organiser et de filtrer efficacement une grande quantité d'informations.

Les objectifs du papier sont de proposer une approche novatrice pour la classification des actualités en utilisant la technique de TF-IDF et l'algorithme de machine à vecteurs de support (SVM). L'objectif principal est de développer une méthode efficace et précise pour la classification des articles d'actualités, en utilisant des caractéristiques extraites à l'aide de TF-IDF et en appliquant SVM comme algorithme de classification.

## Etat de l'art sur lequel s'est basé le travail

Le travail s'est appuyé sur l'état de l'art existant en matière de classification des actualités et de techniques de text mining. Les auteurs ont réalisé une revue de la littérature pour examiner les travaux antérieurs qui ont abordé des problèmes similaires ou liés à la classification des actualités.

L'état de l'art a permis aux auteurs de se familiariser avec les approches existantes, les méthodes utilisées et les résultats obtenus dans le domaine de la classification des actualités. Ils ont identifié les forces et les faiblesses de ces approches et ont cherché à proposer une approche novatrice qui puisse surmonter les limitations des travaux précédents.

Parmi les approches couramment utilisées dans l'état de l'art figuraient :


1. Classification basée sur les fréquences de termes (TF) et les fréquences inverses de documents (IDF) : Cette approche utilise les fréquences des termes dans les documents et le corpus pour représenter les caractéristiques des articles d'actualités. Cependant, elle peut être sensible aux mots fréquents qui ne sont pas discriminants.

2. Approches basées sur les modèles de langage : Ces approches utilisent des modèles statistiques ou probabilistes pour modéliser les séquences de mots dans les articles d'actualités. Elles peuvent être efficaces pour capturer les relations entre les mots, mais peuvent également nécessiter des ressources importantes en termes de calcul et de données d'entraînement.
3. Approches basées sur l'apprentissage automatique : Les méthodes d'apprentissage automatique, telles que les SVM, les réseaux de neurones, les arbres de décision, etc., ont été largement utilisées pour la classification des actualités. Elles permettent d'apprendre à partir des données d'entraînement et de générer des modèles de classification.

## Méthodologie de recherche suivie par le papier

La méthodologie de recherche suivie par le papier comprend plusieurs étapes clés, qui sont les suivantes :

1. Collecte des données : Tout d'abord, les auteurs ont collecté un ensemble de données d'actualités sur lequel ils ont réalisé leurs expériences. La collecte de données peut impliquer la recherche de sources d'actualités en ligne, l'utilisation de bases de données spécifiques, ou même la création d'un ensemble de données annotées manuellement.
2. Prétraitement des données : Une fois les données collectées, les auteurs ont effectué un prétraitement des données afin de les préparer pour l'analyse. Cela peut inclure des étapes telles que la suppression des balises HTML, la normalisation du texte, la suppression des caractères spéciaux, la mise en minuscules, la suppression des mots vides (stop words), et la tokenisation pour diviser le texte en mots individuels.
3. Représentation des données : La technique de TF-IDF a été utilisée pour représenter les données textuelles. TF-IDF permet de quantifier l'importance d'un terme dans un document par rapport à l'ensemble du corpus. Les auteurs ont calculé les valeurs de TF-IDF pour chaque terme dans chaque document d'actualité, ce qui a permis d'obtenir une représentation numérique des caractéristiques des articles.

- 
4. Entraînement du modèle SVM : Les auteurs ont utilisé l'algorithme de machine à vecteurs de support (SVM) comme méthode de classification. Ils ont divisé l'ensemble de données en un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage a été utilisé pour entraîner le modèle SVM en fournissant les vecteurs de caractéristiques TF-IDF des articles et leurs étiquettes de classification correspondantes. L'ensemble de test a été utilisé pour évaluer les performances du modèle une fois qu'il a été entraîné.
  5. Évaluation des performances : Les performances du modèle ont été évaluées à l'aide de différentes mesures de performance telles que la précision, le rappel, la mesure F1 et éventuellement la courbe ROC (Receiver Operating Characteristic). Ces mesures permettent de quantifier la qualité de la classification réalisée par le modèle et d'évaluer son efficacité par rapport aux étiquettes de classification réelles.
  6. Analyse et interprétation des résultats : Enfin, les auteurs ont analysé et interprété les résultats obtenus. Ils ont pu identifier les points forts et les limites de leur approche proposée, comparer leurs résultats avec ceux des travaux précédents, et discuter des implications et des possibilités d'amélioration de leur méthode.

## Techniques utilisées

Techniques utilisées dans le papier :

1. Technique de TF-IDF (Term Frequency-Inverse Document Frequency) : La méthode TF-IDF est utilisée pour représenter les données textuelles. Elle calcule la fréquence du terme dans un document (TF) et l'inverse de la fréquence dans le corpus (IDF). Le TF mesure l'importance du terme dans un document spécifique, tandis que l'IDF mesure l'importance globale du terme dans le corpus. En multipliant le TF par l'IDF, on obtient une valeur qui reflète l'importance relative d'un terme dans un document par rapport à l'ensemble du corpus. Ainsi, TF-IDF permet de représenter les articles d'actualités sous forme de vecteurs de caractéristiques pondérées.
2. Machine à vecteurs de support (SVM) : Les auteurs utilisent la machine à vecteurs de support comme algorithme de classification pour résoudre la problématique de classification des actualités. SVM est un modèle d'apprentissage supervisé qui utilise des vecteurs pour représenter les données d'entrée. Il cherche à trouver une frontière de décision optimale, appelée hyperplan, pour séparer les différentes classes de manière maximale. L'algorithme SVM est capable de gérer des problèmes de classification linéaire et non linéaire en utilisant des fonctions noyau (kernel) appropriées.

Dans le papier, les auteurs entraînent un modèle SVM sur les données d'apprentissage, où les vecteurs de caractéristiques TF-IDF sont utilisés comme entrée. Le modèle apprend à classer les articles d'actualités en fonction de ces caractéristiques. Une fois que le modèle est entraîné, il est capable de prédire la classe des nouveaux articles d'actualités en utilisant les mêmes caractéristiques TF-IDF.

3. Évaluation des performances : Les performances du modèle sont évaluées à l'aide de mesures telles que la précision, le rappel et la mesure F1. La précision mesure la proportion d'articles d'actualités correctement classés par rapport au nombre total d'articles classés dans une classe spécifique. Le rappel mesure la proportion d'articles d'actualités correctement classés par rapport au nombre total d'articles réels dans une classe spécifique. La mesure F1 est une mesure combinée de la précision et du rappel, qui permet d'obtenir une évaluation globale de la performance du modèle. Ces mesures aident à quantifier la qualité de la classification réalisée par le modèle et permettent de comparer différentes approches ou paramètres du modèle.

## Implementation

L'implémentation des deux parties demande sont disponible sur le même dossier pret a etre compiler le premier sous le nom "data\_bbc" et le deuxième sous le nom de "data\_news\_20"

Notant que pour la première le score est de `Accuracy: 0.959731543624161`

Et la deuxième `Accuracy: 0.946969696969697`

## Conclusion

En conclusion, le papier présente une approche innovante pour la classification des actualités en utilisant la technique de TF-IDF et l'algorithme de machine à vecteurs de support (SVM). La méthodologie de recherche suivie a permis aux auteurs de collecter des données d'actualités, de les prétraiter, de les représenter avec TF-IDF, d'entraîner le modèle SVM et d'évaluer ses performances.

Les résultats obtenus ont montré que l'approche proposée était efficace pour la classification des actualités. Les mesures de performance telles que la précision, le rappel et la mesure F1 ont démontré la capacité du modèle SVM à classifier avec précision les articles d'actualités dans différentes catégories. Cette approche a permis d'organiser et de filtrer efficacement une grande quantité d'informations, ce qui est essentiel dans le domaine de l'analyse textuelle.