Une nouvelle approche de text mining basée sur TF-IDF et les machines à vecteurs de support pour la classification des nouvelles

03.06.2024

Abdessamad El ouatiki

1) Définition de la problématique

Dans un contexte où les fournisseurs de nouvelles partagent leurs titres sur divers sites web et blogs, la problématique principale de cette étude est de classifier les nouvelles en différents groupes afin que les utilisateurs puissent identifier le groupe de nouvelles le plus populaire dans le pays souhaité à tout moment. L'objectif de ce papier est de proposer une méthode de classification des nouvelles basée sur la Fréquence de Terme - Fréquence Inverse de Document (TF-IDF) et les Machines à Vecteurs de Support (SVM) pour atteindre cet objectif.

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

Le travail s'appuie sur plusieurs méthodes de classification de nouvelles proposées dans la littérature, telles que la classification des nouvelles financières, la classification de textes courts, la classification automatique des titres de nouvelles, et des approches hybrides de classification de textes intégrant les K-plus proches voisins et les SVM. Ces études montrent que la classification automatique des nouvelles est plus efficace que la classification manuelle, réduisant significativement le temps nécessaire pour organiser les informations.

3) Méthodologie de recherche

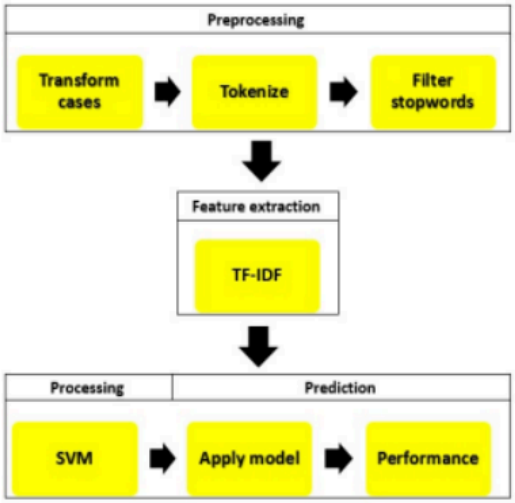

Figure 1.general architecture of the proposed method

L'approche de classification des textes de nouvelles se déroule en trois étapes : le prétraitement des textes, la sélection des caractéristiques basée sur TF-IDF, et la classification à l'aide des SVM.

1. Outils de conception : RapidMiner Studio Professional 6.5 a été choisi pour sa fiabilité, sa stabilité et sa vitesse de calcul pour l'analyse des données.

2. Prétraitement des Textes :

- Nettoyage des données pour éliminer les informations inutiles comme les ponctuations et les phrases non pertinentes.
- Transformation des caractères en minuscules pour uniformiser les mots.
- Tokenisation pour séparer les mots des phrases et éliminer les ponctuations.
- Filtrage des mots vides (stopwords) pour supprimer les mots fréquents et non significatifs.

3. Extraction des caractéristiques : Utilisation de l'algorithme TF-IDF pour calculer le poids des mots en fonction de leur fréquence et de leur importance dans les documents. TF-IDF est calculé comme suit :

$$w_{ij} = tf_{ij} * \log \frac{N}{df_i} \qquad (1)$$

Dans cette équation, (W_{ij}) est le poids du mot (i) dans le document (j), (N) est le nombre de documents dans l'ensemble des documents, et (df_i) est le nombre de documents contenant le mot (i).

4. Classification avec SVM : Utilisation du Support Vector Machine (SVM) avec un noyau RBF et une valeur maximale du paramètre nu pour séparer les échantillons positifs et négatifs, et classifier de nouveaux échantillons.

4) Techniques utilisées

Les principales techniques utilisées dans cette étude sont :

- **TF-IDF (Term Frequency-Inverse Document Frequency)** : Une technique de pondération utilisée pour évaluer l'importance d'un mot dans un document par rapport à un corpus.
- **SVM (Support Vector Machine)** : Un algorithme de classification supervisée utilisé pour analyser les données et reconnaître les motifs, en les classant dans les différentes catégories définies.

5) Comparaison des résultats

Les résultats obtenues:

```
Accuracy: 0.9730337078651685
Classification Report:
                precision    recall  f1-score   support

      business       0.94      0.97      0.96       115
entertainment       0.99      0.99      0.99        72
      politics       0.97      0.96      0.97        76
         sport       1.00      0.99      1.00       102
          tech       0.97      0.95      0.96        80

      accuracy                           0.97       445
     macro avg       0.97      0.97      0.97       445
  weighted avg       0.97      0.97      0.97       445
```

Les résultats du papier:

TableII- values of F obtained for the BBC datasets

Group name	**F-measure**
Business	0.9670
Politics	0.9732
Entertainment	0.9857
Sport	0.9922
Tech	0.9736

6) Critiques du travail

Bien que les résultats soient prometteurs, certaines critiques peuvent être formulées :

- **Dépendance aux jeux de données** : Les résultats obtenus sont spécifiques aux jeux de données utilisés. Il serait intéressant de tester l'approche sur d'autres jeux de données pour évaluer sa généralisation.
- **Complexité computationnelle** : Les techniques TF-IDF et SVM peuvent être coûteuses en termes de calcul, surtout pour de très grands corpus de textes.
- **Prétraitement du texte** : La qualité du prétraitement influence fortement les résultats. Des erreurs dans cette phase peuvent entraîner des baisses de précision.

7) Conclusion

Ce papier présente une méthode efficace pour la classification des nouvelles en utilisant TF-IDF et SVM. Les résultats montrent des précisions élevées, ce qui démontre le potentiel de cette approche. Cependant, pour améliorer et généraliser les résultats, des tests supplémentaires sur différents jeux de données et une optimisation des étapes de prétraitement et de classification sont nécessaires. En dépit de ces limites, cette méthode offre une solution prometteuse pour la classification automatique des nouvelles, facilitant ainsi l'accès rapide à l'information pertinente pour les utilisateurs.