Conclusion

L'arabe est une langue fléchie et morphologiquement très riche qui présente de vrais défis à la classification automatique des documents. Nous avons traité dans cette thèse la question de la classification automatique des documents écrits en caractères arabes en couplant l'apprentissage artificiel avec un traitement automatique de la langue arabe basé sur l'analyseur morphologique de notre équipe de recherche qui fait appel à la riche base de données lexicale DIINAR.1

Nous avons d'abord traité le cas où ces documents ont un contenu monolingue. Dans ce contexte nous avons mené une étude comparative concernant le choix de la nature de l'attribut à adopter pour la représentation vectorielle des documents. Les expérimentations menées utilisent quatre corpus d'apprentissage conçus en utilisant les racines, les lemmes, les mots et les n-grammes, respectivement. Nous avons constaté ainsi que les racines représentent un choix très convenable puisqu'elles sont capables d'alléger significativement l'effet négatif du phénomène des « données éparses ». Elles réduisent la taille du corpus d'apprentissage, ce qui améliore la fidélité et fournit une stabilité de la performance du modèle de prédiction. Toutefois, cette stabilité n'a pas été remarquée lors des expérimentations menées sur des corpus d'apprentissage conçus à partir des n-grammes. Pour cela, nous concluons que l'approche basée sur l'analyse morphologique est la plus fiable pour le choix de la nature des attributs d'autant plus qu'elle profite des avancées dans les recherches dans le traitement automatique des langues. En effet, cela permettra d'améliorer davantage la précision de l'extraction des racines et de surmonter plus de difficultés rencontrées lors du traitement automatique de la langue arabe.

Ensuite nous nous sommes penchés sur la question de l'efficacité de l'algorithme du dopage (Boosting) avec la classification automatique des documents écrits en caractères arabes tout en demeurant dans un cadre monolingue. Nous avons remarqué qu'aucun des travaux menés sur des documents écrits en caractères arabes n'utilise cet algorithme malgré le fait qu'il a été conçu, spécifiquement, pour la classification automatique des documents. Pour cela, nous avons mené une étude comparative entre les arbres de décision dopés et d'autres algorithmes d'apprentissage comme les arbres

de décision (sans dopage), les machines à vecteurs de support (SVM), et les réseaux bayésiens naïfs (NBM). Le dopage a réussi à améliorer significativement la fidélité des arbres de décision C4.5. Toutefois, l'algorithme C4.5 dopé n'a pas pu surpasser la performance et l'exactitude des algorithmes SVM et NBM. Nous attribuons cette faiblesse au fait que les arbres de décision sont très sensibles au moindre changement de leurs données sous-jacentes qui, lors du dopage, sont pondérées et modifiées régulièrement à chaque itération selon les résultats intermédiaires obtenus par le classifieur précédent. Cependant, le dopage reste aussi efficace que jamais.

Nous avons également traité le cas des documents écrits en caractères arabes ayant un contenu multilingue. Quelle que soit la raison pour laquelle les mots écrits en caractères latins existent dans ces documents, nous croyons qu'ils méritent d'être pris en compte comme les mots écrits en caractères arabes. Malheureusement, les pratiques courantes des chercheurs éliminent tous ces mots sans aucune considération quant à leur pertinence ou signification. Malgré leur faible fréquence, nous trouvons que cette pratique élimine des éléments souvent très pertinents au sujet du texte et pouvant contribuer énormément à la subjectivité de la prédiction puisque la majorité de ces mots ont une haute conformité par rapport aux catégories associées à leurs documents. Pour cela, il fallait éliminer uniquement les mots portant peu d'information ou ayant peu de pertinence comme, par exemple, les mots vides puisqu'ils affectent négativement la fidélité et l'efficacité du modèle de prédiction construit. Dans ce paradigme, nous avons aussi remarqué qu'un sous-ensemble important de méthodes de réduction de la taille du vocabulaire connues et largement utilisées dans la littérature (le gain d'information, la mesure de χ^2, le rapport de gain, l'information mutuelle, le rapport des chances et la fréquence des documents) étaient toutes incapables de conserver un nombre significatif des mots écrits en caractères latins sauf si le seuil des nombres d'attributs à conserver est assez large. Cela remet en question l'idée d'utiliser ces méthodes puisque le but est de réduire la taille du corpus tout en conservant les attributs les plus fidèles à la classification et non pas de l'augmenter pour les trouver. Ainsi, nous avons proposé les solutions suivantes qui sont capables de conserver un nombre significatif de ces mots tout en assurant une fidélité de classification satisfaisante :

1. Nous avons proposé une nouvelle méthode de sélection d'attributs, qu'on appelle « 3C » (Coefficient Catégorique Cumulatif), qui vise à mesurer la fréquence d'un attribut dans l'ensemble des catégories ainsi qu'une variante de cette méthode ayant une stratégie de sélection composée qu'on appelle la méthode « 3C-SC ».

2. De plus, nous avons proposé une nouvelle méthode de sélection d'attributs appellée « RFDC » (Rapport de la Fréquence de Documents avec Conformité) qui se base sur la méthode de « 3C » pour calculer les scores des attributs écrits en caractères arabes et qui utilise une autre nouvelle méthode qui pallie la faible fréquence des mots écrits en caractères latins en introduisant dans le calcul de leurs scores la mesure de « conformité » calculée par la méthode de ICF. Pareillement à la méthode « 3C-SC », la méthode « RFDC » adopte une sélection d'attributs composée.

3. Nous avons, en outre, trouvé qu'en modifiant la stratégie de sélection de deux méthodes connues et largement utilisées dans la littérature pour la réduction de la taille du vocabulaire nous avons obtenu une fidélité de classification meilleure que celle obtenue par leur stratégie originale : au lieu de procéder par une sélection séquentielle d'attributs, comme le font tous les chercheurs du domaine, nous appliquons une sélection composée. Autrement dit, après avoir calculé les scores des attributs et avoir trié les résultats par ordre décroissant de score, l'approche choisit d'abord séquentiellement les premiers n mots écrits en caractères arabes et puis revisite la liste établie au départ pour choisir les premiers m mots écrits en caractères latins.

Pour vérifier la validité des nouvelles solutions nous avons mené une large batterie de testes en utilisant plus que 99 corpus d'apprentissage. Nous avons comparé nos solutions avec six méthodes connues. Ainsi, nous avons constaté le suivant :

- Nos solutions sont toujours capables de conserver un nombre significatif de mots écrits en caractères latins supérieur à celui retenu par les méthodes connues.

- Les algorithmes d’apprentissage basés sur les corpus engendrés par nos solutions ont toujours été plus fidèles à la classification automatique que ceux basés sur les corpus engendrés par les méthodes connues.

Cette thèse met en œuvre une approche qui combine le recours à une analyse morphologique puissante basée sur une ressource lexicale très complète (DIINAR.1) et des démarches statistiques fondées sur des techniques de fouille de données récentes et largement connues et utilisées dans le littérature du domaine. Nous espérons avoir illustré dans ce travail l’efficacité de cette double approche dans le cas des documents écrits en caractères arabes, y compris lorsque ces derniers comportent des mots écrits en caractères latins.