كلية العلوم و التقنيات فاس
ⵜⴰⵖⵉⵡⴰⵏⵜ ⵏ ⵜⵎⴰⵙⵙⴰⵏⵉⵏ ⴷ ⵜⵙⵏⵉⴳⵉⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴰⴷ ⴱⵏ ⵄⴱⴷⵓⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

N : 37 / 2019

THÈSE DE DOCTORAT

Présentée Par

Mohammed BEKKALI

Directeur de thèse
Prof. Abdelmonaime LACHKAR

Discipline : Sciences de l'Ingénieur

Spécialité : Informatique

Contributions à l'Amélioration de la Représentation de Texte Court en Langue Arabe : Analyse du contenu des Réseaux Sociaux

---

Soutenue le 12 / 07 / 2019, Devant le jury :

Nom Prénom	Titre	Etablissement	
Abderrahmane SBIHI	PES	Ecole Nationale des Sciences Appliquées de Tanger	Président
Abderrahim ELQADI	PES	Ecole Supérieure de Technologie de Salé	Rapporteur
Hassan BADIR	PES	Ecole Nationale des Sciences Appliquées de Tanger	Rapporteur
Abderrahim BENABBOU	PH	Faculté des Sciences et Techniques de Fès	Rapporteur
Noureddine CHENFOUR	PES	Faculté des Sciences Dhar El Mehraz de Fès	Examinateur
Noureddine RAIS	PES	Faculté des Sciences Dhar El Mehraz de Fès	Examinateur
Abdelhamid EL JIHAD	PH	Institut des études et de Recherches pour l'Arabisation Rabat	Examinateur
Abdelmonaime LACHKAR	PES	Ecole Nationale des Sciences Appliquées de Tanger	Directeur de thèse

Laboratoire d'Accueil : Ingénierie Systèmes et Applications
Etablissement : Ecole Nationale des Sciences Appliquées de Fès

بسم الله الرحمن الرحيم

﴿رَبِّ اشْرَحْ لِي صَدْرِي* وَيَسِّرْ لِي أَمْرِي* وَاحْلُلْ عُقْدَةً مِّن لِّسَانِي * يَفْقَهُوا قَوْلِي﴾

صدق الله العظيم

﴿سورة طه، الآيات: 25-28﴾

Dédicaces

Louange à ALLAH de m'avoir donné la puissance pour accomplir ce travail de recherche. C'est avec sa grâce que ledit travail a été accompli.

Je dédie cette thèse à :

« A mes parents »,

Qui n'ont cessé d'investir pour que je puisse atteindre ce niveau,

Mme Fatima, ma chère mère, le plus grand cœur que j'ai rencontré, le rayon de soleil qui ne cesse de m'éclairer le chemin dans ma vie, sans son affection envers moi, ses oraisons pour moi, et surtout sans son sourire extraordinaire, je n'aurais pas pu frayer mon chemin et continuer à savourer le goût de la réussite et du succès.

M. Mohammed, mon cher père, l'homme le plus généreux du monde, qui n'a jamais tardé à m'encourager, et de me soutenir ; ses aides et ses recommandations m'ont souvent incité à persévérer dans l'effort et à progresser dans ma vie universitaire et professionnelle.

« A ma Sœur et mon Frère »,

Hanane et Brahim, qui n'ont cessé de m'encourager et de me soutenir durant toute cette période d'étude.

Aux Membres de ma famille : Abdessalam Ben Ktib, Safae Sadouk les petits Maroua, Malak et Mehdi.

A toute ma famille, A tous mes amis

Merci…

Mohammed BEKKALI

Remerciements

Le travail de cette thèse est le produit de plusieurs années de recherche durant lesquelles de nombreuses personnes ont marqué leurs empruntes clairement dans les traces réalisées pendant cette période et que je tiens à les remercier.

Je tiens tout d'abord à remercier mon directeur de thèse, Monsieur Abdelmonaime LACHKAR, professeur à l'Ecole Nationale des Sciences Appliquées de Tanger (Ex-Professeur à ENSA-USMBA, Fès), qui a bien voulu m'aider et m'encadrer tout au long de la réalisation de ce travail. Je tiens à le remercier vivement pour ses conseils fructueux qu'il m'a prodigués, pour son orientation ciblée, pour sa disponibilité, même à des heures peu adaptées, et pour la confiance qu'il a bien voulu m'accorder.

Merci également aux membres de jury, pour avoir accepté de juger ce travail auquel ils ont accordé un intérêt particulier.

Je remercie toutes les personnes qui ont participé de manière directe ou indirecte à la concrétisation de ce travail. Qu'ils trouvent ici l'expression de ma reconnaissance.

Publications de l'Auteur

Publications

[1] **Mohammed Bekkali**, Abdelmonaime Lachkar "An effective short text conceptualization based on new short text similarity", Social Network Analysis and Mining 9:1, 2019, DOI : 10.1007/s13278-018-0544-8

[2] Mohammed Rais, **Mohammed Bekkali**, Abdelmonaime Lachkar "An Efficient Method for Biomedical Word Sense Disambiguation Based On Web-Kernel Similarity", International Journal of Healthcare Information Systems and Informatics, (Accepté)

[3] **Mohammed Bekkali**, Abdelmonaime Lachkar "Web Search Engine-Based Representation for Arabic Tweets Categorization", M. Kaya et al. (eds.), From Social Data Mining and Analysis to Prediction and Community Detection, Lecture Notes in Social Networks, 2017, DOI : 10.1007/978-3-319-51367-6_4

Conférences

[1] **Mohammed Bekkali**, Abdelmonaime Lachkar "Arabic Sentiment Analysis based on Topic Modeling", The Second International Conference of The Moroccan Classification Society (SMC'2019), Kénitra, Morocco, 2019

[2] **Mohammed Bekkali**, Abdelmonaime Lachkar "Context-based Arabic Word Sense Disambiguation using Short Text Similarity Measure", 12th International Conference on Intelligent Systems: Theories and Applications (SITA'2018), Rabat, Morocco, 2018

[3] Elfaik Hanane**, Mohammed Bekkali**, Habibi Brahim, Abdelmonaime Lachkar "Arabic Paraphrasing Recognition Based Kernel Function for Measuring the Similarity of Pairs", The 3rd International Conference on Advanced Information Technology, Services And Systems (AIT2S' 2018), Mohammedia, Morocco, 2018

[4] ISSAM, Sahmoudi, **Mohammed Bekkali**, Abdelmonaime Lachkar “Rough Set Theory based a Linguistic Approach for Knowledge reduction in Formal Context”, The 6th International Conference on Arabic Language Processing (ICALP 2017), Fès, Morocco, 2017

[5] **Mohammed Bekkali**, Abdelmonaime Lachkar “An Efficient Hybrid Feature Selection Method based on Rough Set Theory for Short Text Representation”, The 17th International Arab Conference on Information Technology (ACIT'2016), Beni-Mellal, Morocco, 2016

[6] **Mohammed Bekkali**, ISSAM Sahmoudi, Abdelmonaime Lachkar “Enriching Arabic Tweets Representation based on Web Search Engine and the Rough Set Theory”, The IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 2015

[7] **Mohammed Bekkali**, Abdelmonaime Lachkar “Rough Set Theory for an Effective Arabic Tweet Representation”, The Second National Doctoral Symposium On Arabic Language Engineering (JDILA'2015), Fez, Morocco, 2015

[8] **Mohammed Bekkali**, Abdelmonaime Lachkar “Arabic Tweets Categorization Based on Rough Set Theory”, The Third International Conference on Data Mining & Knowledge Management Process (CDKP’2014), Dubai, UAE, 2014

الملخص

أدى الاستخدام الهائل للأجهزة المحمولة، مثل الهواتف الذكية واللوحات الالكترونية، إلى زيادة كبيرة في البيانات التي ينشرها مستخدمي مواقع التواصل الاجتماعي، حيث يتفاعل اليوم حوالي مليار مستخدم يوميًا على الشبكات الاجتماعية عبر الإنترنت، يشاركون المعلومات ويناقشون مجموعة متنوعة من الموضوعات؛ غير ان هذه الشبكات الاجتماعية تسمح للمستخدمين في الغالب بنشر نص قصير، 280 حرفًا كأقصى حد على سبيل المثال في أحد المواقع. تختلف النصوص القصيرة عن النصوص التقليدية باختصارها وعدم توفر معلومات سياقية مما يزيد من صعوبة استخلاص معلومات عالية الجودة من النص القصير؛ علاوة على أن اللغة العربية تزيد من حدة هذه المشاكل باعتبارها واحدة من اللغات الاكثر تعقيدا. ونتيجة لذلك، فإن تقنيات تمثيل النص التقليدي ستكون غير كافية وقد تؤثر سلبًا على أداء اي تطبيق من تطبيقات التنقيب في النصوص. وقصد التغلب على مشاكل النص القصير، تميل أغلبية المقترحات إلى جلب المزيد من المعلومات الدلالية واثراء المحتوى الأصلي من مجموعة البيانات نفسها أو من قاعدة معلومات خارجية.

الهدف من هذه الاطروحة هو تحليل كفاءة إثراء تمثيل النص القصير باللغة العربية، وخصوصا على اثنين من التطبيقات الهامة جدا، نخص بالذكر تصنيف النص القصير وتحليل المشاعر (المعروف أيضا باسم التنقيب في الآراء). بداية قمنا باستخدام نظرية رياضية لإثراء وتعزيز الروابط الدلالية بين النصوص القصيرة في نفس مجموعة البيانات؛ بعد ذلك، قمنا بجلب مزيد من المعلومات الدلالية من الويب من أجل تحديد سياق النص القصير واثرائه. من ناحية أخرى، توجهنا الى محاولة إثراء النص القصير في اللغة العربية باستخدام تمثيل مرتكز على المفاهيم بدلاً من التمثيل المرتكز على المصطلحات من خلال دمج موارد خارجية. وفي هذا الصدد اقترحنا: مساهمة اولى للتغلب على مشكلة تحديد المفهوم اللغوي من خلال قياس تشابه النص القصير، ومساهمة ثانية للتغلب على مشكلة تغطية اللغة العربية في الموارد الخارجية من خلال استعمال التصاميم المبنية على الموضوعات. وأخيرًا، من أجل اختبار مقترحاتنا والتحقق منها، أجرينا سلسلة من التجارب، حيث توضح النتائج المحصل عليها اهمية مساهماتنا.

Résumé

L’utilisation massive des appareils mobiles, tels que les smartphones et les tablettes, a permis une augmentation considérable de la génération de données par les utilisateurs. De nos jours, environ un milliard d'utilisateurs interagissent quotidiennement sur les réseaux sociaux en ligne, où ils partagent des informations et discutent d'une grande variété de sujets. Ces réseaux sociaux permettent aux utilisateurs de publier généralement un texte court, 280 caractères par exemple dans le cas de Twitter. Les textes courts se distinguent des documents traditionnels par leurs brièveté et le manque d’informations contextuelles, en outre la langue arabe complique encore plus ces problèmes vu la richesse de sa morphologie. Par conséquent, les techniques traditionnelles de représentation de texte seront inadéquates et peuvent avoir un impact négatif sur la performance de toute application de fouille de texte. Pour surmonter les problèmes liés au texte court, la majorité des propositions tendent à apporter plus de sémantique au contenu initial à partir de la collection de données elle-même ou bien provenant d’une base de connaissance externe.

Le but de cette thèse est d'analyser l'efficacité de l'enrichissement de la représentation de texte court en langue arabe, en particulier sur deux applications parmi les plus répondues, la catégorisation de texte cout et l’analyse des sentiments (dite aussi fouille d’opinions). Nous avons exploité la théorie des ensembles approximatifs (*Rough Set Theory - RST*) afin d’enrichir et de renforcer les liens sémantique entre les textes courts d’une même collection de données ; par la suite nous avons apporté plus de sémantique à partir du Web dans le but de cerner et d’identifier le contexte d’un texte court. Par la suite, nous avons mis l’accent sur l’enrichissement du texte court en la langue arabe en utilisant une représentation conceptuelle au lieu de celle basée sur les termes en intégrant des bases de connaissances externes. Dans ce sens nous avons proposé, une première contribution pour pallier au problème de désambiguïsions lexicale à travers une mesure de similarité dédiée pour le texte court, une deuxième contribution pour surmonter le problème de taux de couverture de la langue arabe dans les bases de connaissances à travers la modélisation par thèmes (*Topic Modeling*).

Finalement, pour tester et valider nos propositions, nous avons mené une série d’expérimentations, les résultats obtenus illustrent l’intérêt de nos contributions.

Mots clés : Langue Arabe ; Texte Court ; Réseaux Sociaux ; Enrichissement ; Représentation basé Termes ; Représentation basé Concepts ; Mesure de Similarité.

Abstract

With the increase use of mobile devices, such as smartphones and tablets, short text messages, tweets, comments and so on, have become a large portion of the online text data. Today, about one billion users interact daily on online social networks, where they share information and discuss a wide variety of topics. These social networks allow users to post short text, 280 characters for example in the case of Twitter. These short texts are different from the traditional documents in their shortness and sparsity. Moreover, the Arabic language makes these problems even more serious due to its rich morphology. As a result, traditional text representation techniques will be inadequate and may affect negatively the performance of any text mining application. To overcome these problems in short text, two major approaches have been exploited using additional semantics. The expansion information could be from the data collection internally based on Natural Language Processing techniques, or derived from external resources.

The aim of this thesis is to analyze the efficiency of the enrichment representation of short text in Arabic language using two well-known applications: short text categorization and sentiment analysis (also called opinion mining). We have used Rough Set Theory (RST) to enrich and strengthen the semantic links between short texts in the same data collection; in another contribution, we enrich short text representation by leveraging Web search results to provide greater context for the original content of short texts. Otherwise, we have emphasized the short text enrichment in the Arabic language using a conceptual representation instead of the term-based one by integrating an external resource. We have proposed a first contribution to overcome the problem of Word Sens Disambiguation through an Arabic short text similarity measure. Another contribution was proposed to tackle the problem of coverage ratio of the Arabic language in the external semantic knowledge through Topic Modeling.

Finally, to test the effectiveness of our proposition, a series of experiments has been conducted and the obtained results illustrate the interest of our contributions.

Keywords : Arabic Language; Short Text ; Social Networks ; Enrichment ; Term-based Representation ; Concept-based Representation ; Similarity Measure .

Liste des Tableaux

Tableau 1 . Les Consonnes Arabes [72] 33
Tableau 2 . Les Voyelles Arabes [72] 33
Tableau 3 . Caractères qui ne s'attachent pas au suivant [72] 33
Tableau 4 . Différents sens du mot (قلب) [114] 36
Tableau 5 . Catégorie lexicale du mot (ذهب) [114] 37
Tableau 6 . L'encodage de la langue arabe (Unicode vs. CP-1256) [114] 38
Tableau 7 . Exemple d'étiquettes grammaticales attribuées selon la voyellation [114] 41
Tableau 8 . Exemples de lemmatisation pour la langue Arabe 46
Tableau 9 . Exemple d'un Système d'Information 57
Tableau 10 . Exemple d'un Système de Décision 58
Tableau 11 . Espace vectoriel avant d'appliquer le modèle de tolérance 62
Tableau 12 . Espace vectoriel après avoir appliqué le modèle de tolérance 63
Tableau 13 . Exemples des tweets collectés 68
Tableau 14 . Description du corpus 69
Tableau 15 . Exemples des tweets après l'enrichissement basé sur le Web 77
Tableau 16 . Les corpus enrichis générés 79
Tableau 17 . Description du corpus 79
Tableau 18 . Description des deux corpus 84
Tableau 19 . Exemples des mesures de similarité obtenus pour le texte court 95
Tableau 20 . Description des trois corpus 100
Tableau 21 . F1-mesure pour la conceptualisation de texte court 102
Tableau 22 . Comparaison de la couverture de la langue Arabe et Anglaise dans WordNet et BabelNet 106
Tableau 23 . Description du corpus utilisé pour l'analyse des sentiments 116
Tableau 24 . F1-mesure pour l'analyse des sentiments en langue arabe 117

Liste des Figures

Figure 1 . Structure de Thèse 4
Figure 2 . Le problème de l'encodage de la langue arabe [114] 38
Figure 3 . Un exemple de prétraitement (Root / Stem) [114] 44
Figure 4 . Exemple d'un résultat d'analyse de mots «هذا » (hada) par Al-Khalil [81] 49
Figure 5 . La visualisation de l'approximation inférieure et supérieure d'un ensemble X 56
Figure 6 . Méthode Proposée d'Enrichissement de texte court basé sur RST 65
Figure 7 . Exemple illustratif de l'enrichissement 65
Figure 8 . Description des expérimentations de l'enrichissement en utilisant RST 67
Figure 9 . Précision de l'enrichissement basé RST 70
Figure 10 . Rappel de l'enrichissement basé RST 70
Figure 11 . F1-mesure de l'enrichissement basé RST 70
Figure 12 . Méthode d'Enrichissement de texte court basé sur le corpus Web 74
Figure 13 . Exemple d'un Snippet en langue arabe 75
Figure 14 . Description des expérimentations de l'enrichissement sur le Web 78
Figure 15 . Précision de l'enrichissement basé sur le Web 81
Figure 16 . Rappel de l'enrichissement basé sur le Web 81
Figure 17 . F1-mesure de l'enrichissement basé sur le Web 82
Figure 18 . Description des expérimentations de l'enrichissement basé sur le Web comparé avec l'enrichissement basé sur RST 83
Figure 19 . F1-mesure de l'enrichissement basé sur le Web comparé avec l'enrichissement basé sur RST pour le premier corpus 85
Figure 20 . F1-mesure de l'enrichissement basé sur le Web comparé avec l'enrichissement basé sur RST pour le deuxième corpus 86
Figure 21 . La méthode Web-RST pour calculer le score de similarité entre deux textes courts 91
Figure 22 . Méthode proposé pour la conceptualisation de texte court en arabe basée sur la mesure Web-RST 97
Figure 23 . Description des expérimentations de la conceptualisation de texte court en arabe 99
Figure 24 . Schéma Décrivant LDA 107
Figure 25 . Représentation de LDA sous forme de modèle graphique 108
Figure 26 . Fonction de densité de loi de Dirichlet à 2 dimensions (k = 3 topics) sur le triangle 109
Figure 27 . Méthode de conceptualisation de texte court pour l'analyse des sentiments en langue arabe 113
Figure 28 . Description des expérimentations de l'analyse des sentiments en langue arabe 115
Figure 29 . F1-mesure avec des valeurs différentes de Topics (5, 10, 15 et 20) 118

Table des matières

Introduction Générale ... 1
1 Réseaux Sociaux et Forums : Une Nouvelle Source d'Information ... 1
2 Objectifs, Motivations et Défis ... 2
3 Contributions ... 3
4 Plan de Thèse ... 4
Partie 1 : Contexte et Etat de l'Art ... 6
Introduction de la partie 1 ... 7
CHAPITRE 1 ... 8
Fouille de Texte Court ... 8
1 Introduction ... 9
2 Fouille de Texte ... 9
2.1 Prétraitement et Représentation des Documents ... 10
2.1.1 Prétraitement des Documents ... 10
2.1.2 Représentation des Documents ... 11
2.2 Problèmes Liés au Fouille de Texte Court ... 17
2.3 Applications de Texte Court ... 18
2.3.1 L'Extension de Requêtes ... 19
2.3.2 Analyse du Contenu des Réseaux Sociaux ... 20
2.3.3 Classification des SMS ... 20
2.3.4 L'évaluation des Produits ... 21
2.3.5 L'Analyse des Sentiments ... 21
3 Catégorisation de Texte Court ... 22
4 Enrichissement de Texte Court ... 24
5 Mesure de Similarité pour le Texte Court ... 26
6 Conclusion ... 29
CHAPITRE 2 ... 30
Traitement Automatique de la Langue Arabe ... 30
1 Introduction ... 31
2 Description et Caractéristiques de la Langue Arabe ... 31
3 Particularités de la Langue Arabe ... 33
3.1 Alphabet Arabe ... 33
3.2 Structure d'un Mot Arabe ... 34
3.3 Morphologie Arabe ... 35
3.4 Catégories des Mots Arabe ... 35
3.5 Grammaire et Caractéristiques de la Langue Arabe ... 35
4 Difficultés de l'Analyse Automatique de la Langue Arabe ... 36

4.1 Complexités du Traitement Automatique de la Langue Arabe 36
4.2 Exemples de Complexité de la Langue Arabe 36
4.3 Problème d'Encodage 37
4.4 Analyse Morphologique 38
4.5 Segmentation de Texte Arabe 40
4.6 Etiquetage Grammatical 40
4.7 Analyse Syntaxique 41
5 Techniques de Prétraitement 42
5.1 Tokenisation 42
5.2 Elimination des Mots Vides (Stop-Words) 43
5.3 Racinisation (Stemming) 44
5.4 Lemmatisation 45
5.5 Approches et Outils d'Analyse Morphologique 46
5.5.1 Approches d'Analyse Morphologique 46
5.5.2 Analyseurs Morphologiques 47
5.6 Approches d'Etiquetage Grammatical 49
6 Conclusion 50
Partie 2 : Enrichissement de Texte Court basé sur les Termes (Term-based) 51
Introduction de la partie 2 52
CHAPITRE 3 54
Enrichissement de texte court basé sur la Théorie des Ensembles Approximatifs - RST – 54
1 Introduction 55
2 Théorie des Ensembles Approximatifs 55
2.1 Approximation d'un Concept 55
2.1.1 Relation d'indiscernabilité 55
2.1.2 Fonction d'Appartenance 56
2.1.3 Le Système d'Information 56
2.2 Le Modèle de Tolérance des Ensembles Approximatifs (Rough Set Tolerance Model - RSTM) 59
2.2.1 L'Espace de Tolérance des Termes (Tolerance Space) 59
2.2.2 Classe de Tolérance d'un Terme (Tolerance Class) 59
2.2.3 Enrichissement de la Représentation d'un Document 60
2.2.4 Schéma de Pondération Etendu pour l'Approximation Supérieure 60
3 Méthode Proposée pour l'Enrichissement basé sur RST 64
4 Expérimentations et Résultats 66
4.1 Description des Expérimentations 66
4.2 Corpus Utilisé 67
4.3 Résultats & Discussion 69

5 Conclusion 71
CHAPITRE 4 72
Enrichissement de Texte Court basé sur le Corpus WEB 72
1 Introduction 73
2 Méthode Proposée pour l'Enrichissement basé sur le Web 73
3 Expérimentations et Résultats 1 77
3.1 Description des Expérimentations 77
3.2 Corpus Utilisé 79
3.3 Résultats & Discussion 80
4 Expérimentations et Résultats 2 83
4.1 Description des Expérimentations 83
4.3 Résultats & Discussion 85
5 Conclusion 86
Partie 3 : Enrichissement de Texte Court basé sur les Concepts (Concept-based) 87
Introduction de la partie 3 88
CHAPITRE 5 90
Web-RST Mesure de Similarité dédiée au Texte Court 90
1 Introduction 91
2 Web-RST pour Mesurer la Similarité de Textes Court 91
3 Evaluation Subjective de Web-RST 94
4 Intégration de la Mesure WEB-RST pour Résoudre le Problème d'Ambiguïté Lexicale 96
5 Expérimentations et Résultats 98
5.1 Description des Expérimentations 98
5.2 Corpus Utilisés 99
5.3 Résultats & Discussion 101
6 Conclusion 103
CHAPITRE 6 104
Représentation Conceptuelle de Texte Court basé sur LDA 104
1 Introduction 105
2 Problème de Couverture de la Langue Arabe dans les Ressources Sémantiques 105
3 L'Allocation Dirichlet Latente 106
3.1 Définition 106
3.2 Le Modèle 107
3.3 Processus de Génération 108
3.4 Loi de Dirichlet 108
3.5 Probabilité Jointe et Loi Marginale 109
3.6 Inférence et Estimation de Paramètres 110

3.7 Inférence 110
4 Intégration de LDA dans la Conceptualisation de Texte Court 112
5 Expérimentations et Résultats 114
5.1 Description des Expérimentations 114
5.2 Corpus utilisé 116
5.3 Résultats & Discussion 116
6 Conclusion 118
Conclusion Générale et Perspectives 119
Bibliographie 121

Introduction Générale

1 Réseaux Sociaux et Forums : Une Nouvelle Source d'Information

De nos jours, l'utilisation large des appareils mobiles modifie la nature et le traitement de l'information. Premièrement, les réseaux sociaux ont élargi leurs communautés en fournissant une énorme quantité de nouvelles informations. Deuxièmement, grâce aux appareils mobiles qui permettent aux utilisateurs de publier du contenu de partout et à tout moment, les utilisateurs sont poussés à produire un contenu plus court et beaucoup plus fragmenté par rapport au passé. Par exemple, la plateforme Twitter permet aux utilisateurs d'écrire et de partager des textes courts d'une longueur limitée à 140 caractères avant de la doubler à 280 dernièrement[1]. Cette restriction, combinée à une activité d'écriture rapide très fréquente et souvent avec des abréviations omniprésentes et de nouveaux acronymes inventés par les utilisateurs, ouvre de nouveaux défis dans le domaine du traitement de texte. En effet, le contenu des réseaux sociaux peut être analysé à plusieurs objectifs : l'identification des tendances, la catégorisation et le filtrage des nouvelles, etc. D'autres chercheurs tentent à classer les textes courts publiés sur les réseaux sociaux (par exemple, les tweets), en utilisant un contenu tiré du Web, pour comprendre les intérêts des utilisateurs, créer des modèles d'utilisateur, etc.

L'explosion de l'utilisation des réseaux sociaux a créé de nouvelle applications telles que l'analyse des sentiments (dite aussi la fouille des opinions). Ce procédé apparait au début des années 2000 et connait un succès grandissant dû à l'utilisation de données provenant de réseaux sociaux, notamment celles fournies par Twitter. Le but de l'analyse est de déterminer si le sentiment dégagé par une phrase est positif ou négatif par rapport à un objet matériel ou immatériel. Par exemple, l'analyse des sentiments a joué un rôle très important dans les élections présidentielles des états unis par *Barak Obama* ; Selon *Attensity*[2], *Obama* avait principalement un sentiment positif, avec «Go Obama» comme sujet d'actualité[3] durant la période des élections.

Il est clair qu'il existe une nouvelle source importante d'informations potentiellement intéressantes à exploiter. Par conséquent, il est de la plus haute importance de concevoir et de

[1] https://www.01net.com/actualites/twitter-double-la-taille-maximale-des-messages-qui-passent-de-140-a-280-caracteres-1297976.html
[2] C'est une entreprise américaine permettant de fournir des analyses sociales
[3] https://contently.com/2012/10/24/social-media-sentiment-becomes-factor-in-presidential-campaigns/

mettre en œuvre des outils efficace pour obtenir les bonnes informations à partir de ces données brutes collectées. De plus, il peut être utile d'envisager une analyse post-traitement afin d'en déduire des connaissances supplémentaires susceptibles d'améliorer les résultats.

2 Objectifs, Motivations et Défis

Récemment et avec l'explosion de l'utilisation des réseaux sociaux avec plus de 2 milliards d'utilisateurs actifs à travers le monde, ces réseaux sociaux sont des outils idéals pour les utilisateurs qui cherchent à échanger, partager et communiquer dans un seul endroit. Ils permettent aux utilisateurs de publier du contenu avec une taille limitée. Les textes courts se distinguent des documents traditionnels par leur brièveté et le manque d'informations contextuels. Par conséquent, ils ont tendance à être de plus en plus ambigus et les techniques traditionnelles de représentation de texte sont inadéquates et influencent négativement sur la performance de n'importe quelle application de fouille de texte. Cependant, notre étude des différentes méthodes permettant de comprendre et traiter le texte court a pu dégager deux problématiques auxquels nous nous sommes intéressées dans le cadre de ce travail de thèse :

La brièveté du contenu de texte court : Pour faire face au problème de brièveté et le manque d'informations contextuelles dans le texte court, la plupart des solutions proposées visent à apporter une sémantique supplémentaire dans le but d'enrichir la représentation de textes courts. Cette dernière peut provenir de la collection de données elle-même ou à partir d'une base de connaissances externe beaucoup plus large comme Wikipédia[4], WordNet [106] ou bien BabelNet [107]. La première approche requiert des techniques de traitement de langage naturel peu profondes, tandis que la seconde requiert des bases de connaissances beaucoup plus volumineux et appropriés. Pour surmonter ce problème, nous avons proposé deux contributions dans les deux sens (c'est-à-dire un enrichissement à partir de la collection elle-même et un autre en utilisant des ressources externes).

La mesure de similarité pour une paire de texte court : La majorité des mesures de similarité entre deux documents qui ont été proposées utilisent les mots partagés entre eux (par exemple la mesure Overlap [59]). Cette solution peut influencer d'une façon catastrophique par rapport au texte court ; soient les deux textes courts suivants ***« Inb Sina / 'ابن سينا' »*** et ***« Prince des médecins / 'أمير الأطباء' »***, si leurs dégrée de similarité est calculé à base des mots partagés, la similarité sera 0. Sauf que, les deux textes représentent la même personne, alors le dégrée de

[4] https://www.wikipedia.org/

similarité doit théoriquement être très élevé. De ce fait, il est d'une grande nécessité d'adapter les mesures de similarité dédiées au texte court.

En effet, mesurer la similarité de textes courts est devenu une étape cruciale dans le processus de nombreuses tâches. Par conséquent, elle a été étudiée d'une manière intensive durant la dernière décennie [62 - 69]. Pour surmonter ce problème, plusieurs solutions tentent à étendre la représentation de texte court avec du contenu supplémentaire provenant d'une ressource externe dans le but d'engendrer le contexte de texte et par la suite calculer la similarité entre les deux textes étendus. Ces solutions concernent uniquement les langues latins, par contre aucun travail n'a été réaliser dans le cadre de la langue arabe. Nous allons étudier comment utiliser le Web comme étant un corpus très large et ouvert pour améliorer le calcul de similarité entre deux texte courts en langue arabe.

Dans le cadre de cette thèse, nous avons mis en œuvre des systèmes qui sont les plus utilisés dans les applications de textes court :

- La catégorisation de texte court
- L'analyse des sentiments (dite aussi fouille d'opinions)

3 Contributions

Dans l'objectif d'améliorer et d'enrichir la représentation de texte court en langue arabe, nous avons procédé par deux types d'enrichissement :

Un enrichissement basé sur les termes (Term-based)

La nouvelle information ajoutée sera des termes liés sémantiquement au contenu initial, deux contributions ont été proposées dans le cadre de la présente thèse :

- Enrichir la représentation des tweets en arabe en se basant sur la théorie des ensembles approximatifs (Rough Set Theory - RST) [92], qui est un outil mathématique permettant de traiter les données imprécises et vagues.
- Enrichir la représentation des tweets en arabe en se basant sur le corpus Web comme étant une ressource d'informations large et dynamique.

Un enrichissement basé sur les concepts (Concept-based)

La nouvelle information ajoutée sera des concepts extraits à partir d'une base de connaissances externe, deux contributions ont été proposées dans le cadre de la présente thèse :

- Une représentation conceptuelle de texte court en arabe en se basant sur BabelNet [107] comme une ressource externe. Durant la phase de conceptualisation, nous avons proposé une mesure de similarité pour le texte court qui a été intégré dans le processus de désambiguïsation lexicale (Word Sense Disambiguation - WSD).
- Nous avons utilisé l'Analyse Latente de Dirichlet (Latent Dirichlet Analysis - LDA) [64] afin de surmonter le problème de taux de couverture de la langue arabe dans les ressources de connaissance externes.

4 Plan de Thèse

Le présent rapport est composé de trois parties comme illustré sur la figure 1, précédées par une introduction générale, et suivies d'une conclusion générale et des perspectives.

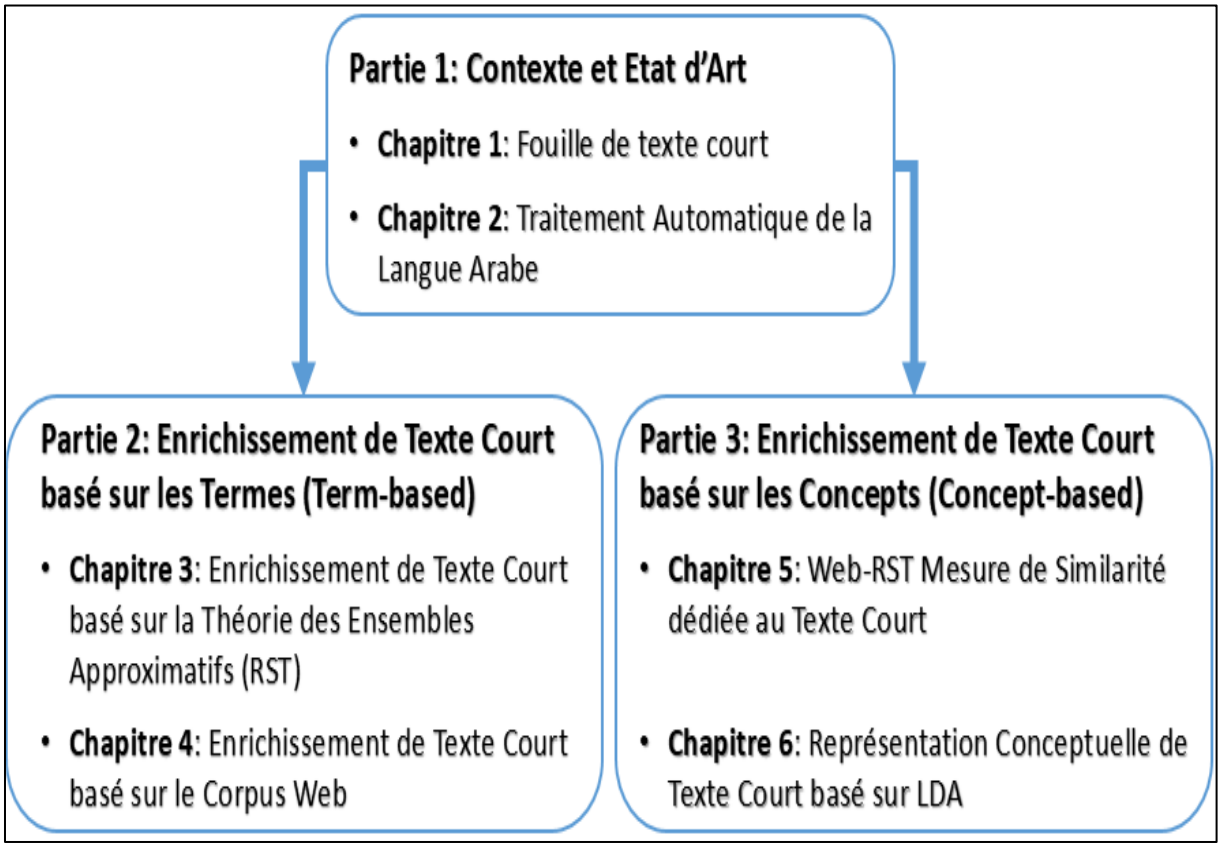


Figure 1 . Structure de Thèse

La première partie intitulée Contexte et Etat de l'Art : dans laquelle nous présentons le contexte générale et les travaux connexes qui ont été proposé dans la littérature et qui sont liés à notre travail. Elle est composée de deux chapitres 1 et 2 :

- Chapitre 1 : Fouille de texte court, présente un tour d'horizon du domaine de fouille de texte (*Text Mining*), il décrit tout d'abord les concepts de base de ce domaine. Puis, il introduit les problèmes liés à la fouille de texte court. Par la suite, il présente l'état d'art des travaux qui ont été proposé pour la catégorisation de texte court, l'enrichissement de texte court et finalement les mesures de similarité pour le texte court.
- Chapitre 2 : présente les différentes caractéristiques, particularité et problèmes qui sont associés au Traitement Automatique des Langues Naturelles (TALN) et particulièrement la langue arabe.

La deuxième partie intitulée Enrichissement de texte court basé sur les termes : elle regroupe les différentes contributions que nous avons proposées pour enrichir la représentation de texte court en langue arabe en se basant sur les termes. Elle est composée de deux chapitres 3 et 4:

- Chapitre 3 : Enrichissement de texte court basé sur la Théorie des Ensembles Approximatifs, présente les concepts de base de la RST et comment nous avons l'utilisé afin d'enrichir la représentation de tweets en langue arabe.
- Chapitre 4 : Enrichissement de texte court basé sur le Web, décrit comment le corpus Web comme étant une ressource d'informations large et dynamique peut être exploité pour améliorer la représentation des tweets en langue arabe.

La troisième partie intitulée Enrichissement de texte court basé sur les concepts : elle présente les différentes contributions que nous avons proposées pour enrichir la représentation de texte court en langue arabe en se basant sur les concepts. Elle est composée de deux chapitres 5 et 6 :

- Chapitre 5 : Web-RST Mesure de Similarité dédiée au Texte Court, présente une mesure de similarité pour le texte court en langue arabe. Cette mesure sera utilisée dans le processus de conceptualisation de texte court afin de surmonter le problème de désambiguïsation lexicale (WSD).
- Chapitre 6 : Représentation Conceptuelle de Texte Court basé sur LDA, décrit comment nous avons incorporé la modélisation par thèmes (Topic Modeling) pour pallier au problème de taux de couverture de la langue arabe dans les ressources externes.

Partie 1 : Contexte et Etat de l'Art

- **Chapitre 1 : Fouille de Texte Court**

- **Chapitre 2 : Traitement Automatique de la Langue Arabe**

Introduction de la partie 1

La fouille de texte (Text Mining) ou la Découverte de Connaissances à partir de Texte (Knowledge Discovery from Text (KDT)), mentionné pour la première fois dans Feldman et Dagan [1] comme étant l'application des algorithmes, des méthodes d'apprentissage automatique et des statistiques dans le but de faciliter l'extraction et la compréhension d'informations à partir des collections de documents.

Cette partie présente les concepts de base de la fouille de texte et décrit les processus du prétraitement et représentation des données textuelles ; et vu que la brièveté et le manque d'informations contextuelles dans le texte court, les modèles de représentation des documents peuvent être inadéquates et influencer négativement sur la performance de n'importe quel système de fouille de texte court. De ce fait, cette partie introduit aussi les problèmes liés au traitement du texte court. Par la suite, elle fournit une investigation sur les travaux de recherche liés à : la catégorisation de textes courts, les différentes approches permettant d'enrichir la représentation de texte court, et les mesures de similarité pour le texte court.

Il faut noter que la nature de la langue traitée par un système de fouille de texte reste un élément décisif sur les résultats obtenus ; et puisque la langue arabe se caractérise comme étant l'une des langues les plus riches morphologiquement, la présente partie donne aussi une idée claire et globale sur les différents points concernant le Traitement Automatique de Langue Naturel en traitant le cas de la langue arabe. Les différentes caractéristiques, particularité et problèmes de cette langue ont été étudiés dans le but d'améliorer, avec les résultats présentés dans cette thèse, la représentation de texte court en langue Arabe.

CHAPITRE 1
Fouille de Texte Court

1 Introduction 9
2 Fouille de Texte 9
2.1 Prétraitement et Représentation des Documents 10
2.1.1 Prétraitement des Documents 10
2.1.2 Représentation des Documents 11
2.2 Problèmes Liés au Fouille de Texte Court 17
2.3 Applications de Texte Court 18
2.3.1 L'Extension de Requêtes 19
2.3.2 Analyse du Contenu des Réseaux Sociaux 20
2.3.3 Classification des SMS 20
2.3.4 L'évaluation des Produits 21
2.3.5 L'Analyse des Sentiments 21
3 Catégorisation de Texte Court 22
4 Enrichissement de Texte Court 24
5 Mesure de Similarité pour le Texte Court 26
6 Conclusion 29

1 Introduction

Ce chapitre présente un tour d'horizon du domaine de fouille de texte. Il décrit les processus du prétraitement et représentation des données textuelles. Puis, il introduit les problèmes liés au traitement du texte court ainsi que ses applications. Par la suite, il fournit un une investigation sur les travaux de recherche liés à :

- la catégorisation de textes courts
- les différentes approches permettant d'enrichir la représentation de texte court
- les mesures de similarité pour le texte court

2 Fouille de Texte

Dès l'invention des ordinateurs, les hommes sont à la recherche d'une manière efficace de gérer, stocker, diffuser et rechercher l'information. D'où est née l'idée d'utiliser des machines pour automatiser la recherche d'information. Cela fut notamment popularisé en 1945 par Vanner Bush dans son célèbre article « *As We May Think* » [2]. Plusieurs méthodes et techniques de gestion et de traitement d'information ont été développées. A nos jours, on peut estimer que le niveau d'informatisation des données manipulées dans nos différentes organisations est évolué grâce au développement et à la maîtrise de la technologie (soit celle des matériaux, soit celle de la communication, soit celle de la construction des logiciels ou de la gestion et du traitement de l'information) dont l'Internet est un exemple flagrant. Malgré cette évolution, la progression des moyens de recherche efficaces est encore insuffisante, ce qui a conduit à révéler le domaine du fouille de texte (*Text Mining*) au grand jour notamment par le biais des moteurs de recherche connu sous le nom des Systèmes de Recherche d'Information (SRI) [3] [5].

Le *Text Mining*, mentionné pour la première fois dans *Feldman et Dagan* [1], est l'ensemble des technologies et méthodes destinées au traitement automatique de données textuelles en vue d'en dégager et structurer le contenu, les thèmes dans une perspective d'analyse rapide de découverte d'informations cachées ou de prise automatique de décision. C'est un procédé consistant à synthétiser (classer, structurer, résumer, …) les textes en analysant les relations, les patterns, et les règles entre les unités textuelles (mots, groupes, phrases, documents).

Les applications du *Text Mining* ont pour objectif de simplifier l'accès à l'information, son stockage et sa réutilisation. L'information peut être de différentes natures, les informations de type textuel viennent naturellement à l'esprit, mais il y a aussi la recherche ou l'extraction de

médias (images, vidéos, sons), ou d'informations au sein d'ensembles de données numériques, séquentielles ou graphiques. Afin d'atteindre cet objectif, des Systèmes d'Indexation et de Recherche d'Information doivent représenter, stocker et organiser l'information, puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête.

2.1 Prétraitement et Représentation des Documents

2.1.1 Prétraitement des Documents

Une collection de données textuelle se compose d'un volume considérable de documents collectés à partir de sources de données hétérogènes. Le prétraitement de ces données peut influencer positivement ou négativement sur la performance de n'importe quelle application de fouille de texte. D'où la nécessité de mettre en place une chaine de Traitement Automatique du Langage Naturel (TALN) appropriée à la langue et à l'application en question.

Le prétraitement des données peut également avoir pour objectif la réduction de dimension (élimination de l'information inutiles) et de temps de traitement, de sorte que les données issues de l'étape de prétraitement présentent une représentation logique des unités lexicales (tokens) les plus représentatifs. Le prétraitement des données textuelles se base sur les techniques de l'analyse morphologique.

L'analyse morphologique s'intéresse à la formation des mots au travers des processus de flexion (marques de genre, nombre, de conjugaison...) [114], dérivation (formation de vendeur à partir de vend) et composition. Étant donné un mot, il s'agit de déterminer quelles sont les unités minimales de sens qui le composent. Ces unités minimales de sens sont appelées morphèmes et se déclinent en termes de racine et d'affixes. Une analyse morphologique complète précise en plus la catégorie grammaticale de la racine et associe aux affixes des informations sémantiques et flexionnelles. Elle est réalisée soit en utilisant des bases lexicales existantes, soit à l'aide de véritables systèmes d'analyse, plus à même de traiter les formes non répertoriées. L'analyse morphologique est présentée dans de nombreuses applications du Traitement Automatique des Langues et elle permet de reconnaître la présence d'un même mot sous des formes de surface différentes. L'une des analyses morphologiques les plus simples mises en œuvre en recherche d'information est la Racinisation (*Stemming*). En extraction de connaissance, l'analyse morphologique la plus utilisée est la lemmatisation qui permet d'associer à une forme fléchie une forme conventionnelle ainsi que de calculer les traits flexionnels. Ainsi on peut trouver d'autres techniques pour effectuer l'analyse morphologique du texte comme la tokenisation, translitération, etc…

La Langue Arabe est l'une des langues les plus utilisées dans le monde, c'est une langue sémitique qui a une morphologie plus complexe que l'Anglais, elle est composée de trois types de mots : noms, verbes et particules. Les noms et les verbes sont dérivés d'un ensemble limité d'environ 10000 racines [4], et qu'en raison de cette morphologie complexe, il y a des étapes indispensables pour le prétraitement du texte arabe :

- La tokenisation une chaîne de mots en *tokens* ;
- La normalisation de ces *tokens* ;
- La suppression des mots vides (stops-words) ;
- L'application d'un algorithme de Racinisation ou Lemmatisation ;
- Le calcul de la pondération pour chaque mot dans le document.

Par la suite un modèle peut être utilisé pour représenter l'ensemble des documents de la collection. Les modèles de représentation les plus connus dans la littérature seront présentés dans la section suivante.

2.1.2 Représentation des Documents

La représentation des documents dans un espace vectoriel (Vector Space Model [108]) est très pratique car elle met les documents dans un espace euclidien où des mesures de similarité tels que le produit scalaire et la similarité Cosinus ou la distance euclidienne sont immédiatement disponibles. Les vecteurs de documents sont construits dans un espace de très haute dimension. De ce fait, il y a eu un intérêt considérable pour la représentation de documents dans des dimensions réduites pour surmonter les inconvénients de la représentation traditionnelle des documents en vecteur «sac de mots» [5].

Par ailleurs, il y a eu d'autres propositions pour la représentation de textes courts qui tendent à surmonter le problème de brièveté et aussi le manque d'information contextuelle.

a) Représentation basé sur les Termes

À chaque fois qu'il est question de définir un problème de façon à assurer un traitement automatique, il est impossible de passer outre l'étape où il faut choisir la façon dont on va représenter le problème. Dans le cas de la classification automatique de textes, on doit opter vers une façon efficace pour représenter les instances à traiter, soit les textes. Un grand nombre de chercheurs dans le domaine ont choisi d'utiliser une représentation vectorielle basé sur les termes dans laquelle chaque texte est représenté par un vecteur de *n* termes pondérés. À la base, les *n* termes sont tout simplement les *n* différents mots apparaissant dans les textes de

l'ensemble d'entraînement. Cette approche est aussi appelée «*Bag of Word*» («sac de mots»). Par la suite, il faut utiliser un système de pondération permettant d'associer un poids à chaque terme. Il peut être tout simplement binaire (1 si le mot est présent dans le texte, 0 dans le cas échant) ; comme aussi il peut représenter le nombre d'occurrences du mot dans le texte. Cependant, on peut donner une très grande importance aux termes qui apparaissent très souvent à travers toutes les classes et qui sont peu représentatifs d'une classe en particulier.

Le système de pondération le plus utilisé pour calculer le poids d'un terme est TF-IDF (de l'anglais ***Term Frequency - Inverse Document Frequency***) [113]. Issue du monde de la Recherche d'Information [6], ce système de pondération donne plus d'importance aux mots qui apparaissent souvent à l'intérieur d'un même texte, ce qui correspond bien à l'idée intuitive que ces mots sont les plus représentatifs du document. Par ailleurs, il donne également moins de poids aux mots qui appartiennent à plusieurs documents, pour refléter le fait que ces mots ont un faible pouvoir de discrimination entre les classes.

Le poids d'un terme t_k dans un document d_j est calculé ainsi :

$$TF\text{-}IDF\ (t_k, d_j) = f(t_k, d_j) * log(\ |T_r| / f(t_k)\ ) \quad (1)$$

où :

- $f\ (t_k, d_j)$ est le nombre d'occurrences de t_k dans d_j
- $|Tr|$ est le nombre de documents
- $f\ (t_k)$ est le nombre de documents dans lesquels t_k apparaît au moins une fois.

Si on désire avoir des poids entre 0 et 1, on peut appliquer une normalisation, ce qui est souvent le cas. La fonction TF-IDF a montré une bonne efficacité dans des tâches de catégorisation de textes, et, en plus, son calcul est simple. Il faut savoir que d'autres méthodes plus performantes peuvent aussi être utilisées.

Il est a noté que la représentation «*Bag of Word*» a été largement utilisée pour représenter les textes, un bon nombre de chercheurs ont expérimenté l'utilisation des modes de représentation alternatifs. Une des motivations est que cette approche met de côté une quantité considérable d'information contenue dans le document original. Les paragraphes, les phrases et l'ordre des mots sont complètement écartés. Toute l'information sur les structures syntaxiques est perdue. La question est bien sûr de savoir, si la qualité de la classification automatique des textes va être affectée négativement par cette façon d'agir ? Est-ce que tout ce qui est mis à l'écart au profit d'une représentation «*Bag of Word*» est véritablement pertinent lors de l'assignation d'un

texte à une catégorie ? La question à savoir aussi, est ce que les mots pris séparément représentent toujours des unités sémantiques atomiques n'est plus à débattre ? Il est clair que ce n'est pas le cas. Personne ne peut contredire le fait que les mots «apprentissage» et «automatique» ensemble ont une sémantique différente que lorsque ils sont utilisés séparément. On peut donc oser croire que des modes de représentation conservant l'information sur les liens entre les mots pourraient améliorer la classification de textes, sachant que cette tâche est directement liée au sens du message véhiculé par le texte, c'est-à-dire à sa sémantique.

b) Représentation basé sur les Concepts

La discussion précédente mis en évidence le compromis entre l'utilisation des mots séparément et la clarté sémantique. Les concepts - unités de connaissance - fournissent une solution qui équilibre ce compromis. Une autre méthode de représentation [7], bien que se base aussi sur le formalisme vectoriel pour représenter les documents, reste fondamentalement différente de la représentation « *Bag of Word* ». Les dimensions de l'espace vectoriel ne sont pas associées ici à des termes mais à des concepts.

Les concepts sont non ambigus : chaque concept représente un sens unique. Les synonymes sont représentés par le même concept, tandis que les termes ayant plusieurs significations sont représentés par différents concepts correspondant à leurs significations. Les concepts fournissent également une représentation compacte et efficace : le nombre de concepts dans les textes est généralement beaucoup plus petit que les mots.

Créer une représentation basée sur les concepts est plus compliqué que de créer une basée sur les mots, et constitue en soi un problème de recherche continu. En général, les concepts sont sélectionnés dans un système (ou une ressource) de concepts externe et attribués à un texte en fonction de son contenu. Chaque ressource a ses caractéristiques uniques en termes de type de concept, de structure et d'organisation de concepts, ainsi que le taux de couverture.

Les ressources externes (ou bien les bases de connaissances) les plus utilisées peuvent se résumer comme suit :

Wikipédia [5]***:***

Wikipédia est un projet d'encyclopédie collective en ligne, universelle, et multilingue. Ce projet, contenant plus que 20 million articles avec plus 282 langues, vise à offrir un contenu librement réutilisable, objectif et vérifiable, que chacun peut modifier et améliorer. Wikipédia

[5] https://www.wikipedia.org/

a prouvé son utilité afin d'enrichir la représentation des documents textuelles et par conséquent, l'amélioration de la performance de plusieurs applications dans le domaine de fouille de texte.

WordNet [6]***:***

WordNet est une grande base de données lexicale en anglais. Les noms, verbes, adjectifs et adverbes sont regroupés en ensembles de synonymes (synsets), chacun exprimant un concept distinct. Les synsets sont liés entre eux au moyen de relations conceptuelles sémantiques et lexicales. La structure de WordNet en fait reste un outil très utile pour améliorer le traitement automatique du langage naturel.

WordNet regroupe les mots en fonction de leur signification. Cependant, il existe des distinctions importantes. Premièrement, WordNet relie non seulement les formes de mots (d'est-à-dire des chaînes de lettres) mais principalement les sens spécifiques des mots. Par conséquent, les mots qui se trouvent à proximité les uns des autres dans le réseau sont sémantiquement désambiguïsés. Deuxièmement, WordNet marque les relations sémantiques entre les mots, alors que le regroupement de mots par exemple dans un thésaurus ne suit aucun modèle explicite autre que la similarité de sens.

BabelNet [7]***:***

BabelNet est le plus grand dictionnaire encyclopédique multilingue et réseau sémantique créé par l'intégration transparente de grandes encyclopédies Web multilingues, telle que Wikipédia et WordNet. Le résultat est un dictionnaire encyclopédique qui fournit des synsets (*Babel synsets*), c'est-à-dire des concepts et des entités nommées lexicalisés dans de nombreuses langues et liés à un grand nombre de relations sémantiques.

Open Directory Project (ODP)[8]***:***

ODP est un annuaire multi-langue, composé de quelques millions de pages Web pré-classées et organisées en arborescence. Pour la langue arabe, l'ODP inclut 4781 snippets pré-classés en 459 catégories par un groupe d'experts.

[6] https://wordnet.princeton.edu/
[7] https://babelnet.org/
[8] http://odp.org/

DBpedia*[9]*:

DBpedia est un outil permettant d'extraire un contenu structuré à partir des informations créées dans divers projets de Wikipédia. Ces informations structurées ressemblent à un graphe de connaissances ouvert (*Open Knowledge Graph*), accessible à tous sur le Web. Un graphe de connaissances est un type particulier de base de données qui stocke les connaissances sous une forme lisible par une machine et constitue un moyen de collecter, d'organiser, de partager, de rechercher et d'utiliser des informations.

Les Moteurs de Recherche Web:

Les moteurs de recherche Web comme Google et Bing sont des applications Web permettant de trouver des ressources à partir d'une requête sous forme de mots. Les ressources peuvent être des pages web, des articles de forums, des images, des vidéos, des fichiers, etc. Durant les dernières années ces moteurs ont prouvé leur utilité comme étant une source d'informations large et dynamique permettant d'apporter de grandes améliorations pour les applications de fouille de texte.

c) Représentation Mixte

L'avantage de la représentation conceptuelle est en particulier, de réduire les effets synonymiques du vocabulaire. Par exemple, «pic», «cime», «sommet», «crête» possèdent des sens en commun. Lors d'une représentation vectorielle, chaque mot sera associé à une dimension. Il n'y aura donc aucune ressemblance entre des textes utilisant ces différents mots. L'avantage de la représentation conceptuelle est que des mots synonymes partagent au moins un concept.

Cependant, l'inconvénient majeur de la représentation conceptuelle est que les noms propres du document ne sont pas pris en compte. En effet les noms propres, étant sémantiquement vides par définition, ne possèdent pas de représentation au sein du thésaurus. Par exemple, le mot «Ferrari» est défini comme un vecteur «nul» alors qu'il peut être utile lors d'un processus de catégorisation, notamment pour une catégorie de type : «Automobile». L'idée de la représentation mixte [7] est donc d'allier, à une représentation conceptuelle pure, une dimension statistique supplémentaire. Cette double représentation des textes a pour avantage de fournir deux informations différentes et complémentaires à un processus de catégorisation. En effet, la représentation statistique permet de mettre en évidence le vocabulaire discriminant tandis que

[9] https://wiki.dbpedia.org/

la représentation conceptuelle permet quant à elle, d'obtenir une vision plus globale du texte en projetant ce dernier sur un ensemble de concepts. Cette projection permet d'en déduire les «champs sémantique» du texte en question.

Chaque document dans le processus de catégorisation mixte sera représenté par un vecteur défini comme la concaténation des deux vecteurs suivants :

1. La première moitié du vecteur mixte correspondra au vecteur statistique.
2. La seconde moitié du vecteur mixte correspondra au vecteur conceptuel.

d) Type des Documents

Documents sur Internet

Sur Internet, la structure est présentée de deux manières : les liens entre les pages en définissent une première (la toile) et la structure interne des pages une autre. Depuis les débuts du développement d'Internet, il a paru intéressant d'intégrer une partie des informations que cette structure pouvait apporter. Google a été le premier moteur de recherche connu à faire une utilisation des liens ; le succès de ces techniques, même s'il commence à être biaisé aujourd'hui car les concepteurs de sites sont sensibles aux indices utilisés, montrent l'intérêt d'utiliser des informations supplémentaires à l'information de nature purement textuelle [8].

Les compétitions TREC ont abordé depuis 1999 [9] la problématique de la recherche sur le Web. Les différentes tâches traitées vont de la recherche d'une page qui a autorité sur un sujet à la recherche simple. Les principales conclusions des différents travaux font que l'utilisation de nouvelles sources d'informations doit être faite avec discernement. Les indices ou statistiques qui ont apporté mieux sont de façon non exhaustive :

- l'utilisation du nombre de liens entrant et sortant de la page ;
- l'utilisation du texte des liens qui renvoient vers la page ;
- l'utilisation de certains éléments de la structure, comme par exemple le titre, les mots clé, etc.

C'est un domaine qui n'est que très peu jalonné. Le nombre d'indices qu'il est possible de prendre en compte est très supérieur à celui de la recherche «plate». Et des techniques très classiques comme Okapi ont des résultats qui sont tout à fait comparables aux autres systèmes.

Texte Libre, Structuré et Semi-Structuré

Selon Ying [10], les Systèmes d'Extraction d'Information de texte libre sont appliqués pour traiter des textes en langage naturel, c'est à dire, des textes grammaticaux. Le traitement syntaxique et l'étiquetage sémantique sont souvent nécessaires pour l'extraction du texte libre.

Le deuxième type possible de texte donné est le texte structuré. Un document structuré a un format particulier et est généralement une page de résultats en interrogeant la base de données, par exemple, le résultat de recherche d'un certain article d'EBay[10], est un tableau comprenant l'image, le titre, le prix, le temps restant, etc., de tous les articles retournés est renvoyé. Pour identifier des informations des textes structurés, le format du document, par exemple, les balises HTML, doit être apprises.

Le texte semi-structuré est un type de texte entre le texte structuré et le texte libre, qui n'est ni structuré ni strictement grammatical. Puisque les textes semi-structurés contiennent des phrases complètes, les systèmes de texte libre en s'appuyant sur l'analyse syntaxique ne pouvaient pas être adaptés.

2.2 Problèmes Liés au Fouille de Texte Court

Les textes courts se distinguent des documents traditionnels par leur brièveté et le manque d'information contextuel (***Shortness and Sparseness***). Ainsi, les techniques traditionnelles de représentation de texte sont inadéquates et peuvent avoir un impact négatif sur la précision et la performance de toute application de fouille de texte en langue arabe.

Pour surmonter les problèmes liés aux textes courts, diverses méthodes ont été proposées dans la littérature. L'une des méthodes les plus courantes consiste à intégrer des informations externes pour améliorer la représentation de données textuelles courtes. Ce type d'approches utilise des sources d'information externes pour enrichir la représentation de texte court. L'information supplémentaire pourrait provenir de la collection de données de textes courts elle-même ou d'une base de connaissances externe beaucoup très large telles que Wikipédia[11], WordNet [106] et BabelNet [107], etc. La première approche requiert des techniques de traitement de langage naturel peu profondes (NLP), tandis que la seconde requiert des ressources externes beaucoup plus volumineux et appropriés.

[10] https://www.ebay.com/
[11] https://www.wikipedia.org/

Spécifiquement, les textes courts seront enrichis avec des mots, des documents ou des concepts liés au contenu initial. Par exemple, lorsqu'on utilise WordNet pour l'enrichissement, on peut extraire des synonymes en tant que termes d'enrichissement et les ajouter à un texte court donné. Si on utilise des bases de connaissances générales telles que Wikipédia pour l'enrichissement, l'explication de certains mots clés de Wikipédia peut être ajoutée aux documents texte originaux. On peut également utiliser le texte court d'origine comme requête de recherche, puis créer un nouveau texte à partir des snippets renvoyés par les moteurs de recherche. En outre, des textes courts peuvent être projetés vers un espace sémantique, et générer de nouvelles représentations dans cet espace sémantique, combinant des informations source et d'autres externes [11] [12] [13]. Ces travaux ont montré que l'intégration d'informations externes est utile pour atténuer le problème de la dispersion (*vagueness*).

Cependant, pour enrichir un texte court, il faut prendre en considération que la représentation enrichie soit sémantiquement similaire au texte original. Étant donné que la plupart des méthodes d'extension sont basées sur des termes (***term-based methods***), c'est-à-dire l'extension d'un texte court en fonction de la similitude entre les vecteurs de termes du texte original et les vecteurs externes. Il est d'une importance vitale de rechercher les termes clés et de filtrer les termes bruit (c'est-à-dire les termes n'ayant aucun lien sémantique avec le contenu original). L'avantage de l'extension du texte pourrait être ignoré une fois que les termes bruit sont inclus, car la plupart des mots ne figurent qu'une fois dans chaque extrait de texte court.

2.3 Applications de Texte Court

L'explosion des applications Web interactives et de la communication mobile a amené les utilisateurs ordinaires à passer d'un consommateur de contenu à un créateur de contenu. La plupart des utilisateurs ordinaires écrivent de textes courts ou des messages en ligne via des appareils mobiles. De plus, le développement des réseaux sociaux a rapidement étendu les ressources de textes courts. La brève analyse textuelle est devenue un nouveau défi, non seulement pour les autorités, mais également pour les sociétés commerciales, qui cherchent à mieux comprendre cette vague d'informations.

Les textes standards (livres et actualités, par exemple) sont créés par des écrivains ou des journalistes professionnels. En revanche, la plupart des textes courts actuels sont rédigés par un nombre considérable d'utilisateurs d'Internet, plutôt que par des écrivains professionnels. À l'aide d'Internet ou d'un téléphone mobile, ces personnes créent de courts textes chaque fois qu'elles utilisent des applications Internet. Les données sont collectées par le fournisseur de

services Internet ou d'autres robots d'exploration pour améliorer l'expérience des utilisateurs. Les principales sources de textes courts peuvent se résumés de la façon suivante :

- les snippets de requêtes générés par les moteurs de recherche, tel que Google[12]
- les réseaux sociaux, tels que Twitter[13] et Facebook[14];
- les systèmes d'évaluation des produits, tels qu'Amazon[15];
- les messages courts (SMS) dans les téléphones portables ;
- les descriptions de données multimédias, tel que YouTube[16];

Les principales applications sur le texte court peuvent se résumées dans les sous-sections suivantes.

2.3.1 L'Extension de Requêtes

La recherche d'informations est le domaine qui étudie la manière de retrouver des informations dans un ensemble de documents. Son importance réside dans le fait qu'elle représente la collecte de documents, la compréhension et le traitement de la requête saisie par l'utilisateur, ainsi que la récupération et la présentation des documents pertinents. Sauf que, la compréhension et le traitement de la requête reste l'étape primordiale dans ce processus, elle peut influencer d'une façon positive ou négative sur les résultats retournés à l'utilisateur par la suite.

Une requête soumise par un internaute à un moteur de recherche peut être vue sous forme de texte court. La mesure de la similarité entre les documents et les requêtes a été largement étudiée dans la recherche d'information. Récemment, El Mahdaouy et al. [14] ont proposé une méthode afin d'augmenter les poids des termes liés sémantiquement aux termes de la requête d'origine pour la langue arabe en utilisant la similarité entre les mots incorporés (*Word Embedding Similarities*). Kuzi et al. [15] ont présenté une série de méthodes d'extension de requêtes en utilisant le modèle (*Word-2-Vec*). Les résultats obtenus montrent que les méthodes d'extension de requête proposées améliorent les modèles de recherche d'information. Dans un autre travail Mahgoub et al. [16] ont proposé une méthode d'extension de requête sémantique en utilisant une ontologie indépendante du domaine construite à partir de Wikipédia. Les expériences sont

[12] https://www.google.com
[13] https://www.twitter.com
[14] https://www.facebook.com
[15] https://www.amazon.com
[16] https://www.youtube.com

effectuées sur TREC arabe 2002. Les expérimentations ont montré de meilleurs résultats par rapport à la méthode de référence de correspondance aux mots clés.

2.3.2 Analyse du Contenu des Réseaux Sociaux

Les réseaux sociaux tels que, Twitter et Facebook ont adoptés dans leur première année une présentation du contenu dans un ordre chronologique. Ce format de présentation est utile à l'utilisateur car les dernières publications sont généralement plus intéressantes qu'un événement qui s'est produit il y a longtemps. Cependant, présenter les publications dans un ordre chronologique peut être une mauvaise solution pour l'utilisateur, surtout s'il a plusieurs abonnés.

Guo et al. [17] ont étudié le problème de fouille et de l'exploitation des corrélations entre textes. Pour cette raison, ils ont introduit une contribution permettant de lier les tweets aux actualités (news) ; en partant de l'hypothèse que deux tweets peuvent être considérés liés ou complémentaires s'ils partagent un hashtag, une entité nommée ou ils sont partagés au même temps. Yerva et al. [18] ont classifié les tweets pour identifier s'ils sont liés à une entreprise ou non en utilisant les profils d'entreprises générés d'une manière semi-automatique à partir d'une ressource Web externe. Yegin et al. [19] ont introduit une technique de classification des tweets basé sur Wikipédia comme une ressource externe. Les auteurs classent les tweets en les projetant sur leurs pages Wikipédia les plus similaires, par la suite, ils calculent la distance sémantique entre les tweets en fonction de la distance entre leurs pages Wikipédia les plus proches. Sriram et al. [20] ont amélioré la classification des tweets en exploitant un ensemble de fonctionnalités extraites du profil de l'auteur et aussi du contenu du tweet.

2.3.3 Classification des SMS

Le service de messages courts (SMS) est l'un des services de communication les plus utilisé dans lequel un message est envoyé d'une façon électronique. Avec la croissance d'utilisation de ce service, le problème courrier indésirable (Spam) est devenu de plus en plus courant. Par conséquent, le problème de classification des SMS devient un axe de recherche qui se développe jour après jour.

Choudhary and Kumar Jain [21] proposent une méthode pour détecter et filtrer les messages de Spam en utilisant les techniques de classification avec apprentissage automatique. Ils ont étudié en détail les caractéristiques des messages SPAM, puis ils ont trouvé dix caractéristiques qui permettent de filtrer efficacement les messages de courrier indésirable. Ezpeleta et al. [22] ont amélioré le filtrage des Spam en utilisant des techniques de reconnaissance de personnalité avec

d’autres d’analyses des sentiments. Jialin et al. [23] ont utilisé la modélisation par thèmes (Topic Modeling) pour le filtrage des Spam afin de surmonter le manque d’information contextuel dans le processus de catégorisation.

2.3.4 L’évaluation des Produits

Avec la croissance rapide du commerce électronique (*e-commerce*), les avis (opinions) concernant certains produits populaires sur le Web ont aussi augmenté rapidement. Ces avis sont très utiles et importants pour la prise de décision dans des applications de différents domaines ; et vu le grand nombre des opinions générés chaque jour, l'automatisation du processus d'exploration d'opinions est devenu une nécessité pour les entreprises.

Song et al. [24] ont étudié la polarité des avis (*reviews*) sur les produits chinois à travers un Framework de classification basé sur SVM (*Support Vector Machine*). Ils ont créé des groupes d’opinions à partir des données d’apprentissage, par la suite ils ont exploité la génération des paraphrases pour étendre la représentation de chaque opinion dans chaque groupe. Liu et al. [25] ont présenté une méthode d’extraction d’opinion basée sur les graphes. Ils modélisent l'extraction d'opinion comme un processus de co-classement (*co-ranking*), dans lequel de multiples relations hétérogènes sont incorporées dans un modèle unifié. Dave et al. [26] ont développé une méthode basée sur la recherche d'information pour l'extraction d'opinion et la classification sémantique des opinions sur des produits. Pang et al. [27] ont mené une analyse des opinions sur les avis des films.

2.3.5 L’Analyse des Sentiments

L’analyse des sentiments (dite aussi : fouille d’opinion) est une discipline qui consiste à extraire et analyser pour un texte donné, les opinions exprimées sur chacun de ses aspects. Une opinion est caractérisée par sa polarité (positive, négative ou neutre), son intensité (par exemple, « excellent » est plus fort que « bon ») et son émetteur, c'est-à-dire celui qui émet l’opinion. De plus, elle porte sur un objet qui peut être un produit, un service, un événement, une personne, une entreprise… etc.

L’analyse des sentiments a attiré un grand intérêt au cours de la dernière décennie, en raison du volume de publications sur de nombreux réseaux sociaux, ce qui fournit une source trop riche d’idées et d’opinions sur divers sujets. Par conséquent, l'analyse des sentiments du contenu des réseaux sociaux est devenue une nécessité pour la langue arabe et bien pour toutes les langues.

Récemment, Abdulaziz et al. [28] ont proposé une nouvelle méthode de classification des sentiments des tweets en langue arabe, en intégrant les réseaux de neurones convolutifs (*Convolutional Neural Networks*) et les réseaux de mémoire à court terme (*Long Short-Term Memory Networks LSTM*). Les résultats obtenus montrent l'amélioration de l'analyse des sentiments sur différents corpus. Tartir et Abdul-Nabi [29] ont introduisé une approche sémantique pour découvrir les attitudes des utilisateurs et les perspectives commerciales à partir des réseaux sociaux, pour la langue arabe standard et dialectal. Ils ont présenté également la première version de l'ontologie arabe des sentiments (ASO), qui contient différents mots utilisés pour l'expression des sentiments. Les résultats obtenus montrent l'intérêt de leur proposition. El-Masri et al. [30] ont présenté un outil d'analyse des sentiments des tweets en arabe en utilisant une combinaison de paramètres : le moment de publication des tweets, des méthodes de prétraitement telles que la racinisation et les retweets, les caractéristiques n-grammes, et les méthodes d'apprentissage automatique. Les utilisateurs peuvent sélectionner un sujet et définir les paramètres souhaités. Le système par la suite, détecte la polarité (négative, positive ou neutre) du sujet à partir des récents tweets associés et affiche les résultats. Al Sallab et al. [31] ont proposé d'intégrer l'apprentissage en profondeur (*Deep Learning*) dans le processus d'analyse des sentiments pour la langue arabe à travers quatre modèles de représentation. Abdul-Mageed et al. [32] ont présenté un système (SAMAR) d'analyse de subjectivité et de sentiment pour le contenu des réseaux sociaux en langue arabe comme l'une des langues les plus complexes sur le plan morphologique. L'avantage de cette étude est que les auteurs ont créé un corpus multi-genres (à partir de quatre types de réseaux sociaux) contenant du texte rédigé en arabe standard et en arabe dialectal.

3 Catégorisation de Texte Court

Les textes courts générés par les réseaux sociaux constituent une source d'information intéressante ; Par exemple, de nombreux chercheurs se concentrent sur l'objectif de déduire les centres d'intérêts des utilisateurs à partir de leurs publications sur ces plateformes. Pour atteindre cet objectif, ils étudient et développent souvent des systèmes capables de catégoriser des textes courts, en identifiant les sujets liés à leur contenu (par exemple, la politique, les voyages, la littérature, etc.).

Récemment, Kundu et al. [33] ont étudié l'utilité des textes courts générés dans les réseaux sociaux aux moments d'une catastrophe. Ils ont constaté que, une bonne catégorisation de ces publications pouvant donner lieu à une action qui peut s'avérer très utile pour définir les moyens

de secours aux personnes se trouvant dans des zones touchées par une catastrophe. Pour cette raison, ils ont proposé un modèle basé sur l'apprentissage en profondeur (*Deep Learning*) pour classer les tweets en différentes classes telles que le besoin et la disponibilité des ressources, et les activités des organisations non gouvernementales, etc. Lui et Li [34] ont proposé une solution pour la prédiction de genre à Weibo[17] (*le réseau social le plus utilisé dans la chine*) en utilisant des caractéristiques linguistiques et d'autres thématiques (*Topic Modeling*) extraites à partir des publications des utilisateurs, et en utilisant SVM comme un classificateur. Ils ont constaté que les hommes et les femmes ont un style d'écriture et un langage très différent qui peut améliorer la prédiction du genre dans les réseaux sociaux. Hucko et al. [35] ont visé à améliorer le processus de communication d'un enseignant avec ses étudiants dans l'objectif d'aider l'enseignant en regroupant les réponses similaires. Ce processus comprend deux étapes principales. Dans la première étape, ils détectent les réponses correctes fournies par la classification de texte. La deuxième étape consiste à regrouper les réponses incorrectes pour les présenter au professeur d'une manière bien organisée. Ils ont évalué leur méthode sur des données réelles en langue Slovaque collectées à partir d'un cours sur les principes de génie logiciel. Chen et al. [36] ont proposé une méthode de classification de texte court en combinant *Word 2 Vec* pour surmonter le problème d'information contextuelle introduit par le texte court d'une part, et TF-ICF (*Term Frequency-Inverse Corpus Frequency*) d'autre part pour mieux déterminer l'importance d'un terme dans le corpus. Des expérimentations sur des données collectées à partir des réseaux sociaux ont été menées. Les résultats obtenus ont montré l'intérêt de leur proposition. Sotthisopha and Vateekul [37] ont proposé un Framework de catégorisation rapide de texte court contenant trois améliorations. Deux d'entre elles sont proposées pour réduire les ressources et le temps de calcul en utilisant l'algorithme K-Means qui peut bien s'adapter à la taille de l'ensemble de données. La troisième est une étape supplémentaire de prétraitement qui vise à augmenter le taux de couverture du vocabulaire lors de l'incorporation de mots (*Word Embeddings*), en associant les termes non couverts avec les termes les plus similaires en faisant appel à la mesure de similarité de Jaro-Winkler [38].

Puisque le modèle de sac de mots (*Bag of Word – BoW*) n'est pas tout à fait approprié pour les textes courts (en raison de leur brièveté et le manque d'information contextuelle), dans [39], les auteurs ont adopté des techniques de réduction de dimensionnalité afin d'améliorer la performance de la catégorisation de texte de court. Au lieu de cela, les auteurs de [40] ont proposé un nouveau modèle de sac de concepts (*Bag of Concept – BoC*) pour la classification

[17] https://www.weibo.com

de textes courts. En effet, ils exploitent une base de connaissances taxonomiques pour associer un modèle de concept à chaque catégorie et un ensemble de concepts pertinents à chaque texte court. Ensuite, ces derniers sont classés avec la catégorie la plus similaire. Dans [41], la classification des tweets est réalisée au moyen d'un autre réseau social (YouTube). En effet, les classes attribuées aux vidéos publiées sur ce dernier servent comme des étiquettes pour les tweets liés à ces vidéos, fournissant ainsi un très grand nombre d'instances étiquetées pouvant être très utiles comme données d'apprentissage.

D'autres travaux exploitent l'extension des documents afin de résoudre les problèmes de manque d'informations dans les textes courts. Dans [42], les auteurs ont proposé une technique d'extension, appelée TIDE, qui peut être appliquée à plusieurs tâches de Machine Learning et de recherche d'information sur des textes courts (tels que la classification de texte, le regroupement, la désambiguïsation d'entités) sans faire appel à des bases de connaissance spécifiques ou de tâches heuristiques spécifiques pour l'extension. Dans [43], et partant de l'hypothèse que le texte court généralement traite un seul sujet, les auteurs envoient le contenu du texte court sous forme d'une requête et analysent les résultats pour enrichir et cerner le sujet traiter par le texte à partir des résultats retournés par un moteur de recherche. Dans [44], les auteurs ont enrichi des textes courts en utilisant l'incorporation de mots (*Word Embbadings*) afin d'améliorer la classification des textes courts. Ils ont utilisé un modèle permettant de construire un nouvel espace de mots incorporés en partant de l'hypothèse que des mots liés sémantiquement auront des vecteurs représentatifs très proches dans le nouvel espace.

4 Enrichissement de Texte Court

L'enrichissement de texte est une technique couramment utilisée pour mieux représenter et classer correctement des textes courts. L'idée est d'enrichir la représentation de texte court avec des informations supplémentaires liées sémantiquement au contenu initial (en interrogeant un moteur de recherche Web ou une autre ressource de connaissance externe) afin de surmonter les problèmes liés aux textes courts.

Récemment, Flisar et Podgorelec [45] ont proposé d'utiliser les outils d'annotation de DBpedia pour identifier les entités pertinentes dans le texte et enrichir les textes courts avec des concepts dérivés de ces entités et issus de l'ontologie DBpedia[18]. Les expérimentations ont été menées sur trois collections de textes courts diverses et en utilisant une variété de classificateurs, les

[18] http://fr.dbpedia.org/

résultats obtenus montrent l'intérêt de leur proposition. Yu et al. [46] ont fourni un moyen permettant de combiner les informations issues des bases de connaissance et les réseaux de neurones profonds (*Deep Neural Network*) pour l'analyse de texte court afin d'aider les machines à mieux comprendre le texte court. McDonald et al. [47] ont proposé une méthode efficace pour l'enrichissement des tweets avec des termes discriminants supplémentaires issus d'un corpus Web contemporain, afin de bénéficier de la relation temporelle entre les documents sources et les publications de Twitter. Ils ont appliqué leur méthode à la classification de réputation, qui vise à classer les tweets qui mentionnent une entité commerciale (par exemple, «performance financière» ou «produits et services»). Meng et al. [48] ont proposé une méthode pour enrichir des textes courts à l'aide de moteurs de recherche publics, en explorant des pages reliées et en obtenant une base de connaissance du texte court original. Tang et al. [49] ont proposé un Framework d'intégration des connaissances multilingue pour le regroupement de texte court. Une méthode d'enrichissement de la représentation par regroupement (CREST) est introduite dans [50] : CREST génère à partir des données d'apprentissage des groupes de topics (de thèmes) en utilisant des méthodes de regroupement. Par la suite, les données des topics sont utilisées pour enrichir la représentation de texte court. Dans [51], les auteurs ont présenté une méthode permettant de relier les publications de Twitter à des articles de presse connexes afin de créer des profils d'utilisateur représentant les activités individuelles de Twitter.

Certains chercheurs ont opté vers l'utilisation des ressources de connaissance externes pour surmonter les problèmes liées aux textes courts. Par exemple, Calisan et Sakar [52] ont proposé une méthode qui consiste à générer de nouvelles caractéristiques de texte court en utilisant une base de connaissance externe, Wikipédia, afin d'améliorer la performance de la classification de textes courts. Li et Zhao [53] ont utilisé Wikipédia comme une base de connaissance pour construire un graphe afin d'enrichir et extraire les mots-clés d'un texte court. La méthode proposée a été testé dans un algorithme de classement de texte (*Text Ranking*). DBPedia est utilisé dans [54] comme source d'enrichissement permettant aux auteurs de classer en temps réel les tweets qui sont liés d'une manière conceptuelle à un sujet donné, afin d'améliorer la précision de l'extraction d'information. Dans [55], les auteurs ont utilisé WordNet et SentiWordNet dans une nouvelle méthode non supervisée de classification de polarité pour les tweets. Plus en détail, ils effectuent des parcours aléatoires sur WordNet en obtenant des scores « *PageRank »* qui seront ensuite utilisés pour pondérer les valeurs des concepts (*synsets*). Les auteurs de [56] partent du constat que les êtres humains, lorsqu'ils interprètent de textes courts, prennent en considération des mots sémantiquement liés au contenu initial. Par conséquent, ils

introduisent un système basé sur le modèle (*Dirichlet Multinomial Mixture*) afin d'améliorer la modélisation des thèmes (*Topic Modeling*) pour le texte court en exploitant les relations sémantiques entre les mots à partir d'un corpus de très grande taille de documents externe. Hu et al. [57] ont proposé un nouveau Framework pour améliorer les performances du regroupement de textes courts en exploitant la sémantique interne du texte original et des concepts externes issus des ressources externes (Wikipédia et WordNet). Gabrilovich et Markovitch [58] ont construit un générateur de caractéristiques à partir des connaissances encyclopédiques (Wikipédia) afin d'améliorer la représentation des documents et par conséquent améliorer la performance de la catégorisation de texte. Le générateur ainsi proposé permet d'identifier les articles les plus pertinents pour chaque document et d'utilise les concepts correspondant à ces articles pour créer de nouvelles caractéristiques qui complètent les termes originaux.

5 Mesure de Similarité pour le Texte Court

La mesure de similarité de textes courts est devenue une étape cruciale dans le processus de nombreuses tâches en raison de leurs brièvetés et le manque d'information contextuelle. Par conséquent, elle a été étudiée d'une manière intensive durant la dernière décennie. Trois approches ont été utilisées dans la littérature pour mesurer la similarité entre deux textes :

1- Approche basée sur les Mots Partagés (*Surface Matching*)

Étant donné une paire de texte (x, y), l'idée est basée sur le nombre de mots qui apparaissent dans les deux textes. Cette technique repose sur l'hypothèse que les textes similaires partagent de mots identiques, ce qui n'est efficace que si les deux textes ont une longueur suffisante. Disons que X et Y sont les ensembles de mots dans les deux textes x et y. Les mesures de similarité présentées dans [59] sont listées comme suit :

- Coefficient de correspondance simple (*Simple Matching*) qui est le nombre de termes partagés. Ce coefficient ne prend pas en compte les tailles de X et Y. Il a été utilisé dans la recherche d'information et peut être défini comme suit :

$$Mat(x, y) = | X \cap Y | \quad (2)$$

- Coefficient de similarité Jaccard (*Jaccad Measure*) est une mesure statistique utilisée pour comparer la similarité et la diversité des ensembles d'échantillons. Le coefficient entre les ensembles d'échantillons est défini comme la taille de l'intersection divisée par la taille de l'union des ensembles d'échantillons :

$$Jacc(x, y) = | X \cap Y | / | X \cup Y | \quad (3)$$

- Le coefficient de Dé (*Dice Measure*) est une mesure de similarité liée à coefficient Jaccard. Il peut être défini comme :

$$Dice(x, y) = 2 * | X \cap Y | / (| X | + | Y |) \quad (4)$$

- Le coefficient de chevauchement (*Overlap Measure*) est une mesure de similarité qui calcule le chevauchement entre deux ensembles. Si l'ensemble X est un sous-ensemble de Y ou l'inverse. Il est défini comme :

$$Over(x, y) = | X \cap Y | / min (| X|, |Y|) \quad (5)$$

- Le Coefficient Cosinus est une mesure de similarité entre deux vecteurs de n dimensions en trouvant l'angle entre eux. Il est souvent utilisé pour comparer des documents dans la recherche d'information et le regroupement de texte. Il peut être défini comme :

$$Cos(X, Y) = | X \cap Y | / \sqrt{|X| * |Y|} \quad (6)$$

Ces techniques ne sont efficaces et adéquates que si on compare des textes comportant un nombre de mots suffisant. Avec des textes courts, ces techniques présenteraient des inconvénients évidents. Par exemple, deux phrases peuvent avoir un sens similaire mais elles sont composées de mots totalement différents ; ou même dans les cas où deux textes courts peuvent partager des termes, ils peuvent être utilisés dans des contextes différents. Il existe également plusieurs extensions possibles de ces méthodes. Par exemple, deux termes peuvent être équivalents s'ils sont synonymes selon certains thésaurus tels que WordNet; ou bien les ensembles X et Y peuvent être construits en utilisant les racines ou les stems au lieu des termes originaux.

2- Approche basée sur l'Historique des Requêtes (*Query log Methods*)

Les moteurs de recherche tels que Google[19] ou Bing[20] traitent des millions de requêtes de recherche par jour. L'historique des requêtes produit est devenu une excellente ressource pour mesurer la similarité entre des textes courts. Ces informations seront par la suite utilisées pour la génération de suggestions de requêtes de recherche, qui devient de plus en plus précise et utile [60]. Un exemple de substitution de requête est présenté dans [61] où Jones et al. ont introduit la notion de substitution de requête, c'est-à-dire la génération d'une nouvelle requête pour remplacer la requête de recherche initiale d'un utilisateur. Ils utilisent des modifications

[19] https://google.com
[20] http://bing.com

basées sur les substitutions habituelles apportées par les internautes à leurs requêtes. De cette manière, la nouvelle requête est fortement liée à la requête d'origine et devra contenir des termes fortement liés à tous les termes d'origine.

3- Approche basée sur les Corpus (*Corpus based Methods*)

Il existe deux types de méthodes populaires basées corpus. Le premier type suppose que la signification d'un mot peut être définie à partir de son utilisation (c'est-à-dire sa distribution dans le texte); L'analyse sémantique latente (LSA) [62], l'analyse sémantique probabiliste latente (PLSA) [63], et par la suite l'allocation de Dirichlet latent (LDA) [64] sont les méthodes les plus connues. Le second type basé sur les corpus utilise des connaissances externes telles que WordNet ou le corpus Web pour calculer la similarité. Cela ne convient pas aux langues qui ne disposent pas de ressources ou des bases de connaissance telles que WordNet.

Récemment, Adel et al. [65] ont proposé FUSE (*FUzzy Similarity mEasure*) une mesure de similarité pour le texte court basée sur un algorithme de la logique flou. Les auteurs ont testé et évalué en utilisant trois corpus et les résultats obtenues montrent l'intérêt de leur proposition. Nagoudi et Schwab [66] ont proposé un système basé sur les termes incorporés (*Word Embedding*) pour calculer la similarité entre deux textes courts. L'idée principale est d'exploiter les vecteurs en tant que représentations de mots dans un espace multidimensionnel afin de dégager les propriétés sémantiques et syntaxiques des mots. Un analyseur morphosyntaxique et un système de pondération TF-IDF sont appliqués pour améliorer l'identification des mots les plus descriptifs d'un texte court. Nakamura et al. [67] ont utilisé Wikipédia pour calculer la similarité entre les textes courts multilingues. Shirakawa et al. [68] ont proposé d'utiliser une analyse probabiliste pour calculer les poids des termes clés et par la suite trouver les entités associées à chaque terme clé dans Wikipédia. Les termes clés et les vecteurs des entités extraites seront utilisés par la suite dans une version étendue du théorème de Naïve Bayes.

Les méthodes de l'approche basée sur les corpus ont prouvé leur efficacité par rapport aux deux premières approches. Parmi les méthodes les plus connues de la littérature, on trouve celle introduite par Sahami et Heilman [69] pour mesurer la similarité entre des textes courts en exploitant les résultats retournés par un moteur de recherche Web pour fournir un contexte plus large au contenu initial. Considérant une paire de texte court, ils traitent chaque élément comme une requête à envoyer vers un moteur de recherche Web, puis à partir des résultats renvoyés, ils créent pour chaque requête un vecteur de contexte contenant de nombreux mots ayant tendance à apparaître dans le contexte des termes d'origine, en utilisant TF * IDF comme

système de pondération. Enfin, la similarité est définie comme le produit scalaire des deux vecteurs de contexte pour les deux textes courts.

Soit un texte court X, le vecteur de contexte de x noté QE(x) (*Query Expansion*) sera calculé de la façon suivante :

1. Envoyer x en tant que requête auprès un moteur de recherche.
2. Soit R (x) l'ensemble des n top documents récupérés d_1; d_2; …; d_n
3. Calculez le vecteur de terme TF-IDF v_i pour chaque document $d_i \in R(x)$
4. Soit C(x) le centre des vecteurs v_i selon la norme L_2 : $C(x) = \sum_{i=1, n} v_i / \| v_i \|$
5. Soit QE(x) la normalisation centre de gravité C (x) : $QE(x) = C(x) / \| C(x) \|$

Considérant deux textes courts x et y, la similarité entre les deux peut être définie comme étant le produit scalaire des deux vecteurs de contexte : *QE(x).QE(y)*

Cette méthode peut être améliorée en apportant quelques traitement aux documents retournés par le moteur de recherche dans l'étape 2, ou bien en adoptant un nouveau système de pondération dans l'étape 3 au lieu du système utilisé TI-IDF. Nous présenterons une version améliorée de cette mesure dans la suite de ce rapport.

6 Conclusion

Dans ce chapitre, nous avons présenté un tour d'horizon du domaine de fouille de texte. Nous avons décrit les processus du prétraitement et représentation des données textuelles. Par la suite, nous avons introduit les problèmes liés au traitement du texte court ainsi que les applications associées de texte cout. Nous avons également présenté une investigation des travaux de recherche principalement liés à : la catégorisation de textes courts, les différentes approches permettant d'enrichir la représentation de texte court et les mesures de similarité pour le texte court.

Dans le chapitre suivant, nous présentons les différentes techniques de Traitement Automatique du Langage Naturel (TALN) dédiée à la langue arabe.

CHAPITRE 2

Traitement Automatique de la Langue Arabe

1 Introduction 31
2 Description et Caractéristiques de la Langue Arabe 31
3 Particularités de la Langue Arabe 33
3.1 Alphabet Arabe 33
3.2 Structure d'un Mot Arabe 34
3.3 Morphologie Arabe 35
3.4 Catégories des Mots Arabe 35
3.5 Grammaire et Caractéristiques de la Langue Arabe 35
4 Difficultés de l'Analyse Automatique de la Langue Arabe 36
4.1 Complexités du Traitement Automatique de la Langue Arabe 36
4.2 Exemples de Complexité de la Langue Arabe 36
4.3 Problème d'Encodage 37
4.4 Analyse Morphologique 38
4.5 Segmentation de Texte Arabe 40
4.6 Etiquetage Grammatical 40
4.7 Analyse Syntaxique 41
5 Techniques de Prétraitement 42
5.1 Tokenisation 42
5.2 Elimination des Mots Vides (Stop-Words) 43
5.3 Racinisation (Stemming) 44
5.4 Lemmatisation 45
5.5 Approches et Outils d'Analyse Morphologique 46
5.5.1 Approches d'Analyse Morphologique 46
5.5.2 Analyseurs Morphologiques 47
5.6 Approches d'Etiquetage Grammatical 49
6 Conclusion 50

1 Introduction

Les techniques du Traitement Automatique du Langage Naturel (TALN) permettent d'extraire à partir des données textuelles des informations plus riches que les simples unités lexicales. Ces informations de nature morphologique, syntaxique et sémantique ont été partiellement utilisées en fouille de texte pour améliorer la représentation du contenu des documents dans divers applications telles que la recherche d'information, la catégorisation, etc.

Dans ce chapitre nous présentons les différents prétraitements dédiés à la langue Arabe. Nous présentons également les différentes caractéristiques, particularité et problèmes de cette langue dans le but d'améliorer la représentation de texte court en langue Arabe.

2 Description et Caractéristiques de la Langue Arabe

La langue arabe est parlée dans plus de 22 pays, du Maroc jusqu'à l'Iraq et dans toute la péninsule arabe. C'est la langue maternelle pour plus de 206 millions de personnes et la langue parlée pour plus de 400 millions de personnes[21]. L'arabe, langue du Saint Coran, est devenue la langue d'une civilisation et ne sert plus seulement à désigner les seuls habitants de la péninsule arabe qui la parlaient [70]. La richesse de la langue arabe dans laquelle se déploient des variétés écrites et orales, rend cette dernière capable de répondre à un spectre très diversifié d'usages sociaux [71]. Mais au-delà de cette diversité, les sociétés arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène. Elles sont attachées à l'intégrité de leur langue, d'où l'importance de l'Arabe Standard Moderne (ASM) qui constitue le terrain commun pour cette large arabe.

D'après Boulaknadel [71], par ses propriétés morphologiques et syntaxiques, le traitement automatique de la langue arabe doit faire face à :

- la nature agglutinante de la langue : l'ensemble des morphèmes collés à l'unité lexicale qui véhiculent plusieurs informations morphosyntaxiques ;
- la richesse flexionnelle de l'arabe ;
- l'absence de voyellation de la majorité des textes arabes écrits : ce phénomène entraîne un nombre important d'ambigüités morphologiques. En arabe, chaque lettre doit prendre un

[21] https://www.internetworldstats.com/stats7.htm

signe de voyellation et les voyelles sont porteuses de certains traits morphosyntaxiques comme la déclinaison, le mode et le cas.

En outre des propriétés linguistiques, l'arabe recense un nombre de ressources linguistiques comprenant des lexiques monolingues et multilingues ainsi que des corpus de langue générale et des corpus de spécialité consacrés à une situation de communication ou à un domaine de la connaissance [71]. L'Arabe compte aussi un certain nombre d'outils linguistiques à savoir les analyseurs morphologiques ainsi que les extracteurs des racines basés essentiellement sur une procédure de désuffixation qui consiste à supprimer les suffixes qui différencient les flexions des unités lexicales (les formes conjuguées d'un verbe par exemple) [71].

L'arabe peut être considéré comme un terme générique rassemblant plusieurs variétés [71] :

- *l'Arabe Classique*[22] *:* est une des formes de l'arabe utilisée au moyen âge dans les textes littéraires pendant le califat omeyyade et abbasside (entre les VIIe et IXe siècles). Elle évolue au fil du temps de l'arabe précoranique à l'arabe coranique, puis à l'arabe post-coranique auquel est parfois réservée l'appellation « *arabe classique* ».
- *l'Arabe Standard Moderne (l'ASM) :* une forme un peu différente de l'arabe classique, et qui constitue la langue écrite de tous les pays arabophones. *L'ASM* reste le langage de la presse, de la littérature et de la correspondance formelle, alors que l'arabe classique appartient au domaine religieux et est pratiqué par les membres du clergé ;
- *les Dialectes Arabes :* malgré l'existence d'une langue commune, chaque pays a développé son propre dialecte. Issus de l'arabe classique, leurs systèmes grammaticaux respectifs affichent de nettes divergences avec celui de *l'ASM*. Boulaknadel [71] regroupe ces dialectes en quatre grands groupes :
  1. *les Dialectes Arabes*, parlés dans la Péninsule Arabique : dialectes du Golfe, dialecte du najd, yéménite ;
  2. *les Dialectes Maghrébins :* algérien, marocain, tunisien, hassaniya de Mauritanie ;
  3. *les Dialectes proche-orientaux :* égyptien, soudanais, syro-libano-palestinien, irakien (nord et sud) ;
  4. la langue *Maltaise* est également considérée comme un dialecte arabe.

[22] http://dictionnaire.sensagent.leparisien.fr/Arabe%20classique/fr-fr/

3 Particularités de la Langue Arabe

3.1 Alphabet Arabe

L'alphabet arabe comporte 28 consonnes (tableau 1) qui correspondent chacune à un phonème, 14 consonnes lunaires qui n'assimilent pas le ل de l'article et 14 consonnes solaires qui assimilent le ل de l'article.

Tableau 1 . Les Consonnes Arabes [72]

Consonnes Solaires	Consonnes Lunaires
ت ث د ذ ر ز س ش ص ض ط ظ ل ن	أ ب ج ح خ ع غ ف ق ك آ ه م و ي

De même, il existe six voyelles en arabe standard (tableau 2), trois longues et trois courtes, la durée phonétique d'une voyelle longue étant environ le double de celle d'une voyelle courte [72].

Tableau 2 . Les Voyelles Arabes [72]

Voyelles longues	Voyelles courtes
ا و ي	َ ُ ِ

Les lettres arabes changent de forme selon leur position dans le mot. Toutes les lettres de l'alphabet (à l'exception de ء), qu'elles soient manuscrites ou imprimées, sont agglutinées soit au début soit au milieu soit à la fin. Cependant, six d'entre elles s'attachent uniquement aux lettres précédentes mais pas aux lettres suivantes (tableau 3) [72].

Tableau 3 . Caractères qui ne s'attachent pas au suivant [72]

Les caractères qui ne s'attachent pas au suivant
و ز ر ذ د ا

La lettre «hamza-أ » peut également prendre ces formes : «أ،إ،آ،لأ،لإ», également pour la lettre «wāw-و» qui peut prendre cette forme «ؤ» [72].

3.2 Structure d’un Mot Arabe

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d’éléments de la grammaire [73]. La structure du mot arabe est donc décomposable en cinq éléments qui sont des morphèmes : **proclitique, préfixe, base, suffixe et enclitique** qui donnent des informations et des traits grammaticaux.

Les proclitiques et les enclitiques s'attachent au mot minimal par une seconde dérivation externe pour former le mot maximal. Par proclitique, on désigne les proclitiques simples (morphèmes à une lettre) et les proclitiques composés (morphèmes à plusieurs lettres). Les premiers sont des coordonnants, des conjonctions, des prépositions, etc. Les seconds sont obtenus par combinaison des premiers. L'article (ال : al) est également considéré comme proclitique simple, bien qu'il comporte deux lettres. L'enclitique est un pronom personnel complément qui peut être simple ou double. Il est attaché au mot qui le précède pour ne former qu'un seul mot.

La représentation suivante schématise la structure possible du mot «أتعاقبوننا» qui exprime la phrase suivante en français : «Est-ce que vous nous punissez ?». Il se décompose comme suit [72] :

- proclitique أ : conjonction d’interrogation ;
- préfixe ت : préfixe verbal du temps de l’inaccompli ;
- base عاقب : corps schématique dérivé de la racine ع ق ب;
- suffixe ون : suffixe verbal exprimant le pluriel ;
- enclitique نا : pronom, suffixe complément du nom.

a) Préfixes

Le préfixe est un morphème verbal relatif à l'inaccompli (actif ou passif) placé avant la base. Pour les noms et les autres aspects de la conjugaison nous dirons que le préfixe est vide. Nous prenons la racine à la 3ème personne du singulier de l’accompli, par exemple : كَتَبَ. A cette forme nous ajoutons un pronom personnel préfixe, par exemple « il » ي, le verbe passe alors au cas nominatif : يَكْتُبُ.

b) Suffixes

Le suffixe est un morphème situé immédiatement après la base. La base peut être une base verbale suivie d'un suffixe verbal ou une base nominale suivie d'un suffixe nominal. Un suffixe peut être une combinaison de deux suffixes tel que (ياء النسبة : Yâ ALNiSBa).

3.3 Morphologie Arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et les noms sont le plus souvent dérivés d'une racine à trois consonnes radicales [74]. Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est une caractéristique de la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine. On peut dériver un grand nombre de noms, de formes et de temps verbaux.

La majorité des verbes arabes ont une racine composée de trois consonnes. L'arabe comprend environ 150 schèmes ou patrons dont certains plus complexes, tel le redoublement d'une consonne ou l'allongement d'une voyelle de la racine, l'adjonction d'un ou de plusieurs éléments ou la combinaison des deux. Une autre caractéristique est le caractère flexionnel des mots : les terminaisons permettent de distinguer le mode des verbes et le cas des noms [74].

3.4 Catégories des Mots Arabe

La langue arabe comprend trois catégories de mots : le verbe, le nom et la particule [72] :

- le verbe est une entité qui exprime un sens dépendant de l'un des trois temps, l'accompli, l'inaccompli et l'impératif. Les verbes se divisent en deux catégories, les verbes «sains» et les verbes «défectueux» :
  - les verbes sains ne contiennent aucune des lettres défectueuses dans leur radical, à savoir les trois voyelles «alif-ا », «yāʾ-ي», «wāw-و».
  - les verbes défectueux contiennent au moins une lettre défectueuse.
- le nom désigne un objet ou un être exprimant une notion indépendante du temps ;
- les particules servent à lier les noms, les verbes et les parties de la phrase mais également à indiquer le temps.

3.5 Grammaire et Caractéristiques de la Langue Arabe

La grammaire traditionnelle se divise en deux branches [71] :

1. La morphologie, الصرف, qui comprend :
   - *Morphologie dérivationnelle*, qui étudie la construction des unités lexicales et leur transformation selon le sens voulu. Ainsi, la dérivation morphologique est décrite sur une

base morphosémantique : d'une même racine, se dérivent différentes unités lexicales selon des schèmes qui sont des adjonctions et des manipulations de la racine.

- *Morphologie flexionnelle* concerne le marquage casuel pour le nom et l'adjectif ou la conjugaison du verbe, appelé "الإعراب".

2. La syntaxe, "النحو ", qui étudie la construction correcte des phrases garantit la grammaticalité de la phrase en analysant :
   - la position des unités lexicales les unes par rapport aux autres, déterminant ainsi l'ordre des unités lexicales ;
   - le marquage casuel des unités lexicales de la phrase. Ainsi, la fonction syntaxique de l'unité lexicale est déterminée en s'appuyant sur la morphologie.

4 Difficultés de l'Analyse Automatique de la Langue Arabe

4.1 Complexités du Traitement Automatique de la Langue Arabe

La langue arabe est une langue difficile pour un certain nombre de raisons [75] :

- L'orthographique avec des signes diacritiques est moins ambiguë et plus phonétique en arabe, certaines combinaisons de caractères peuvent être écrites de différentes façons.
- La langue arabe a des voyelles courtes qui aident à des prononciations différentes. Grammaticalement ils sont nécessaires, mais omis dans les textes arabes écrits.
- Manque des corpus arabes librement accessibles.

4.2 Exemples de Complexité de la Langue Arabe

a) Sens des Mots

Le sens d'un mot d'après son contexte, le tableau 4 présente un exemple pour le mot قلب.

Tableau 4 . Différents sens du mot (قلب) [114]

Le sens du mot	La phrase
L'actualité importante	في قلب الأحداث
Le cœur	أجرى عملية قلب مفتوح
Le cœur du stade	الكرة في قلب الملعب

b) Variations dans la Catégorie Lexicale

Le sens d'un mot d'après son catégorie lexicale, le tableau 5 présente un exemple pour le mot ذهب.

Tableau 5 . Catégorie lexicale du mot (ذهب) [114]

Le sens du mot	La catégorie du mot	Phrase
Or	Nom	ذهب خالص
Aller	Verbe	ذهب إلى المدرسة

c) Synonymes

Les langues ont beaucoup de mots qui sont considérés comme des synonymes. Grâce à un corpus donné, les chercheurs peuvent utiliser les outils de l'analyse morphologique pour connaitre les synonymes d'un mot, la fréquence de chaque mot de ces synonymes et le plus connu entre eux. Des exemples des synonymes en arabe sont (وهب أعطى، منح، بذل،) qui signifie (donner), (عائلة, أسرة) qui signifie (famille), et (صف, فصل) qui signifie (classe) [75].

d) Forme de Mot en Fonction de son Cas

La forme de quelques mots arabes peut changer en fonction de leurs modes de cas (nominatif, accusatif ou génitif). Par exemple le pluriel du mot (مسافر) qui signifie (voyageur) peut avoir la forme (مسافرون) dans le cas du nominative (مرفوعة) et la forme (مسافرين) dans le cas de l'accusatif / génitif (مجرورة / منصوبة) [75].

e) Caractéristiques Morphologiques

Un mot arabe peut être composé d'un lemme, des affixes et des clitiques. Le lemme est obtenu à partir d'une racine consonantique (جذر) et d'un modèle. Les affixes flexionnels comprennent des marqueurs (علامات أو حركات إعرابية) pour le temps, le sexe et/ou le nombre. Les clitiques comprennent certaines prépositions (حروف جر), conjonctions (حروف العطف), déterminants (محددات), et pronoms (ضمائر). Les clitiques attachés au début d'un stem sont appelés proclitique et ceux fixés à l'extrémité de celui-ci sont appelés enclitiques [75].

4.3 Problème d'Encodage

La langue arabe a des problèmes d'affichage parce que son encodage diffère selon la plate-forme de la machine. La figure 2 montre le problème d'encodage où toutes les cellules ombrées sont affichées correctement tandis que les autres cellules ne sont pas correctes. Le prétraitement et la classification du texte arabe avec un encodage incorrect peut conduire à des résultats

erronés. Le tableau 6 présente les caractéristiques des deux systèmes communs d'encodage arabes : Unicode et CP-1256 [75].

Tableau 6 . L'encodage de la langue arabe (Unicode vs. CP-1256) [114]

Unicode	**CP-1256 (code page 1256 Arabic windows)**
Devenir la norme standard de plus en plus	couramment utilisé
Caractères codés sur 2 octets	Caractères codés sur 1 octet
Entrée / affichage largement soutenu	Entrée / affichage largement soutenu
Prise en charge étendue des caractères arabes	Ne peut pas supporter les longs caractères arabes
Représentation multi-script	Supporte bi-script (Roman / arabe)
Prise en charge des formulaires de présentation (formes et des ligatures)	Supporte trilingue : arabe, français, anglais (ANSI)

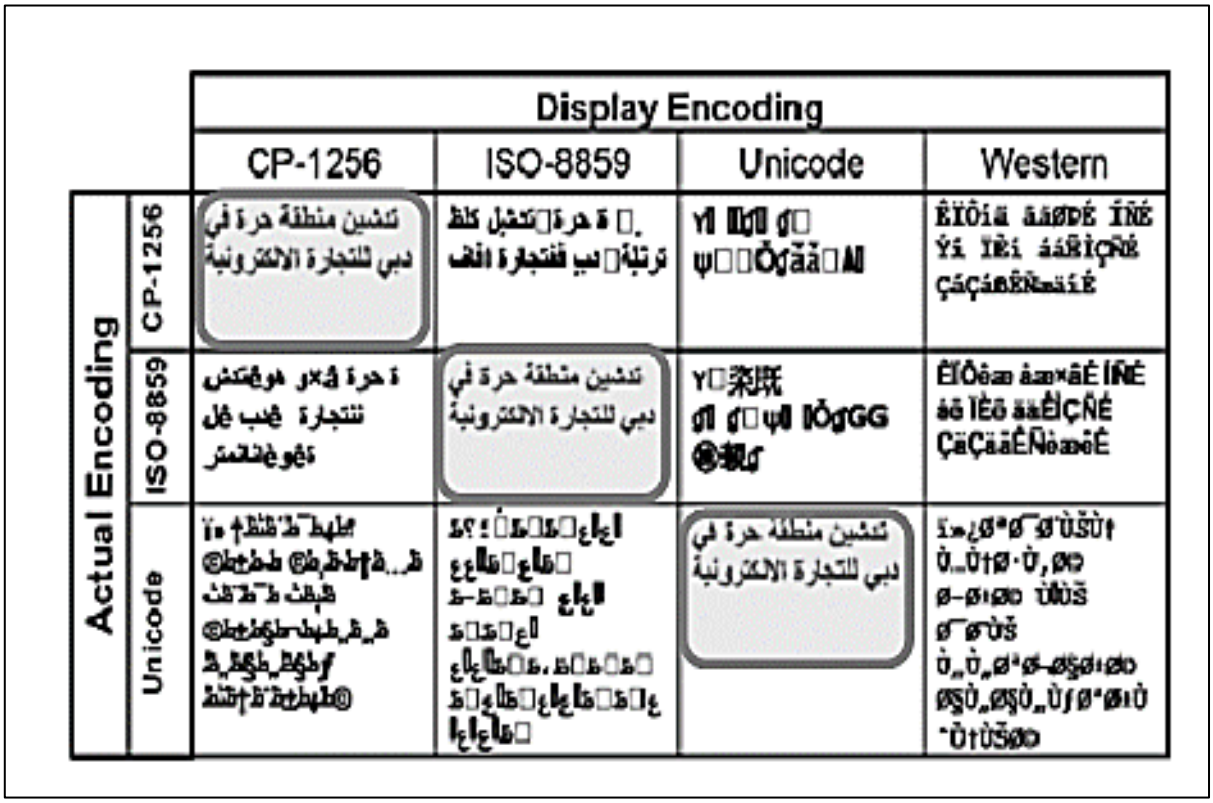


Figure 2 . Le problème de l'encodage de la langue arabe [114]

4.4 Analyse Morphologique

L'analyse morphologique s'intéresse à la formation des mots à travers des processus de flexion (marques de genre, nombre, de conjugaison...), dérivation et composition. Étant donné un mot, il s'agit de déterminer quelles sont les unités minimales de sens qui le composent. Ces unités

minimales de sens sont appelées morphèmes et se déclinent en termes de racine et d'affixes. Une analyse morphologique complète précise en plus la catégorie grammaticale de la racine et associe aux affixes des informations sémantiques et flexionnelles. Elle est réalisée soit en utilisant des bases lexicales existantes, soit à l'aide de véritables systèmes d'analyse, plus à même de traiter les formes non répertoriées.

Le problème principal de cette analyse réside dans l'agglutination et l'absence de voyellation. Pour l'agglutination en arabe et contrairement aux langues latines, les pronoms, les prépositions, les articles, les conjonctions, et autres particules collent aux noms, verbes, adjectifs et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française [114].

Cette caractéristique engendre une ambiguïté morphologique au cours de l'analyse. Ainsi, la reconnaissance des unités lexicales qui composent une unité morphologique n'est pas toujours facile à détecter. Le problème est de reconnaître que la bonne segmentation réside ainsi, dans la difficulté de distinction entre un proclitique ou enclitique et un caractère original du mot.

Par exemple, le caractère "و" dans le mot « il est arrivé | وصل » est un caractère original alors que dans le mot "et il a ouvert | وفتح ", il s'agit plutôt d'un proclitique.

L'absence de voyellation pose un autre problème important. En effet, les mots non voyellés engendrent beaucoup de cas ambigus au cours de l'analyse (e.g. le mot non voyellé "فصل" pris hors contexte peut être un verbe au passé conjugué à la troisième personne du singulier "il a licencié | فَصَلَ", ou un nom masculin singulier "chapitre/ saison | فَصْلٌ", ou encore une concaténation de la conjonction de coordination "puis | ف " avec le verbe "صل" : impératif du verbe lier conjugué à la deuxième personne du singulier masculin) [70].

L'analyse morphologique est réalisée soit en utilisant des bases lexicales existantes, soit à l'aide de véritables systèmes d'analyse, plus à même de traiter les formes non répertoriées. Elle est présentée dans de nombreuses applications du Traitement Automatique des Langues (TAL) et permet de reconnaître la présence d'un même mot sous des formes de surface différentes. L'une des analyses morphologiques les plus simples mises en œuvre en Recherche d'Information est la Racinisation. En extraction de connaissance, l'analyse morphologique la plus employée est la Lemmatisation qui permet d'associer à une forme fléchie une forme conventionnelle ainsi que de calculer les traits flexionnels.

4.5 Segmentation de Texte Arabe

La segmentation d'un texte est une étape fondamentale pour son traitement automatique ; son rôle est de découper le texte en unités simples d'informations. En effet, l'opération de segmentation d'un texte consiste à délimiter les segments de ses éléments de base qui sont les caractères, en éléments constituants différents niveaux structurels tels que : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème, etc [70].

Pour la langue arabe, Maâloul et al. [70] affirment que les particularités de cette langue, rend la segmentation arabe toujours beaucoup plus difficile par rapport aux autres langues, étant donné qu'il n'y a pas de majuscules qui marquent le début d'une nouvelle phrase. De plus, les signes de ponctuation, ne sont pas utilisés de façon régulière.

D'après l'étude réalisée par Belguith et al. [76], certaines particules comme "et |و", "donc| ف", etc. jouent un rôle principal dans la séparation de phrases et peuvent être déterminantes pour guider la segmentation.

Pour la segmentation de texte Ouersighni et al. [77] utilisent :

- Une segmentation morphologique basée sur la ponctuation,
- Une segmentation basée sur la reconnaissance de marqueurs morphosyntaxiques ou des mots fonctionnels comme : حتى ، لكن ، أي ، و ، أو (ou, et, mais, quand).

4.6 Etiquetage Grammatical

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (non, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît [70].

Dans le même cadre, la difficulté de l'étiquetage grammatical [70] peut s'amplifier lorsque les textes visés se présentent sous leur forme non pas voyellée, mais partiellement voyellée ou encore totalement non voyellée, ce qui correspond au cas le plus courant.

Ce problème de voyellation pour un mot [70] est ainsi posé du fait que le choix de l'accentuation qui convient au mot est difficile et dépend essentiellement du contexte.

Le tableau 7 présente le problème d'ambiguïté grammaticale rencontrée lors de l'attribution catégorique d'un mot non voyéllé "كتب ", qui admet au moins quatre étiquettes grammaticales qui sont les suivantes [70] :

Tableau 7 . Exemple d'étiquettes grammaticales attribuées selon la voyellation [114]

Exemple de voyellation	Étiquettes grammaticales
كُتُبٌ kutubun : des livres	substantif, masculin, pluriel
كَتَبَ kataba : il à écrit	verbe, 3ème personne masculin, singulier de l'accompli actif
كُتِبَ kutiba : il a été écrit	verbe, 3ème personne masculin, singulier de l'accompli passif
كَتِّبْ kattib : fais écrire	verbe à l'impératif, 2ème personne masculin, singulier

4.7 Analyse Syntaxique

Maâloul et al. [70] définissent l'analyse syntaxique comme une analyse qui permet d'associer à un énoncé sa ou ses structures syntaxiques possibles, en identifiant ses différents constituants et les rôles que ces derniers entretiennent entre eux. Cette analyse syntaxique prend en entrée le résultat de l'analyse lexicale (éventuellement de l'étiquetage morphosyntaxique) et fournit en sortie une structure hiérarchisée des groupements structurels et des relations fonctionnelles qui unissent les groupements.

Il a signalé aussi, dans la même voie de recherche, que les ambiguïtés vocaliques et grammaticales, relatives au non voyellation des mots, posent des difficultés au niveau de l'analyse syntaxique. Ainsi, une phrase, en absence de la voyellation, peut être interprétée et traduite selon plusieurs interprétations qui sont toutes syntaxiquement correctes [70].

a) Absence de Voyelles

L'absence de la voyellation dans les textes arabes présente un problème important lors de l'analyse automatique de ces textes. Maâloul et al. [70] affirment que l'ambiguïté grammaticale augmente si le mot est non voyellé. Cela est dû au fait qu'un mot non voyellé possède plusieurs voyellations possibles, et pour chaque voyellation est associée à une liste différente de catégories grammaticales.

b) Agglutination

Dans la langue arabe, les articles, les prépositions, les pronoms, etc. se collent, contrairement aux langues latines, aux adjectifs, aux noms, aux verbes et aux particules auxquels ils se rapportent. Comparé au français, un mot en arabe peut parfois correspondre à une phrase en français. Cette caractéristique peut engendrer une ambiguïté au niveau morphologique pour cette langue. En effet, il est parfois difficile de distinguer entre un proclitique ou enclitique et un caractère original du mot [70].

c) Irrégularité de l'Ordre des Mots dans la Phrase

Généralement l'ordre des mots dans une phrase écrite en langue arabe est relativement libre. Nous pouvons mettre au début de la phrase le mot sur lequel on veut attirer l'attention et l'on termine sur le terme le plus long ou le plus riche en sens ou en sonorité [70]. Cet ordre provoque des ambiguïtés syntaxiques artificielles, dans la mesure où il faut prévoir dans la grammaire toutes les règles de combinaisons possibles d'inversion de l'ordre des mots dans la phrase [76]. Ainsi par exemple, le changement de l'ordre des mots dans la phrase ne produit pas deux phrases ayant des sens différents.

d) Absence de Ponctuation Régulière

Pour la langue Arabe Standard Moderne (ASM), utilisée actuellement sur le Web et les différents outils de médias sociaux, il y'a un manque de signes de ponctuations et des marqueurs typographiques ; et même dans le cas où ils y figurent, ils ne sont pas gérés par des règles précises d'utilisation.

Par ailleurs, les paragraphes écrits avec l'ASM ne contiennent aucun signe de ponctuation à part un point à la fin de ce paragraphe. Ainsi, il est à noter que la présence des signes de ponctuation ne peut pas guider la segmentation comme c'est le cas pour d'autres langues, telles que le français ou l'anglais. Ainsi, la segmentation de textes arabes doit être guidée non seulement par les signes de ponctuations et les marqueurs typographiques mais aussi par des particules et certains mots tels que les conjonctions de coordination, etc. [76].

5 Techniques de Prétraitement

5.1 Tokenisation

La première étape dans un processus de traitement d'un gros corpus au moyen d'un outil statistique est de subdiviser le texte à traiter en plusieurs unités d'information appelées tokens qui sont, traditionnellement, des mots simples. Ce processus de tokenisation [82] pose une question primordiale : sur le plan informatique, comment repérer un mot ? En d'autres termes, quels sont les indicateurs formels de surface, non ambigus, qui peuvent délimiter un mot ? Si pour le français ou l'anglais littéraire, ou des langues apparentées, la réponse est presque triviale — à savoir que toute chaîne de caractères précédée et suivie d'un espace est considérée comme un mot simple — il en va tout autrement pour d'autres langues.

Dans le cas de termes composés en langue Allemande comme, par exemple, lebensversicherungsgesellschaftsangestellter (“employé d’une compagnie d’assurance vie”), ou pour la langue Arabe dans laquelle les pronoms sujets et compléments sont en général attachés aux verbes et une seule chaîne de caractères représente ainsi une phrase comme, par exemple, «كتبته » (je l’ai écrit), cette notion de tokens devient inadéquate.

Dans une perspective de classification numérique à des fins d’extraction de connaissances, la définition d’une unité d’information est tributaire des contraintes suivantes :

- L’unité d'information doit être une portion du texte soumis à l’analyse numérique. Il doit être facile sur le plan informatique de repérer les unités d’information.
- La définition d’une unité d’information doit être indépendante de la langue dans laquelle le texte est écrit. Une telle définition permet à l’analyse numérique, moyennant des modifications minimes, de couvrir un large éventail de langues.
- Les unités d’information doivent être statistiquement comparables. Il doit être aisé d’en calculer les fréquences d’apparition dans les différentes parties du texte et par conséquent d’estimer leur distribution et la régularité à laquelle plusieurs unités cooccurrent dans les mêmes parties du texte.

Que l’unité linguistique dans une analyse de classification numérique soit linguistiquement comprise dans une certaine mesure hors de son contexte n’est pas en soi une contrainte lors de la phase de classification. Toutefois à l’affichage des résultats, ce facteur devient important dès lors que l’interprétation faite par l’utilisateur en dépend. La plupart des analyseurs statistiques fondés sur le calcul de la fréquence des cooccurrences utilisent le mot comme unité d’information, même si celui-ci ne répond pas à toutes les contraintes énumérées ici.

Cependant, l’importance de l’ergonomie interprétative des mots a prévalu sur tout autre aspect, particulièrement ceux liées aux aspects multilingues de l’analyse. Ce dernier facteur devient aujourd’hui incontournable : l’essor du Web confirme ce besoin de multilinguisme. Il semble donc impératif que les modèles pour l’analyse de corpus, qu’ils soient numériques ou linguistiques, tiennent compte du caractère multilingue des textes à analyser.

5.2 Elimination des Mots Vides (Stop-Words)

Les mots les plus communs dans n’importe quel document texte ne fournissent pas de signification aux documents, ceux sont des prépositions, des articles et des pronoms, etc. Ces mots sont traités comme des mots vides. Parce que chaque document texte est traité avec ces

mots qui ne sont pas nécessaires pour les applications de fouille de texte. Ces mots sont éliminés.

Tout groupe de mots peut être choisi comme des Stop-words pour un but donné. Ce procédé réduit également les données de texte et améliore les performances du système [82]. Exemple de mots vides : "إلى ", "في ", "متى " etc.

5.3 Racinisation (Stemming)

La Racinisation (Stemming) [78] (figure 3) est une procédure qui vise à regrouper les mots sémantiquement proches à partir de ressemblances. On peut dire aussi que la racinisation cherche à rassembler les différentes variantes d'un mot autour d'une racine. Cette procédure traite à la fois des cas relevant de la flexion et de la dérivation. Les techniques utilisées pour procéder à la racinisation reposent généralement sur une liste d'affixes (suffixes, préfixes, post fixe, antéfixes) de la langue considérée et sur un ensemble de règles de dé-suffixation construites a priori qui permettent, étant donné un mot de trouver son stem/racine.

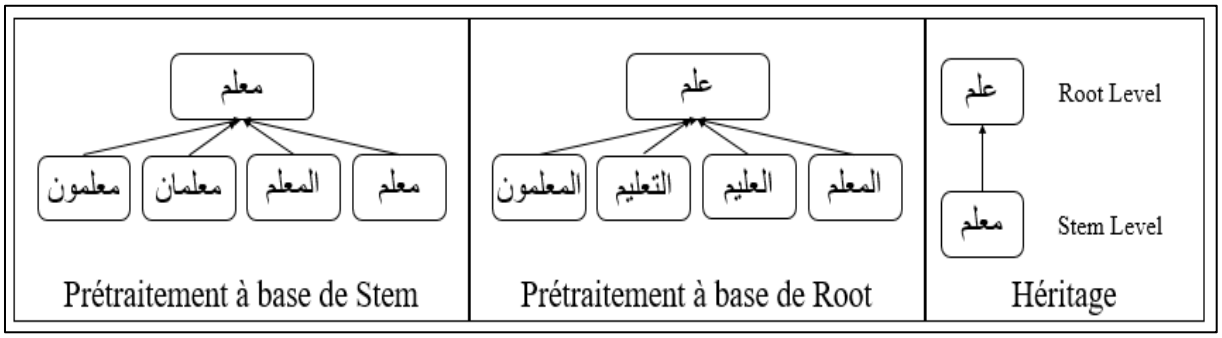


Figure 3 . Un exemple de prétraitement (Root / Stem) [114]

Bien que le principal avantage de ces outils réside dans leur simplicité, l'absence de contraintes linguistiques fortes engendre néanmoins des erreurs. La racinisation peut se faire par approximation des phénomènes linguistiques en jeu. Le principe général de tous les algorithmes classiques de racinisation repose sur deux opérations :

- Désuffixer: supprimer les suffixes qui différencient les flexions d'un mot (par exemple les formes conjuguées d'un verbe) et les mots d'une même famille morphologique;
- Recoder: regrouper les différentes variantes graphiques d'une même racine (ses allomorphes).

Les algorithmes de racinisation sont nécessaires dans de nombreuses applications telles que le Traitement du Langage Naturel (TAL), la compression de données, et les Systèmes de

Recherche d'Information. Dans la plupart des cas, ces variantes ont une interprétation sémantique similaire et peuvent être traitées d'une manière équivalente pour les applications de fouille de texte. Les algorithmes de racinisation peuvent être utilisés pour réduire les termes ont leurs racines.

Pour la langue Arabe, il existe deux techniques d'analyse morphologiques différentes, la racinisation et la racinisation légère (Light Stemming). La racinisation réduit les mots à leurs racines [79]. La racinisation légère (Light Stemming), en revanche, supprime les affixes communs de mots sans les réduire à leurs racines. La racinisation réduirait les mots arabes (المكتبة، الكاتب، الكتاب) qui signifie (la bibliothèque, l'écrivain, et le livre), respectivement, à la racine (كتب), ce qui signifie (écrire).

L'algorithme de racinisation de Khoja [79], est le stemmer arabe le plus connu. Le stemmer de Khoja supprime le plus long suffixe et le plus long préfixe. Il compare ensuite le mot restant avec des modèles verbaux et nominaux, pour extraire la racine. Le stemmer fait usage de plusieurs fichiers linguistiques de données comme une liste de tous les caractères diacritiques, caractères de ponctuation, des articles précis et les mots vides (stop words). Al- Shalabi, Kanaan et Al- Serhan [80] ont développé un algorithme d'extraction de la racine (extraction de la racine tri- littéral) qui n'utilise pas de dictionnaire. Cela dépend de l'assignation des pondérations pour les lettres d'un mot multiplié par la position de la lettre, les consonnes ont été affectés d'un poids de zéro et des différents poids ont été affectés aux lettres regroupées dans le mot (سألتمونها) où tous les affixes sont formés par des combinaisons de ces lettres. L'algorithme sélectionne les lettres avec le poids le plus bas comme des lettres de racines. Dernièrement, Zeroual et al. [81] ont développé Al-Khalil Stemmer, c'est un stemmer léger avec un algorithme composé de deux phases importantes : la suppression de tous les clitiques possibles et le recherche des racines correctes à l'aide des dictionnaires de ressources linguistiques arabes.

5.4 Lemmatisation

La lemmatisation ou étiquetage morphologique consiste à assigner à chaque mot d'un texte son lemme, éventuellement associé à des informations flexionnelles. Un lemme est une forme lexicale existante choisie par convention au sein d'un paradigme flexionnel. Ainsi pour le Français, le lemme d'un nom correspond à la forme au singulier, celui d'un adjectif à la forme masculin singulier et celui d'un verbe à l'infinitif. La lemmatisation s'effectue en examinant le mot seul, ou à partir du mot accompagné de sa catégorie grammaticale. Un programme utilisant le mot et son étiquette peut être vu comme un sous-produit immédiat de l'étiquetage

morphosyntaxique puisqu'un lemme correspond à une forme fléchie unique dès que sa catégorie grammaticale est connue, sauf pour un nombre fini d'exceptions. Un exemple d'exception est la forme fléchie « fils » qui peut correspondre à deux lemmes différents, « fil» et «fils». Il existe deux types principaux de lemmatiseurs : ceux utilisant une base de données lexicales et ceux utilisant un analyseur morphologique.

Pour la langue Arabe la lemmatisation (tableau 8) se définit par l'identification d'une forme canonique correspondant à différentes formes flexionnelles ou dérivationnelles d'un mot donné (dérivation du pluriel, du singulier, du féminin ou du masculin, etc.). L'application de la lemmatisation en recherche d'information en langue Arabe ne donne pas toujours les résultats escomptés car le système régissant la dérivation en Arabe est plus complexe et ne se résume pas, souvent, à une simple suffixation.

Tableau 8 . Exemples de lemmatisation pour la langue Arabe

Mot en arabe	Sens du mot	Lemme	Sens du lemme
رجال	Les hommes	رجل	Un homme
كتابات	Les écritures	كتابة	Une écriture

5.5 Approches et Outils d'Analyse Morphologique

5.5.1 Approches d'Analyse Morphologique

L'analyse morphologique a fait l'objet de plusieurs travaux de recherche qui sont classés généralement en deux approches [70] :

- La première approche utilise un lexique comportant tous les mots (sous leurs différentes formes possibles) avec leurs caractéristiques associées. Dans ce cas, l'identification des caractéristiques morphologique des mots s'effectue par une simple consultation du lexique. Cette approche suggère de stocker tous les mots avec leurs différentes formes. Dans ce cas, les dictionnaires se doivent de consigner toutes les unités attestées possibles, simples et construites. Ce type de lexique est appelé lexique en extension.
- La deuxième approche réduit le lexique aux seules informations non calculables (exemple : formes canoniques, racines, etc.) et utilise des règles qui mettent en jeu un savoir linguistique dans les traitements pour connaître le reste des informations. Cette approche consiste à choisir un lexique en intension, c'est-à-dire le noyau du lexique dénué de toute redondance

pouvant servir de base à l'inventaire complet des lexèmes de la langue si on se donne un ensemble approprié de règles morphologiques.

Dans ce qui suit, nous présentons quelques systèmes d'analyse morphologique de la langue Arabe tels que le système *ARAMORPH* de *Tim Buckwalter* [83][23], le système *MORPH-2* de *Chaâben* qui a été réalisé au sein du laboratoire *MIRACL* [70] et l'analyseur morphologique Al-Khalil [81].

5.5.2 Analyseurs Morphologiques

a) Buckwalter

L'analyseur de Buckwalter [83] développé par LDC (Linguistic Data Consortium) permet de segmenter chaque unité lexicale en une séquence du type préfixe-stem-suffixe. Le préfixe est une combinaison de *0-4* caractères, le suffixe est composé de *0* à *6* caractères et le stem comprend un à plusieurs caractères. Il est constitué principalement de trois lexiques : préfixes (*548* entrées), suffixes (*906* entrées), et stem (*78839* entrées). Les lexiques sont complétés par trois tables de compatibilité utilisés pour couvrir toutes les possibilités de combinaisons préfixe-stem (*2435* entrées), suffixe-stem (*1612* entrées) et préfixe-suffixe (*1138* entrées). Ainsi, l'analyseur donne en sortie l'unité lexicale, sa catégorie morphosyntaxique et sa traduction anglaise.

b) Aramorph

Aramorph, est un analyseur morphologique réalisé par *Tim Buckwalter*, qui utilise en entrée un texte qui doit être translittéré en *ASCII* avant tout traitement, et le résultat doit être reconverti en langue Arabe pour qu'il soit compréhensible. Ce système ne permet pas l'analyse des textes contenant des chiffres et il se limite uniquement à l'analyse des mots qui figurent dans les dictionnaires arabes.

L'analyseur morphologique *Aramorph* [84] segmente les unités lexicales, repère les différents composants et atteste son appartenance à la langue. Pour cela, le système est assisté par le lexique *DIINAR.1* [85] pour éviter les analyses théoriquement possibles et inexistantes dans la langue. Par la suite, l'analyseur donne une liste des traits associés à l'unité lexicale en entrée. Il offre deux types d'options. La première vise les traits morphosyntaxiques, la seconde concerne l'analyse des préfixes et suffixes.

[23] http://www.nongnu.org/aramorph/french/

En plus des étiquettes morphosyntaxiques, il donne en sortie d'autres informations comme la base, l'unité lexicale minimale vocalisé ou non ainsi que la forme complète supposée vocalisée ou non.

Analyser les préfixes revient à décrire ses découpages possibles et d'examiner les compositions des clitiques. Ceci amène le système à faire la distinction entre les clitiques ayant la même forme mais appartenant à des catégories syntaxiques différentes.

c) MORPH-2

L'analyseur morphologique *MORPH-2* de textes arabes non voyellés a comme objectif la détermination des caractéristiques morphosyntaxiques de chaque mot composant le texte. *MORPH-2* utilise un lexique constitué de *3266* racines trilitères et quadrilatères de la langue Arabe et repose sur cinq étapes à savoir : la segmentation de la phrase en mots, le prétraitement morphologique, l'analyse affixale, l'analyse morphologique et le posttraitement. La première étape est une étape de prétraitement permettant de segmenter la phrase en mots. La deuxième étape est aussi une étape de prétraitement mais elle s'effectue au niveau du mot. Elle permet de supprimer les enclitiques et les proclitiques agglutinés au mot pour obtenir sa forme minimale. Ensuite, une étape de filtrage permettra l'élimination des particules de la liste des mots qui vont subir l'analyse affixale. La troisième étape permet la détermination des préfixes, infixes et suffixes du mot, ainsi que la racine, s'il s'agit d'un verbe, ou la forme canonique s'il s'agit d'un nom. La quatrième étape consiste à déterminer, à partir de chaque combinaison, les caractéristiques morphosyntaxiques correspondantes et de les affecter au mot correspondant. La cinquième et dernière étape consiste à déterminer les groupements de mots possibles dans la phrase.

Ce système a été évalué sur un corpus de textes non voyellés pris d'un livre scolaire Tunisien (composé de *81* textes formant *899* paragraphes et *29 188* mots) et a donné un taux de rappel global de *69,77%* et un taux de précision global de *68,51%* [76].

d) Al-Khalil

L'analyseur morphologique d'*Al-Khalil* [81] est un fruit de coopération entre l'organisation arabe pour l'éducation et culture et science au Saudia et l'université du premier Mohamed au Maroc. C'est le deuxième logiciel gratuitement téléchargé après l'analyseur de *Buckwalter* mais il est plus profond avec l'extraction des racines et des schèmes d'un lexème tout en gérant les entrées et les sorties en deux langues anglaise et arabe.

Al-Khalil analyse le mot arabe aux composants principaux comme : la racine, le schème et le lemme ainsi les préfixes et les suffixes, la vocalisation de mot et le type. La priorité qui est utilisé pour sortir des résultats associés au type de mot est définie d'une façon algorithmique [81]. Il est développé sous java et produit des résultats comme dans (Figure 4).

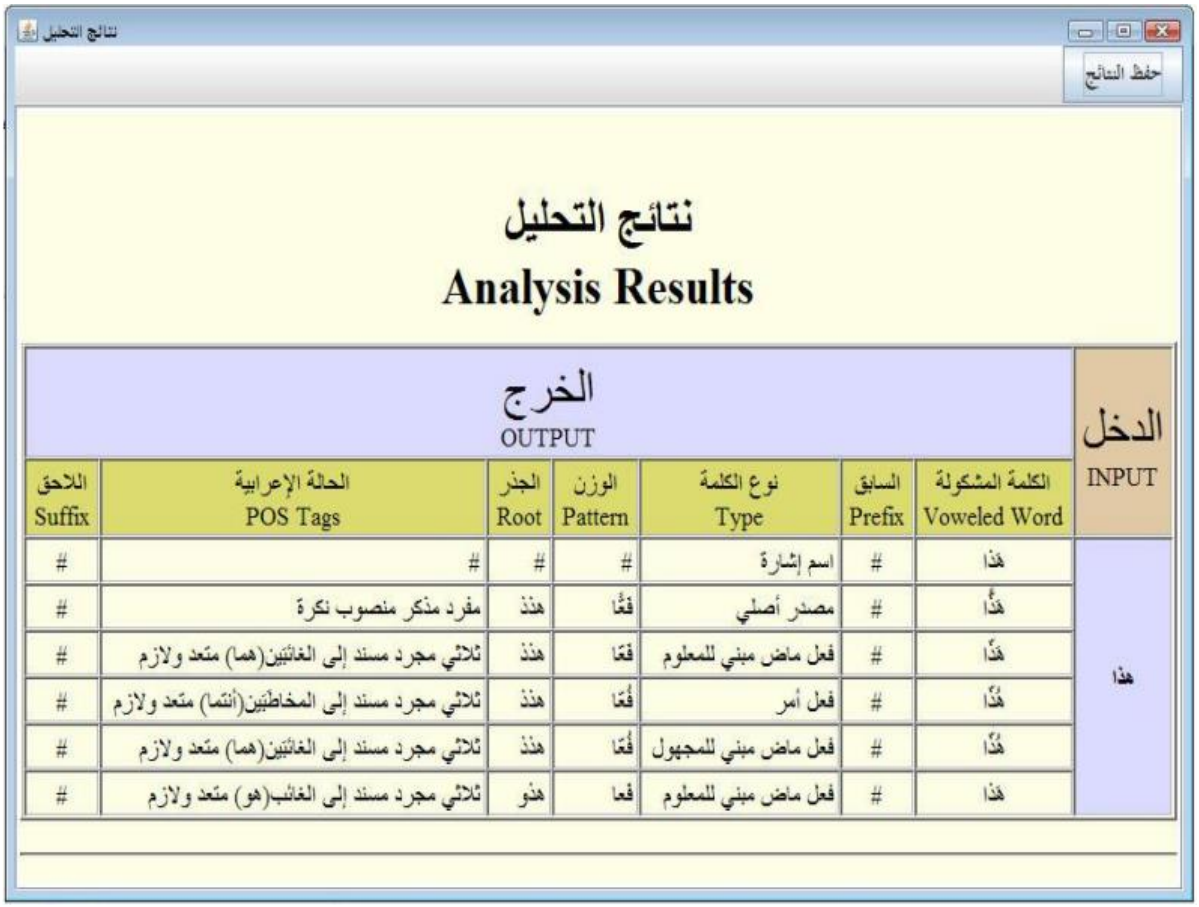


Figure 4 . Exemple d'un résultat d'analyse de mots «هذا » (hada) par Al-Khalil [81]

5.6 Approches d'Etiquetage Grammatical

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (non, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît [70].

Les approches d'étiquetage grammatical peuvent être divisées en deux principales catégories : numériques et probabilistes. L'approche numérique fait intervenir des règles qui portent sur les successions permises ou non de deux, trois ou *n* étiquettes grammaticales. Toutefois, pour l'approche probabiliste, une formulation probabiliste est utilisée pour la résolution de l'ambiguïté dans le choix de la règle. Ainsi, les règles se sont vu adjoindre des poids statistiques afin de choisir les résolutions les plus probables. Ces règles peuvent être lues de plusieurs façons

: on peut dire par exemple qu'après une telle étiquette, ce sont telles ou telles étiquettes qui peuvent suivre ; mais si l'on considère la dernière étiquette, on peut également dire qu'elle dépend de celles qui la précèdent. Ainsi est-ce que la formulation probabiliste utilisant les chaînes de Markov comme modèle qui s'est très vite répandue dès la fin des années 70 [53] peut conduire à des meilleurs résultats.

6 Conclusion

Dans ce chapitre nous avons présenté une idée claire et globale sur les différents points concernant le Traitement Automatique du Langage Naturel (TALN) en traitant le cas de la langue Arabe. Les différentes caractéristiques, particularité et problèmes de cette langue ont été étudiés dans ce présent chapitre dans le but d'améliorer la représentation de texte court en la langue Arabe.

A travers le tour d'horizon sur les plusieurs techniques du *TALN* appliquées à la langue Arabe, présenté dans ce chapitre, nous avons constaté que ces techniques permettent d'acquérir, à partir des documents, des informations plus riche que les unités lexicales simples. Ces informations permettent une meilleure représentation du contenu textuel.

Partie 2 : Enrichissement de Texte Court basé sur les Termes (Term-based)

- **Chapitre 3 : Enrichissement de Texte Court basé sur la Théorie des Ensembles Approximatifs (RST)**

- **Chapitre 4 : Enrichissement de Texte Court basé sur le Corpus Web**

Introduction de la partie 2

Récemment, avec l'explosion de l'utilisation des réseaux sociaux avec plus de 2 milliards d'utilisateurs actifs dans le monde entier, ces réseaux sociaux constituent un outil collaboratif idéal pour les utilisateurs qui chercher à échanger, partager de l'information, communiquer et aussi s'amuser dans un seul endroit comme Twitter, Facebook, etc. [86]. Ces réseaux sociaux permettent aux utilisateurs de publier un texte court. Les textes courts sont différents des documents traditionnels par leur brièveté et la rareté des termes partagés (***Shortness and Sparseness***). Par conséquent, l'absence d'informations contextuelles augmente le degré d'ambiguïté, surtout pour la langue arabe qui est caractérisé comme étant une langue flexionnelle où un seul mot peut avoir plusieurs significations. Par exemple, le mot arabe «علم» possède plus de trois sens ; cela peut exprimer : science (nom), drapeau (nom), enseigner (verbe) ... etc. De ce fait, les techniques traditionnelles de représentation de texte seront inadéquates et peuvent avoir un impact négatif sur la précision de toute application de fouille de texte, telle que la catégorisation.

Pour résoudre les problèmes de brièveté et le manque d'information contextuelle dans le texte court, deux approches principales ont été exploitées dans le but d'enrichir sa représentation avec l'intégration des informations sémantiques supplémentaires. Les informations d'extension pourraient provenir de la collection de données elle-même en se basant sur des techniques de traitement automatique du langage naturel ou provenant à partir des ressources sémantiques externes telles que Wikipédia[24], WordNet [106], BabelNet [107] et les moteurs de recherche tels que Google[25] et Bing[26].

Cette partie comprend deux chapitres qui présentent deux contributions concernant l'enrichissement de la représentation du texte court en arabe. Dans le premier chapitre, nous proposons une méthode pour améliorer et renforcer les liens sémantiques entre les éléments de collection de données à travers la Théorie des Ensembles Approximatifs (*Rough Set Theory – RST*) [91]. C'est outil mathématique qui a été proposé pour faire face à l'imprécision et à l'incertitude. Le deuxième chapitre présente une autre contribution qui se base principalement sur l'utilisation du Web comme étant un corpus ouvert et une source d'information dynamique.

[24] https://www.wikipedia.org/
[25] https://www.google.com/
[26] https://www.bing.com/

En utilisant le grand volume de documents dans le Web, nous pouvons ajouter des termes contextuels qui aident à fournir un contexte plus précis au texte court original en fonction des résultats renvoyés par le moteur de recherche.

CHAPITRE 3

Enrichissement de texte court basé sur la Théorie des Ensembles Approximatifs - RST –

1 Introduction 55
2 Théorie des Ensembles Approximatifs 55
2.1 Approximation d'un Concept 55
2.1.1 Relation d'indiscernabilité 55
2.1.2 Fonction d'Appartenance 56
2.1.3 Le Système d'Information 56
2.2 Le Modèle de Tolérance des Ensembles Approximatifs (Rough Set Tolerance Model-RSTM) 59
2.2.1 L'Espace de Tolérance des Termes (Tolerance Space) 59
2.2.2 Classe de Tolérance d'un Terme (Tolerance Class) 59
2.2.3 Enrichissement de la Représentation d'un Document 60
2.2.4 Schéma de Pondération Etendu pour l'Approximation Supérieure 60
3 Méthode Proposée pour l'Enrichissement basé sur RST 64
4 Expérimentations et Résultats 66
4.1 Description des Expérimentations 66
4.2 Corpus Utilisé 67
4.3 Résultats & Discussion 69
5 Conclusion 71

1 Introduction

Dans ce chapitre, nous proposons d'enrichir la représentation de texte court en utilisant la théorie des ensembles approximatifs comme étant un outil mathématique pour traiter les données vagues et imprécises. Notre proposition sera intégrée dans un système de catégorisation des tweets arabe. L'efficacité de notre système a été évaluée et comparée en termes de f1-mesure en utilisant deux classificateurs Naïve Bayésien (NB) [87] et Support Vector Machine (SVM) [88] [89] ; les résultats obtenus montrent l'intérêt de notre contribution.

2 Théorie des Ensembles Approximatifs

2.1 Approximation d'un Concept

La théorie des ensembles approximatifs a été développée à l'origine par un informaticien polonais Zdzisław Pawlak en tant qu'un outil d'analyse et de classification des données. Elle a été appliquée avec succès dans diverses tâches, telles que la sélection / extraction de caractéristiques et la classification [91] [92]. Dans ce chapitre, nous présentons les concepts fondamentaux des ensembles approximatifs avec des exemples illustratifs. On outre, nous décrivons le modèle de tolérance dédié aux traitements de données textuelles.

Considérons un ensemble d'objet non vide U appelé l'univers. Supposons qu'on veut définir un concept sur l'univers des objets U ; et que notre concept peut être représenté comme un sous-ensemble X de U. Le point central de la théorie des ensembles approximatifs est la notion d'approximation d'ensemble : tout ensemble dans U peut être approché par son approximation inférieure et supérieure.

2.1.1 Relation d'indiscernabilité

A l'origine [92], pour définir une approximation inférieure et supérieure, il faut introduire une relation d'indiscernabilité : $R \subset U \times U$ (où R peut être n'importe quelle relation d'équivalence : réflexive, symétrique et transitive). Pour deux objets $x, y \in U$, si xRy (c'est-à-dire qu'il y a une relation entre x et y) alors on dit que x et y sont indiscernables l'un de l'autre. La relation d'indiscernabilité R conduit à une partition complète de l'univers U en classes équivalentes $[x]_R$, avec $x \in U$.

On définit l'approximation inférieure et supérieure de l'ensemble X, par rapport à un espace d'approximation désigné par A = (U ; R), respectivement comme suit :

$$L_R(X) = \{x \in U: [x]_R \subset X\} \quad (7)$$

$$U_R(X) = \{x \in U: [x]_R \cap X \neq \emptyset\} \quad (8)$$

Intuitivement, on peut déduire que l'approximation inférieure contient des objets qui appartiennent certainement à notre concept, tandis que l'approximation supérieure contient des objets qui peuvent appartenir à notre concept (Figure 5).

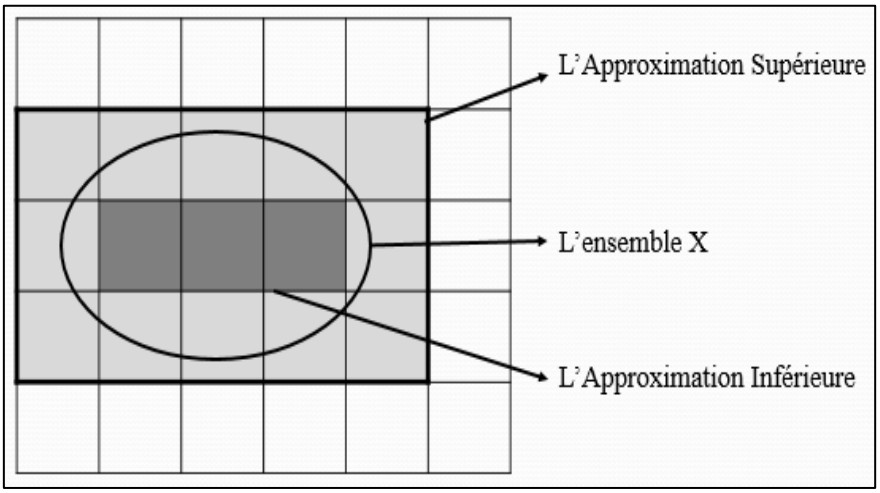


Figure 5 . La visualisation de l'approximation inférieure et supérieure d'un ensemble X

2.1.2 Fonction d'Appartenance

Les approximations peuvent aussi être définies par la fonction d'appartenance approximative [95]. Étant donné la fonction d'appartenance approximative $\mu_X : U \rightarrow [0, 1]$ d'un ensemble $X \subset U$, les approximations inférieure et supérieure sont définies comme suit :

$$L\mu(X) = \{x \in U: \mu(x, X) = 1\} \quad (9)$$

$$U\mu(X) = \{x \in U: \mu(x, X) > 0\} \quad (10)$$

Notez que, la fonction d'appartenance approximative est définie comme $\mu(x, X) = |[X]_R \cap X| / |[X]_R|$, cette définition est en réalité équivalente à (7) et (8) respectivement.

2.1.3 Le Système d'Information

Un système d'information est décrit comme une paire (U, A) ; où U est un ensemble fini et non vide d'objets appelé l'univers, A est un ensemble fini non vide des attributs qui décrivent les objets. Pour tout $a \in A$ correspond une fonction $f_a : U \rightarrow V_a$ (où V_a est l'ensemble des valeurs de a). Généralement, un système d'information est représenté par une table à deux dimensions où les lignes représentent les objets et les colonnes représentent les attributs.

Exemple Illustratif

Le tableau 9 présente un exemple de système d'information d'un univers de patients hospitalisés U = {p1, p2, p3, p4, p5, p6, p7, p8}. Les patients sont décrits par leurs symptômes de la maladie. L'ensemble des attributs est définie comme : A= {Mal à la tête, Douleur Musculaire, Température}, avec :

- $V_{\text{Mal à la tête}}$ = {Oui, Non}
- $V_{\text{Douleur Musculaire}}$ = {Oui, Non}
- $V_{\text{Température}}$ = {Normal, Élevé, Trop Élevé}

Tableau 9 . Exemple d'un Système d'Information

Patient	**Attributs**		
	Mal à la tête	**Douleur Musculaire**	**Température**
p1	Oui	Oui	Normal
p2	Oui	Oui	Élevé
p3	Oui	Oui	Trop Élevé
p4	Non	Oui	Normal
p5	Non	Non	Élevé
p6	Non	Oui	Trop Élevé
p7	Non	Non	Élevé
p8	Non	Oui	Trop Élevé

Généralement, le tableau d'information est donnée avec une colonne additionnelle *d* dite colonne de décision, dans ce cas on parle d'un *Système de Décision* : I = (U, A ∪ {d}). Au système d'information, on va ajouter une nouvelle colonne de décision {grippe} avec : V_{grippe} = {Oui, Non}. Voici la nouvelle représentation de notre système de décision (Tableau 10).

Soit une relation d'équivalence R définie par l'égalité des deux attributs 'Mal à la tête' et 'Température' ($xRy \Leftrightarrow f_{\text{Mal à la tête}}(x) = f_{\text{Mal à la tête}}(y) \wedge f_{\text{Température}}(x) = f_{\text{Température}}(y)$). La relation d'équivalence R partitionne l'univers U en des classes dont les éléments partagent les mêmes propriétés. Voici les classes obtenues : {p1}, {p2}, {p3}, {p4}, {p5, p7} et {p6, p8}. On peut voir par exemple, que p5 et p7 sont indiscernables à l'égard de relation R.

Tableau 10 . Exemple d'un Système de Décision

Patient	Attributs			Grippe
	Mal à la tête	Douleur Musculaire	Température	
p1	Oui	Oui	Normal	Non
p2	Oui	Oui	Élevé	Oui
p3	Oui	Oui	Trop Élevé	Oui
p4	Non	Oui	Normal	Non
p5	Non	Non	Élevé	Non
p6	Non	Oui	Trop Élevé	Oui
p7	Non	Non	Élevé	Oui
p8	Non	Oui	Trop Élevé	Non

Supposant qu'un concept X qu'on veut décrire est donné ici comme une décision, si un patient a la grippe. Dans ce cas le concept X sera défini comme suit : X = {p2, p3, p6, p7}. Pour déduire les approximations de X, il faut calculer pour chaque objet de U sa classe d'équivalence suivant la relation R fixée précédemment. Voici les classes d'équivalence obtenues :

- $[p1]_R = \{p1\}$
- $[p2]_R = \{p2\}$
- $[p3]_R = \{p3\}$
- $[p4]_R = \{p4\}$
- $[p5]_R = \{p5, p7\}$
- $[p6]_R = \{p6, p8\}$
- $[p7]_R = \{p5, p7\}$
- $[p8]_R = \{p6, p8\}$

Finalement, les approximations inférieure et supérieure de X selon les équations (7) et (8) seront :

$L_R(X) = \{p2, p3\}$
$U_R(X) = \{p2, p3, p5, p6, p7, p8\}$

Pour certaines applications, l'exigence d'une relation équivalente s'est révélée trop stricte. Par conséquent, un espace généralisé de tolérance est introduit par Skowron [90], permettant de passer d'une relation d'équivalence à une autre relation, dite de tolérance, où la propriété de transitivité n'est plus requise.

2.2 Le Modèle de Tolérance des Ensembles Approximatifs (Rough Set Tolerance Model - RSTM)

Le modèle de tolérance des ensembles approximatifs a été mis au point [93] [94] pour modéliser les documents et les termes dans la recherche d'information, fouille de textes, etc. Avec sa capacité de gérer l'incertitude et l'imprécision, le modèle de tolérance paraît être un bon outil pour modéliser la relation entre les termes et les documents.

2.2.1 L'Espace de Tolérance des Termes (Tolerance Space)

On considère $D = \{d_1, d_2, \ldots. d_n\}$ un ensemble de documents, $T = \{t_1, t_2, \ldots. t_m\}$ est l'ensemble des termes de D. En adoptant le modèle d'espace vectoriel, chaque document sera représenté par un vecteur de poids $[w_{i1}, w_{i2}, \ldots. w_{ij}]$, où w_{ij} représente le poids du terme j dans le document i. Dans le modèle de tolérance, l'espace de tolérance est défini sur un univers de tous les termes de la façon suivante : $U = T = \{t_1, t_2, \ldots. t_m\}$.

L'idée est de reproduire les termes indices qui sont reliés sémantiquement dans des classes. Pour ce faire, la relation de tolérance est définie comme la cooccurrence des termes indices dans tous les documents de D. Le choix de cette relation est motivé par sa capacité de prendre en considération la relation sémantique entre les termes. C'est-à-dire que, plus deux termes se reproduisent ensemble dans plusieurs document plus que ces deux termes sont liés sémantiquement.

2.2.2 Classe de Tolérance d'un Terme (Tolerance Class)

On considère $f_D(t_i, t_j)$ le nombre de documents dans D, dans lesquels apparaitront les deux termes t_i et t_j. La fonction d'incertitude *I* avec un seuil θ est définie par :

$$I_\theta(t_i) = \{t_j \mid f_D(t_i, t_j) > \theta\} \cup \{t_i\} \qquad (11)$$

Clairement, on peut déduire que la fonction ci-dessus satisfait aux deux conditions d'être réflexive ($t_i \in I_\theta(t_i)$) et symétrique ($t_i \in I_\theta(t_j) \Leftrightarrow t_j \in I_\theta(t_i)$ pour tout $t_i, t_j \in T$). Donc, I_θ (ti) est la classe de tolérance du terme index t_i.

Une classe de tolérance représente un concept caractérisé par les termes qu'il contient. En faisant varier le seuil θ (par exemple par rapport à la taille de la collection de documents), on peut contrôler le degré de liaisons des mots dans les classes de tolérance.

La fonction d'appartenance μ pout tout $t_i \in T$, $X \subseteq T$ est définie par :

$$\mu_X (t_i, X) = v (I_\theta (t_i), X) = | I_\theta (t_i) \cap X| / | I_\theta (t_i) | \quad (12)$$

Finalement, les deux approximations inférieure et supérieure de tout sous-ensemble $X \subset T$ peuvent être déterminées de la façon suivante :

$$L_R(X) = \{t_i \in T : v (I_\theta (t_i), X) = 1\} \quad (13)$$

$$U_R(X) = \{t_i \in T : v (I_\theta (t_i), X) > 0\} \quad (14)$$

Si on considère X comme un concept qui est décrit d'une façon vague par les termes qui les constitue, dans ce cas on peut dire que l'approximation supérieure est l'ensemble des termes qui partagent des liens sémantique avec X, alors que l'approximation inférieure représente le corps du concept X.

2.2.3 Enrichissement de la Représentation d'un Document

Dans le modèle d'espace vectoriel (*Vector Space Model*), un document est vu comme un sac de termes (mots) pondérés ; ceci est représenté en assignant des valeurs de poids, dans le vecteur de document, différentes de zéro pour les termes qui se produisent dans le document [108]. Avec le modèle de tolérance des ensembles approximatifs, l'objectif est d'enrichir la représentation du document en tenant compte non seulement aux termes du document, mais également aux d'autres termes avec lesquels existent des liens sémantiques.

Une représentation plus «riche» du document peut être acquise en le représentant comme un ensemble de classes de tolérance des termes qu'il contient. Ceci est réalisé simplement en représentant le document avec son approximation supérieure. Soit : $d_i = \{t_{i1}, t_{i2}, \ldots. t_{ik}\}$ un document dans l'ensemble des documents D, avec $t_{i1}, t_{i2}, \ldots. t_{ik}$ ses termes ; alors l'approximation supérieure du document d_i est :

$$U_R (d_i) = \{t_i \in T : v (I_\theta (t_i), d_i) > 0\} \quad (15)$$

2.2.4 Schéma de Pondération Etendu pour l'Approximation Supérieure

Pour construire le vecteur de poids du document il suffit d'utiliser la méthode TF*IDF ; mais afin d'utiliser les approximations du document, le système de pondération doit faire la différence entre les termes qui existent juste dans son approximation supérieure et non pas dans

le document lui-même (c'est-à-dire les termes ajoutés après le calcul de l'approximation supérieure). Le système de pondération est défini à travers l'équation (16) comme [109] :

$$w_{ij} = \begin{cases} (1 + \ln(f_{di}(t_i))) * \ln\frac{N}{f_D(t_j)} & ; t_j \in d_i \\ \min t_k \in w_{ij} * \frac{\ln(\frac{N}{f_D(t_j)})}{1+\ln(\frac{N}{f_D(t_j)})} & ; t_j \in U_R(d_i) / d_i \\ 0 & ; t_j \notin U_R(d_i) \end{cases} \quad (16)$$

Où w_{ij} est le poids de terme t_j dans le document d_i. L'extension veille à ce que les termes apparaissant dans l'approximation supérieure de d_i mais pas en d_i, ont un poids plus petit que le poids des termes du document d_i. Après le calcul des poids avec le système de pondération étendu, une normalisation par la longueur de vecteur est ensuite appliquée à tous les vecteurs des documents en utilisant l'équation (17) [109] :

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_{t_k \epsilon U_R(d_i)} (w_{ij})^2}} \quad (17)$$

Exemple Illustratif

Nous présentons un exemple illustratif du modèle de tolérance de la théorie des ensembles approximatifs (*RST*). L'objectif de ce modèle est d'enrichir la représentation des documents en tenant compte non seulement aux termes originaux d'un document, mais également d'autres termes avec lesquels existent des liens sémantiques. Pour cette raison, nous allons représenter un ensemble de documents avant et après l'application du modèle de tolérance.

Le tableau 11 présente l'espace vectoriel d'un sous ensemble de document (8 documents) dans une collection de texte court. L'espace vectoriel est représenté sous forme d'une matrice dont les colonnes correspondent aux documents, les lignes correspondent aux termes et les cellules correspondent à une relation entre un terme et un document (relation d'appartenance : c.-à-d. est-ce que un terme existe dans un document ou non).

Le tableau 12 présente le même espace vectoriel après l'enrichissement effectué à travers le modèle de tolérance. Il faut noter que, les relations avec le signe 'X' représentent les termes orignaux des documents, tandis que les relations avec le signe '✓' représentent les termes ajoutés après l'enrichissement.

Tableau 11 . Espace vectoriel avant d'appliquer le modèle de tolérance

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
سيارات voitures	X			X			X	X
البيع Vente								
تجمع Rassemblement	X							
الصور Photos							X	
نمر Tigre		X			X			
القوة Puissance		X						
سر Secret		X						
العجيبة Étrange		X						
الأسود Lions			X		X	X		
العاب Jeux				X				
قطع Pièces								X
يجد Trouver							X	
مُسن Agé							X	
غيار Rechange								X

Par exemple, le document numéro 2 contenant le terme « tigre / نمر » parmi autres termes, l'utilisation du modèle de tolérance permet d'ajouter le terme « lions / الأسود», nous pouvons remarquer la forte relation sémantique existante entre les deux termes comme étant deux animaux très proches dans leurs structure physique et les proies préférées. Le même traitement a été appliqué aussi dans le document 6. Dans le document 8, nous remarquons que le terme « Vente / البيع» a été ajouté aux contenu initial du document dans lequel nous trouvons les termes « Voitures / سيارات», «Pièces / قطع» et « Rechange / غيار» ; ces termes sont fréquemment utilisés ensemble ce qui montrent l'intérêt de l'utilisation de l'enrichissement basé sur le modèle de tolérance.

Tableau 12 . Espace vectoriel après avoir appliqué le modèle de tolérance

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
سيارات Voitures	X			X			X	X
البيع Vente	✓			✓			✓	✓
تجمع Rassemblement	X							
الصور Photos	✓			✓			X	✓
نمر Tigre		X	✓		X	✓		
القوة Puissance		X						
سر Secret		X						
العجيبة Étrange		X						
الأسود Lions		✓	X		X	X		
العاب Jeux				X				
قطع Pièces								X
يجد Trouver							X	
مُسن Agé							X	
غيار Rechange								X

Avantage du modèle de tolérance

Les avantages principaux du modèle de tolérance des ensembles approximatifs sont les suivantes :

- Elle n'a pas besoin des informations préliminaires ou complémentaires sur les données, comme par exemple la probabilité dans les statistiques
- Elle fournit des méthodes efficaces, des algorithmes et des outils pour trouver des modèles cachés dans les données
- Il offre une interprétation simple des résultats obtenus

- Elle permet de réduire les données d'origine en calculant les sous-ensembles d'attributs, appelés réductions, qui préservent la qualité des données du système d'information de départ.

Dans la prochaine section, nous allons voir comment nous pouvons bénéficier du modèle de tolérance dans l'enrichissement de texte court en arabe.

3 Méthode Proposée pour l'Enrichissement basé sur RST

Dans cette section, nous présentons en détail notre méthode pour enrichir la représentation du texte court en se basant sur la théorie des ensembles approximatifs. Cet enrichissement tend à remplacer la représentation d'un document avec son approximation supérieur en utilisant d'autres termes du corpus avec lesquels existent des liens sémantiques.

La figure 6 présente un organigramme de notre méthode pour l'enrichissement de texte court, elle contient deux composants principaux :

- Le bloc de prétraitement où chaque texte court dans le corpus sera nettoyé en supprimant les mots latins et les caractères spéciaux comme (/, #, $, etc.…), par la suite les mots vides en arabe seront aussi éliminés. Les mots vides ont une fréquence élevée et une faible discrimination et doivent être filtrés. Vers la fin un algorithme de racinisation ou de lemmatisation peut être appliqué pour réduire les mots à leurs radicaux (stem) / ou racines (root). La sortie de ce composant est une représentation vectorielle du document dans laquelle nous proposons de garder que les termes, nommés, indices.
- Le bloc d'enrichissement où nous allons appliquer la théorie des ensembles approximatifs en passant par les étapes suivantes :
  - Déterminer l'ensemble des termes du corpus
  - Calculer la fréquence de chaque terme dans le document ainsi que sa fréquence totale dans tout le corpus
  - Générer pour chaque terme sa classe de tolérance contenant les termes qui se produisent ensemble un nombre de fois supérieur à un seuil bien fixé. La classe de tolérance est déterminée à l'aide de la formule (11).
  - Obtenir l'approximation supérieure de chaque document à l'aide de la formule (15).

L'utilisation de l'approximation supérieure permet de renforcer les liens sémantiques entre les documents en enrichissant leurs représentations à partir du même corpus sans faire appel à aucune ressource externe.

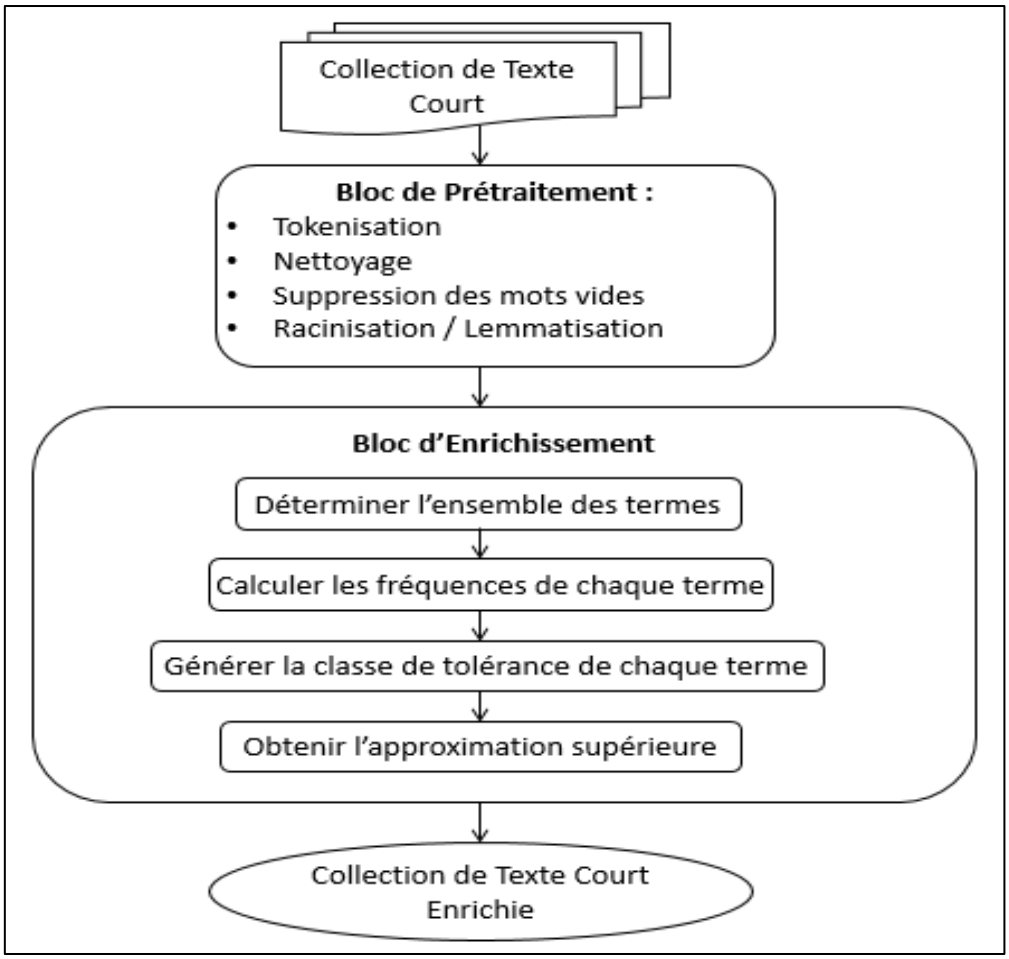


Figure 6 . Méthode Proposée d'Enrichissement de texte court basée sur RST

Pour illustrer les liens sémantiques découverts par l'approximation supérieure, la figure 7 présente un tweet avec ses termes initiales contenant le mot (« spectacle » / «عرض»). L'étape d'enrichissement a ajouté deux termes, l'un des deux est (« cinéma » / «سينما»), nous pouvons remarquez la liaison sémantique qui existe entre les deux termes « spectacle » et « cinéma ».

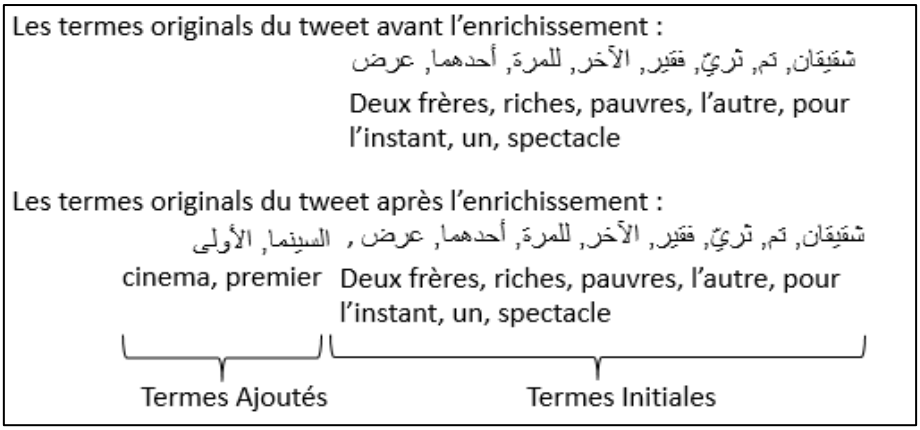


Figure 7 . Exemple illustratif de l'enrichissement

Finalement, chaque tweet sera représenter par un vecteur de poids qui reflète le degré d'importance de chaque terme dans le tweet, toute en prenant en considération que le système de pondération fait la différence entre les termes qui existent seulement dans l'approximation supérieure et non pas dans le tweet lui-même. Le système de pondération est défini à l'aide de la formule (16).

4 Expérimentations et Résultats

Dans cette section, nous présentons une description détaillée des expérimentations que nous avons mené afin de valide notre contribution, ainsi que le corpus utilisé et les résultats obtenus.

4.1 Description des Expérimentations

Pour montrer l'intérêt de notre proposition que l'utilisation de l'approximation supérieure améliore la représentation de texte court et par conséquent améliore la performance de n'importe quel système de fouille de texte, une série d'expérimentations a été menée. Pour cette raison, nous avons construit un système de catégorisation (classification) des tweets en arabe.

Les systèmes de catégorisation de texte tentent de trouver une relation entre un ensemble de textes et un ensemble de catégories (classes). L'apprentissage automatique est l'outil qui permet de décider si un texte appartient à un ensemble de catégories prédéfinies. Le classificateur est construit en apprenant les caractéristiques de chaque catégorie à partir d'un ensemble de documents d'apprentissage. Après avoir construit le classificateur, son efficacité est testée en l'appliquant à l'ensemble de test et vérifie le degré de correspondance entre les deux ensembles.

Figure 8 décrit nos expérimentations où nous avons testé dans un premier temps la catégorisation des tweets avec la représentation ordinaire (c'est-à-dire sans enrichissement) ; et pour valider notre proposition, nous avons effectué la catégorisation en utilisant les tweets enrichis.

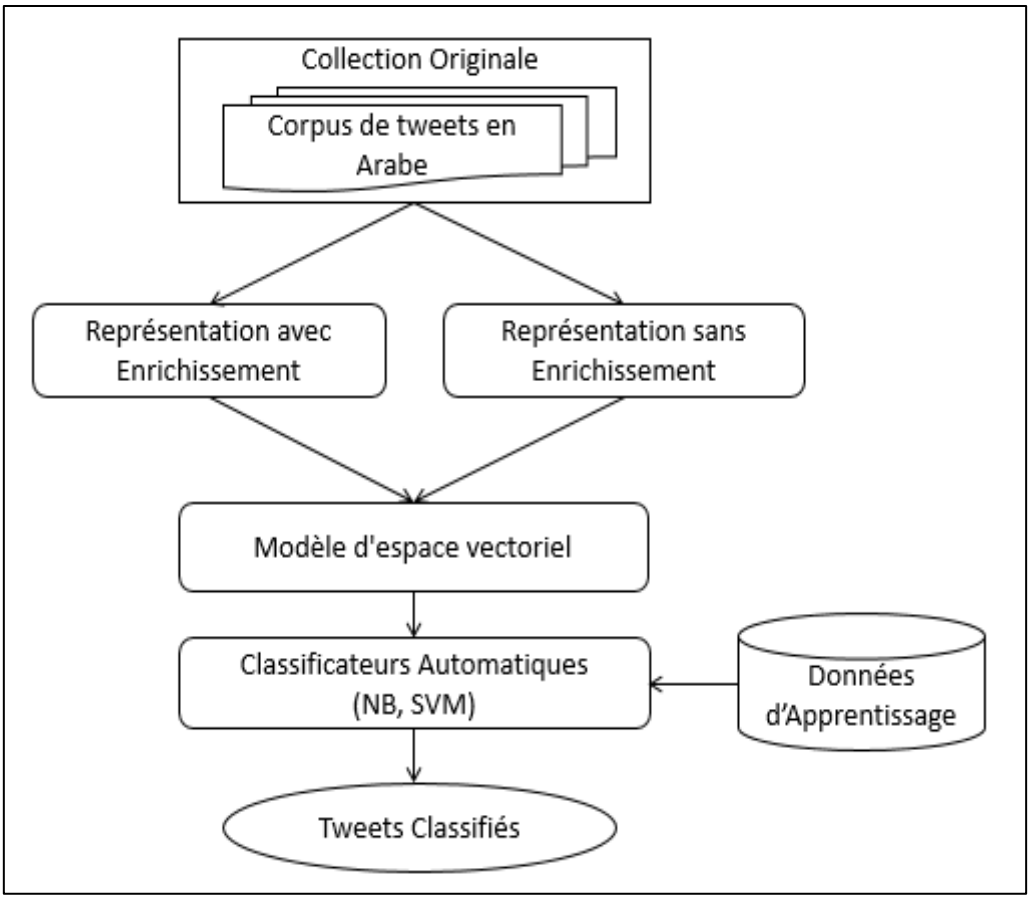


Figure 8 . Description des expérimentations de l'enrichissement en utilisant RST

4.2 Corpus Utilisé

Le corpus utilisé dans nos expérimentations est collecté à partir de Twitter en utilisant un outil libre *NodeXL Excel Template*[27] qui rend la collecte des tweets super facile. Le tableau 13 présente quelques exemples des tweets collectés. Ce corpus est classé manuellement en six catégories (Tableau 14).

L'efficacité de notre système a été évaluée et comparée en termes de la mesure f1-mesure à l'aide de deux classificateurs NB [87] et SVM [88] [89].

F1-measure peut être calculée en utilisant la précision et le rappel comme suit :

$$\text{F1-measure} = 2 * ((\text{Précision} * \text{Rappel}) / (\text{Précision} + \text{Rappel})) \quad (18)$$

La Précision et Rappel sont définis respectivement avec les deux formules :

$$\text{Précision} = TP / ((TP+FP)) \quad (19)$$

$$\text{Rappel} = TP / ((TP+FN)) \quad (20)$$

[27] http://social-dynamics.org/twitter-network-data/

Avec :

- TP : fait référence à l'ensemble des tweets correctement affectés à la catégorie donnée.
- FP : fait référence à l'ensemble des tweets qui sont incorrectement affectés à la catégorie.
- FN : fait référence à l'ensemble des tweets qui ne sont pas correctement affectés à la catégorie.

Tableau 13 . Exemples des tweets collectés

Catégories	Tweets en Arabe	Tweets en Français
Cinéma / سينما	تغرق في عالم يكتنفه الجنون والقتل والقوى الخارقة للطبيعة	Elle s'enfonce dans un monde de folie, de meurtre et de pouvoirs surnaturels
	القصة الحقيقية لمخطط اغتيال الرئيس ريتشارد نيكسون	L'histoire vraie du complot visant à assassiner le président Richard Nixon
Documentaire وثائقي/	تحت المجهر "حرب السدود": نهر النيل الآن على الجزيرة الوثائقية	Au microscope "La guerre des barrages" : le Nil est maintenant sur Al Jazeera documentaire
	بقي 6 أيام ويعرض البرنامج العالمي "الكون الكبير" على قناة ناشونال جيوغرافيك: أبو ظبي	Il reste 6 jours pour montrer le programme mondial "The Big Universe" sur la chaîne National Geographic : Abu Dhabi
Economie / اقتصاد	صعود الذهب إلى أعلى مستوياته ثلاثة أسابيع ونصف بعد موافقة واشنطن على توجيه جوية على العراق	La montée de l'or à son plus haut niveau, trois semaines et demi après l'approbation de Washington pour diriger des vols sur l'Irak
	المبيعات الإجمالية لشركات الاسمنت تتراجع بنسبة 2% مقارنة بالشهر الماضي	Les ventes totales des cimenteries diminuent de 2% par rapport au mois dernier
Santé / الصحة	ذكرنا أعراض وعلامات السكتة القلبية في تغريدتنا السابقة في تمام الساعة 10 صباحا	Nous avons mentionné des symptômes et des signes de crise cardiaque dans le précédent Tweet à 10h00
	أخصائية تغذية لعمل نظام غذائي مرتفع السعرات ومناسب لك	Un diététicien fait un régime riche en calories pour vous
Actualité / أخبار	سقوط قذيفة هاون افضى إلى إصابة محافظ الأنبار بجروح بليغة	Un obus de mortier conduit à une blessure grave du gouverneur d'Anbar
	حوارات وتقارير: في تونس سباق محموم للظفر بمنسب الرئيس	Interviews et reportages : En Tunisie, une course fébrile pour remporter le poste de président
Tourisme / السياحة	الأماكن السياحية في تركيا: التفاصيل	Places touristiques en Turquie : Détails
	منتجع شاموني يعتبر أحد أكثر المنتجعات شهرة في فرنسا	Chamonix Resort est l'une des stations les plus célèbres de France

Tableau 14 . Description du corpus

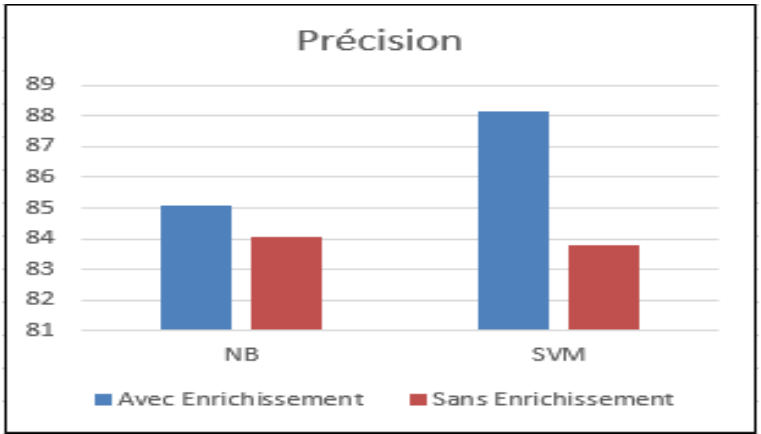
Catégorie	Nombre de Tweets
Cinéma / سينما	79
Documentaire / وثائقي	64
Economie / اقتصاد	63
Santé / الصحة	71
Actualité / أخبار	90
Tourisme / السياحة	83
Total	**450**

Le corpus a été divisé en deux parties : un corpus d'apprentissage (avec 70% des tweets par catégorie) et un autre corpus de test (avec 30% des tweets par catégorie).

4.3 Résultats & Discussion

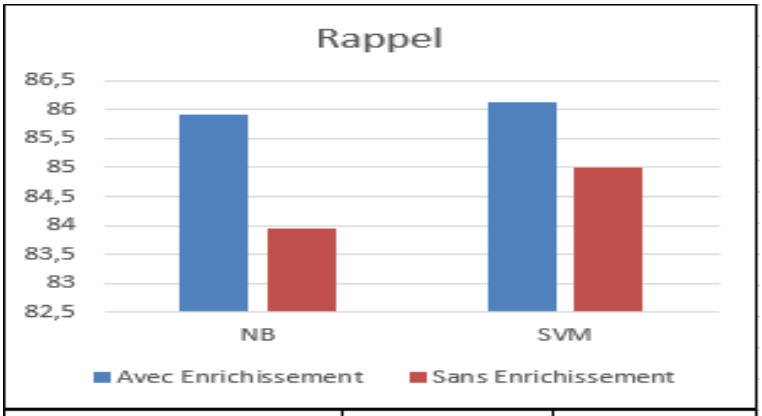
Les figures 9, 10 et 11 présentent les résultats obtenus de précision, rappel et f1-measure respectivement. Ces résultats montrent bien que l'enrichissement des tweets en utilisant l'approximation supérieure conduit à l'amélioration des résultats de la catégorisation. L'application de la théorie des ensembles approximatifs permet de renforcer les liens sémantiques entre les tweets en ajoutant des termes qui se produisent fréquemment avec les termes orignaux d'un tweet.

De plus, notre méthode pour enrichir la représentation des tweets en arabe a été testée en utilisant deux classificateurs NB et SVM. Les résultats obtenus avec les deux classificateurs montrent l'intérêt de notre contribution.



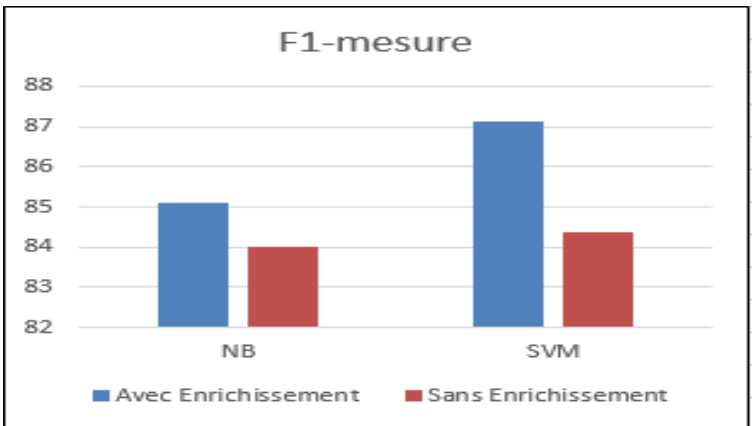
Précision	NB	SVM
Avec Enrichissement	85,06	88,16
Sans Enrichissement	84,04	83,77

Figure 9 . Précision de l'enrichissement basé RST



Rappel	NB	SVM
Avec Enrichissement	85,93	86,13
Sans Enrichissement	83,96	85,01

Figure 10 . Rappel de l'enrichissement basé RST



F1-mesure	NB	SVM
Avec Enrichissement	85,09	87,13
Sans Enrichissement	84	84,39

Figure 11 . F1-mesure de l'enrichissement basé RST

5 Conclusion

Dans ce chapitre, nous avons proposé d’enrichir la représentation de texte court en utilisant la théorie des ensembles approximatifs (RST) comme étant un outil mathématique pour traiter les données vagues et imprécises. Afin de valider notre proposition, un système de catégorisation des tweets arabe a été construit en intégrant notre représentation enrichie de texte court. L'efficacité de notre système a été évaluée et comparée en utilisant deux classificateurs NB et SVM. Les résultats obtenus sont très encourageants et misent en évidence l’intérêt de notre contribution.

CHAPITRE 4

Enrichissement de Texte Court basé sur le Corpus WEB

1 Introduction 73
2 Méthode Proposée pour l'Enrichissement basé sur le Web 73
3 Expérimentations et Résultats 1 77
3.1 Description des Expérimentations 77
3.2 Corpus Utilisé 79
3.3 Résultats & Discussion 80
4 Expérimentations et Résultats 2 83
4.1 Description des Expérimentations 83
4.3 Résultats & Discussion 85
5 Conclusion 86

1 Introduction

Dans ce chapitre, nous proposons une deuxième contribution pour enrichir la représentation de texte court en utilisant le corpus Web comme étant un grand corpus ouvert et dynamique avec la théorie des ensembles approximatifs (RST) en tant qu'un outil mathématique pour le traitement des données vagues et imprécises. Pour mettre en évidence l'intérêt de notre proposition, une série d'expérimentations a été réalisé à travers un système de catégorisation des tweets en arabe avec trois classificateurs NB, SVM et les arbres de décisions (Decision Tree - DT -); les résultats obtenus sont très encourageants.

2 Méthode Proposée pour l'Enrichissement basé sur le Web

Dans cette section, nous présentons en détail notre méthode pour enrichir la représentation des textes courts en arabe en se basant d'une part, sur le corpus Web comme une ressource externe ouvert et dynamique. En utilisant le grand volume de documents sur le Web, nous pouvons enrichir le texte court avec d'autres termes dans le même contexte que les termes d'origine en bénéficiant des snippets renvoyés par le moteur de recherche. D'autre part, afin de renforcer les liens sémantiques existants entre le texte original et les snippets retournés par le moteur de recherche Web et de n'ajouter au texte court que les termes les plus proches sémantiquement, nous faisons appel à la théorie des ensembles approximatifs.

La figure 12 présente un organigramme de notre méthode pour l'enrichissement de texte court. Elle contient deux composants principaux :

Le composant de prétraitement, comme c'est déjà mentionné au chapitre 3, où chaque texte dans le corpus sera nettoyé en supprimant les mots latins et les caractères spéciaux comme (/, #, $, etc.…), par la suite les mots vides en arabe seront aussi éliminés. Les mots vides ont une fréquence élevée et une faible discrimination et doivent être filtrés. Vers la fin de cette étape, un algorithme de racinisation ou de lemmatisation peut être appliqué pour réduire les mots à leurs radicaux (stem) / ou racines (root). La sortie de ce composant est une représentation vectorielle du texte dans laquelle nous gardons que les termes nommés les termes indices.

Le composant d'enrichissement présente le processus d'enrichissement de la représentation du texte. Dans notre contribution, nous enrichissons la représentation d'un texte court par des termes supplémentaires existant dans les N-Top snippets renvoyés par deux des moteurs de

recherche Web les plus populaires, Google[28] et Bing[29], en envoyant le contenu du texte court en tant que requête au moteur de recherche Web.

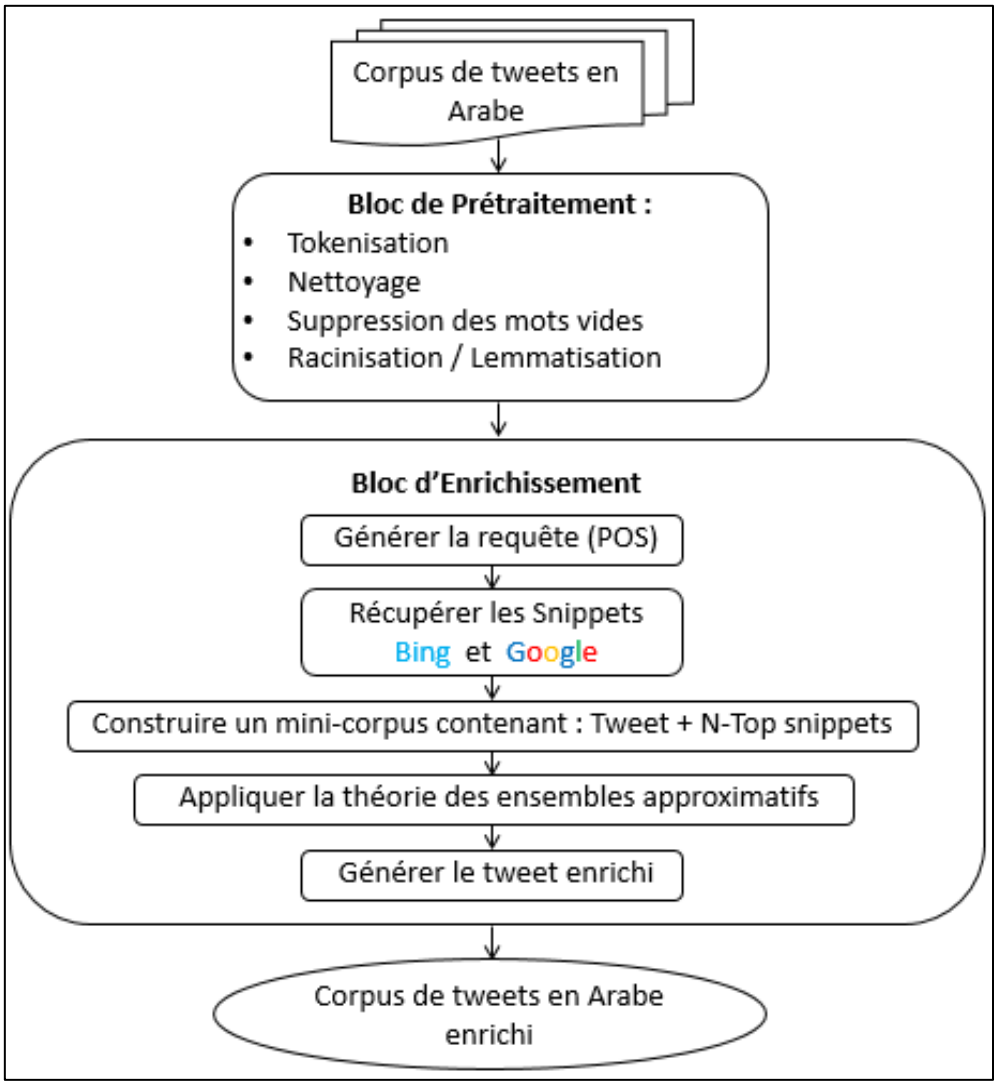


Figure 12 . Méthode d'Enrichissement de texte court basée sur le corpus Web

Pour chaque texte court dans le corpus nous appliquer le traitement suivant :

- La plupart des textes courts contiennent entre 10 et 30 termes ; Si nous envoyons l'intégralité du contenu du texte en tant que requête, le moteur de recherche Web risque de renvoyer une liste vide. Pour surmonter ce problème, nous proposons d'utiliser un analyseur morphologique pour extraire seulement les mots nominaux vu qu'ils sont considérés comme les termes les plus discriminants. Dans notre méthode nous avons utilisé l'analyseur morphologique « Stanford POSTagger » [96].

[28] www.google.com
[29] www.bing.com

- Récupérer les N-Top snippets retournés par le moteur de recherche en envoyant les termes nominaux issue par l'analyseur morphologique comme requête. Il faut noter que nous avons établi un mécanisme d'éliminer la redondance en fonction de l'URL des snippets retourné par les deux moteurs de recherche. Les snippets retournés sont de la forme : *Id, body et title* (Figure 13).
- Construire un mini-corpus formé du texte court original et les N-Top snippets récupérés par les moteurs de recherche.
- Appliquer le processus d'enrichissement sur le mini-corpus issu de l'étape précédente, ce processus est basé sur la théorie des ensembles approximatifs en calculant les approximations supérieures des éléments du mini-corpus à l'aide de la formule (15).



Figure 13 . Exemple d'un Snippet en langue arabe

A la fin de ce traitement, nous pouvons garantir que la représentation des textes courts sera enrichie avec d'autres termes liés sémantiquement avec le contenu initial. Parce que en envoyant les termes les plus discriminatifs du texte court original, le moteur de recherche va renvoyer des résultats liés à la requête ; de l'autre côté, en utilisant la théorie des ensembles approximatifs, nous pouvons sélectionner de ces snippets les termes ayant une forte liaison sémantique avec le contenu original. L'algorithme du processus d'enrichissement peut être résumé et présenté ci-dessous (algorithme 1).

Après avoir appliqué l'algorithme 1 pour tous le corpus initial, nous allons obtenir un nouveau corpus enrichi prêt à être traité.

Finalement, chaque texte court sera représenté par un vecteur de poids qui reflète le degré d'importance de chaque terme. Il faut noter que le système de pondération fait la différence

entre les termes qui existent seulement dans l'approximation supérieure et non pas dans le texte court lui-même. Le système de pondération est défini à l'aide de la formule (16).

Algorithme 1 : Méthode d'Enrichissement Proposé

```
Début
    Initialiser le corpus des tweets dt
    Initialiser le corpus des tweets enrichi comme étant un ensemble vide ds
    Pour chaque tweet twi dans le corpus dt
        requete ← séléctionnezLesTermesNominaux(twi)
        List<Snippets> res ← NTopRésultats(requete)
        Initialiser un mini-Corpus u = twi U res
        genererApproximationSuperieur(twi, u)
        ajouter Approximattion Superieur de twi au Corpus Enrichi ds
    Fin Pour
fin
void genererApproximationSuperieur (tw, u)
Début fonction
    Prétaitement du corpus u
    Initialiser l'ensemble des termes s
    Initialiser la matrice des co-occurrences
    Pour chaque terme ti dans s do
        Pour chaque terme tj dans s do
            Si le nombre de co-occurrence de ti et tj > θ alors
                ajouter tj à la classe de tolérance de ti
            fin si
        Fin Pour
    Fin Pour
    Pour chaque terme tj dans s do
        tct = la classe de tolerance de tj
        cof = |tct ∩ tw| / |tct|
        Si cof > 0 alors
            Ajouter tj à l'approximation supérieur de tw
        fin si
    Fin Pour
Fin fonction
```

Le tableau 15 présente deux exemples pour illustrer les liens sémantiques découverts après l'enrichissement. Le premier exemple présente un tweet sur (l'économie / 'اقتصاد') qui contient le terme ('huile' / 'النفط'). Le terme ('ministre' / 'وزير') a été ajouté à l'approximation supérieure générée, les deux termes sont utilisées ensembles dans certains contextes.

Le deuxième exemple présente un tweet sur (l'éducation / 'التعليم'') qui contient le terme ('ministère' / 'وزارة'), les deux termes ('supérieur' / 'العالي') et ('éducation' / 'التعليم') ont été ajoutés à l'approximation supérieur qui fait référence au ministère de l'enseignement supérieur.

Dans le but de valider notre proposition, nous avons conduit une série d'expérimentations en utilisant un système de catégorisation des tweets en langue arabe. La description des expérimentations avec les résultats obtenus seront discuté dans la section prochaine.

Tableau 15 . Exemples des tweets après l'enrichissement basé sur le Web

Catégorie	Contenu de Tweet		Termes initiales	Termes Ajoutés
Economie / الاقتصاد	Arabe	صعود الذهب إلى أعلى مستوياته ثلاثة أسابيع ونصف بعد موافقة واشنطن على توجيه جوية على العراق	صعود، الذهب، أعلى، مستويات، ثلاثة، أسابيع، واشنطن، نصف، موافقة، ضربات، جوية، العراق	عالم، وزير
	Traduction en Français	La montée de l'or à son plus haut niveau, trois semaines et demi après l'approbation de Washington pour diriger des vols sur l'Irak	Montée, or, plus, niveaux, trois, Washington, semaines, demi, accord, coups, air, Iraq	Monde, ministre
Education / التعليم	Arabe	الوزارة تنفذ برنامج منسوبيها لذوي الاحتياجات الخاصة	وزارة، تنفذ، برنامج، منسوبيها، ذوي، الاحتياجات، خاص	تعليم، العالي
	Traduction en Français	Le ministère met en œuvre son programme d'employés ayant des besoins spéciaux	Ministère, met en œuvre, programme, employées, avec, besoins, spécial	Education, supérieur

3 Expérimentations et Résultats 1

Dans cette section, nous présentons un premier cas de figure des expérimentations que nous avons mené afin de valider notre contribution, ainsi que les corpus utilisés et les résultats obtenus.

3.1 Description des Expérimentations

Dans notre contribution nous avons utilisé le Web, un corpus ouvert et dynamique, comme une ressource externe et l'approximation supérieure comme étant un moyen pour sélectionner les termes les plus proches sémantiquement aux termes initiales du tweet. Pour montrer que notre proposition améliore la représentation des tweets en Arabe, une série d'expérimentations a été menée. Pour cette raison, nous avons construit un système de catégorisation des tweets en arabe. Par ailleurs, nous voulons tester l'impact que le nombre des snippets retournés par les moteurs de recherche peut importer à l'enrichissement des tweets. C'est pour cette raison, nous avons fait les expérimentations avec différent nombres de snippets (N-Top résultats prend les valeurs 10, 20, 30, 40 et 50 respectivement). En revanche, pour récupérer les snippets, la requête sera

envoyée aux moteurs de recherche Web Google et Bing en utilisant les services offerts par l'API Google[30] et l'API Bing[31].

Figure 14 décrit les expérimentations que nous avons effectué, premièrement nous avons testé la catégorisation avec une représentation ordinaire des tweets (c'est-à-dire sans enrichissement) ; par la suite nous avons effectué la catégorisation avec l'enrichissement et utilisant les différentes valeurs des snippets discutés précédemment (c'est-à-dire que nous avons générer cinq corpus enrichi correspond aux cinq valeurs des N-top snippets retournés par les moteurs de recherche comme il est décrit dans le tableau 16).

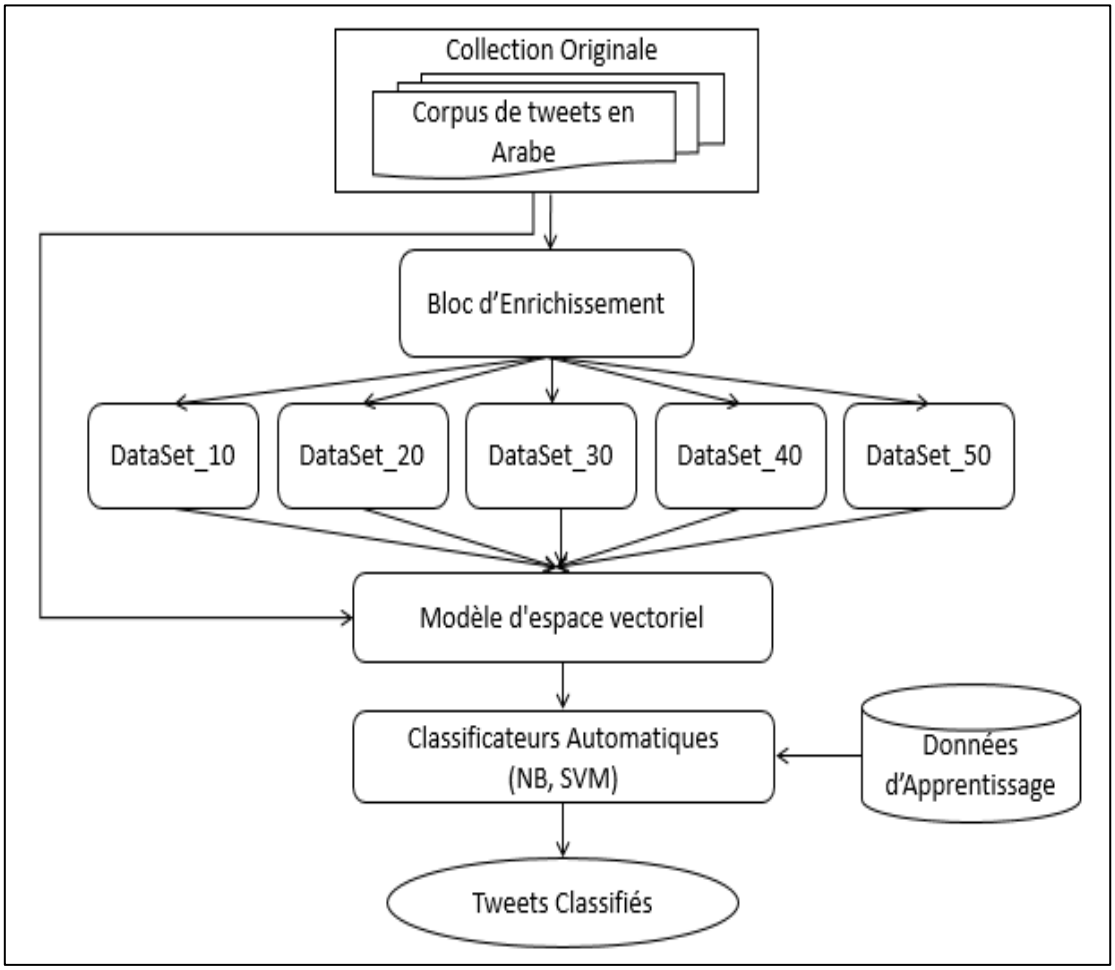


Figure 14 . Description des expérimentations de l'enrichissement basé sur le Web

[30] https://developers.google.com/custom-search/docs/start
[31] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44

Tableau 16 . Les corpus enrichis générés

Nombre de snippets	Nom de Corpus
10	Dataset_10
20	Dataset_20
30	Dataset_30
40	Dataset_40
50	Dataset_50

3.2 Corpus Utilisé

Le corpus utilisé dans nos expérimentations est collecté à partir de Twitter en utilisant l'outil *NodeXL Excel Template*[32] qui rend la collecte des tweets super facile (comme c'est déjà mentionné dans le chapitre 3). Ce corpus est classé manuellement en huit catégories (Tableau 17).

L'efficacité de notre système a été évaluée et comparée en termes de la mesure de f1-mesure (définie précédemment dans ce rapport au chapitre 3) à l'aide de deux classificateurs NB et SVM.

Le corpus a été divisé en deux parties : un corpus d'apprentissage (avec 70% des tweets par catégorie) et un autre de test (avec 30% des tweets par catégorie).

Tableau 17 . Description du corpus

Catégorie	Nombre de Tweets
Cinéma / سينما	130
Documentaire / وثائقي	139
Education / التعليم	147
Economie / اقتصاد	131
Santé / الصحة	132
Actualité / أخبار	120
Sport / الرياضة	139
Tourisme / السياحة	125
Total	**1063**

[32] http://social-dynamics.org/twitter-network-data/

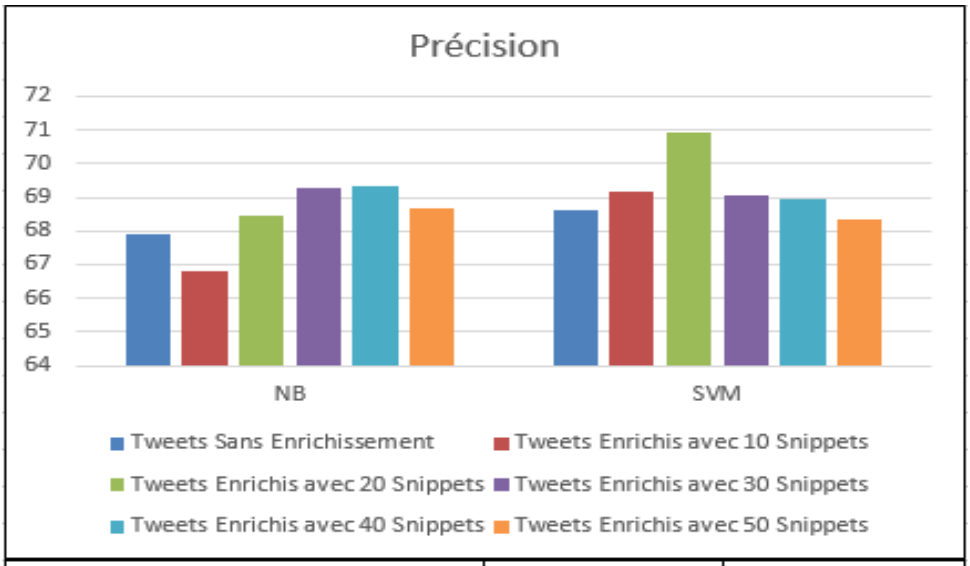
3.3 Résultats & Discussion

Les figures 15, 16 et 17 présentent les résultats obtenues de précision, rappel et f1-measure respectivement. Nous pouvons noter les remarques suivantes :

En utilisant le classificateur SVM par exemple : la précision avec les tweets originaux est de 68, 64%. En revanche, lorsque nous appliquons notre méthode pour enrichir leur contenu, la précision obtenue est de 69, 19% en utilisant les 10-Top snippets, 70, 91% (avec 20-Top snippets), 69, 05% (avec 30-Top snippets), 68, 95% (40-Top snippets) et 68, 33% (50-Top snippets). Nous remarquons que l'amélioration de la représentation des tweets a un impact positif sur la précision de notre système de catégorisation des tweets en arabe. Ces résultats démontrent que les 20 premiers snippets contiennent des termes plus pertinents qui ont une influence positive sur le processus de catégorisation.

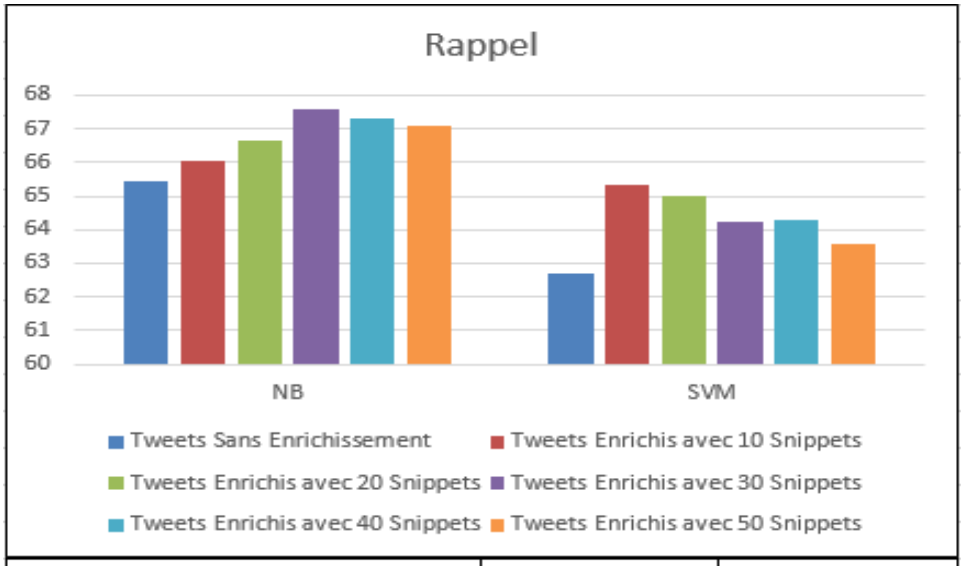
Le rappel avec les tweets originaux est de 62, 67%. Avec l'application de notre méthode d'enrichissement, le rappel obtenu est de 65, 35% en utilisant les 10-Top snippets, 64, 98% (avec 20-Top snippets), 64, 24% (avec 30-Top snippets), 65, 28% (avec 40-Top snippets) et 63, 56% (avec 50-Top snippets); Ces résultats montrent que l'ajout des termes contextuels permet de mieux situer les tweets d'origine et, par conséquent, de réduire l'effet du problème de texte court et le manque d'information contextuelle.

La f1-mesure avec les tweets originaux est de 61,7%. Lorsque nous appliquons notre méthode pour enrichir leur contenu, la f1-mesure obtenue est de 63, 94% en utilisant 10-Top snippets, 64, 39% (avec 20-Top snippets), 63, 48% (avec 30-Top snippets), 64, 19% (avec 40-Top snippets) et 62, 66% (avec 50-Top snippets). Nous pouvons conclure que l'enrichissement de la représentation des tweets par des termes supplémentaires provenant du Web améliore la performance de notre système de catégorisation des tweets en arabe.



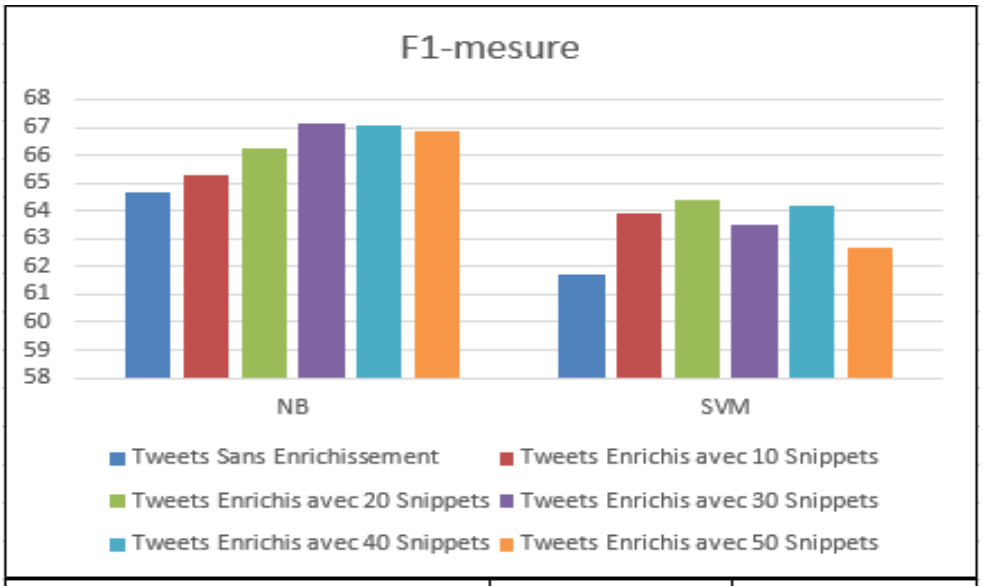
Précision	NB	SVM
Tweets Sans Enrichissement	67,92	68,64
Tweets Enrichis avec 10 Snippets	66,78	69,19
Tweets Enrichis avec 20 Snippets	68,44	70,91
Tweets Enrichis avec 30 Snippets	69,26	69,05
Tweets Enrichis avec 40 Snippets	69,32	68,95
Tweets Enrichis avec 50 Snippets	68,7	68,33

Figure 15 . Précision de l'enrichissement basé sur le Web



Rappel	NB	SVM
Tweets Sans Enrichissement	65,42	62,67
Tweets Enrichis avec 10 Snippets	66,07	65,35
Tweets Enrichis avec 20 Snippets	66,66	64,98
Tweets Enrichis avec 30 Snippets	67,6	64,24
Tweets Enrichis avec 40 Snippets	67,33	64,28
Tweets Enrichis avec 50 Snippets	67,09	63,56

Figure 16 . Rappel de l'enrichissement basé sur le Web



F1-mesure	NB	SVM
Tweets Sans Enrichissement	64,69	61,7
Tweets Enrichis avec 10 Snippets	65,3	63,94
Tweets Enrichis avec 20 Snippets	66,25	64,39
Tweets Enrichis avec 30 Snippets	67,12	63,48
Tweets Enrichis avec 40 Snippets	67,05	64,19
Tweets Enrichis avec 50 Snippets	66,85	62,66

Figure 17 . F1-mesure de l'enrichissement basé sur le Web

Notre méthode d'enrichir la représentation des tweets en Arabe a été testée aussi en utilisant le classificateur NB. Les résultats obtenus avec les deux classificateurs démontrent l'intérêt de notre contribution.

D'un autre côté, dans nos expérimentations nous avons utilisé cinq valeurs de snippets retournés par les moteurs de recherche en raison d'étudier l'influence de ce nombre sur l'enrichissement des tweets. Nous pouvons noter les remarques suivantes :

- L'enrichissement avec les différents nombres de snippets utilisé conduit à l'amélioration des résultats de catégorisation
- La variation des résultats par rapport au nombre de snippets utilisé est relativement minimale : avec le classificateur NB (respectivement SVM), le résultat minimale est 65,3 % et c'était avec 10 snippets (respectivement 62,66% et c'était avec 50 snippets), tandis que le résultat maximale est 67, 12% et c'était en utilisant 30 snippets (respectivement 64,39% et c'était avec 20 snippets)
- Les meilleurs résultats sont obtenus généralement avec un nombre de snippets entre 20 et 30, parce qu'au-delà de ce nombre les snippets retournés par les moteurs de recherche seront légèrement liés à la requête et peuvent présenter un peu de bruit.

Dans la section suivante, nous présentons d'autres expérimentations que nous avons mené afin de comparer les deux propositions pour faire l'enrichissement : l'un à partir de collection de données elle-même à travers RST et l'autre en utilisant le corpus Web.

4 Expérimentations et Résultats 2

Dans cette section, nous présentons un deuxième cas de figure des expérimentations que nous avons mené afin de comparer les deux modes d'enrichissement, ainsi que les corpus utilisés et les résultats obtenus.

4.1 Description des Expérimentations

Dans la section précédente, nous avons testé notre proposition d'enrichir la représentation des tweets en arabe en se basant sur le Web et la théorie des approximations approximatifs (RST). Les résultats obtenus ont montré l'amélioration de la performance de catégorisation des tweets. Dans cette section, nous nous sommes intéressés par l'impact de l'utilisation du Web sur l'enrichissement des tweets. C'est pour cette raison, nous allons mener une série d'expérimentations où l'enrichissement sera fait seulement avec la théorie des ensembles approximatifs, de l'autre côté sera fait en impliquant le Web comme une ressource externe.

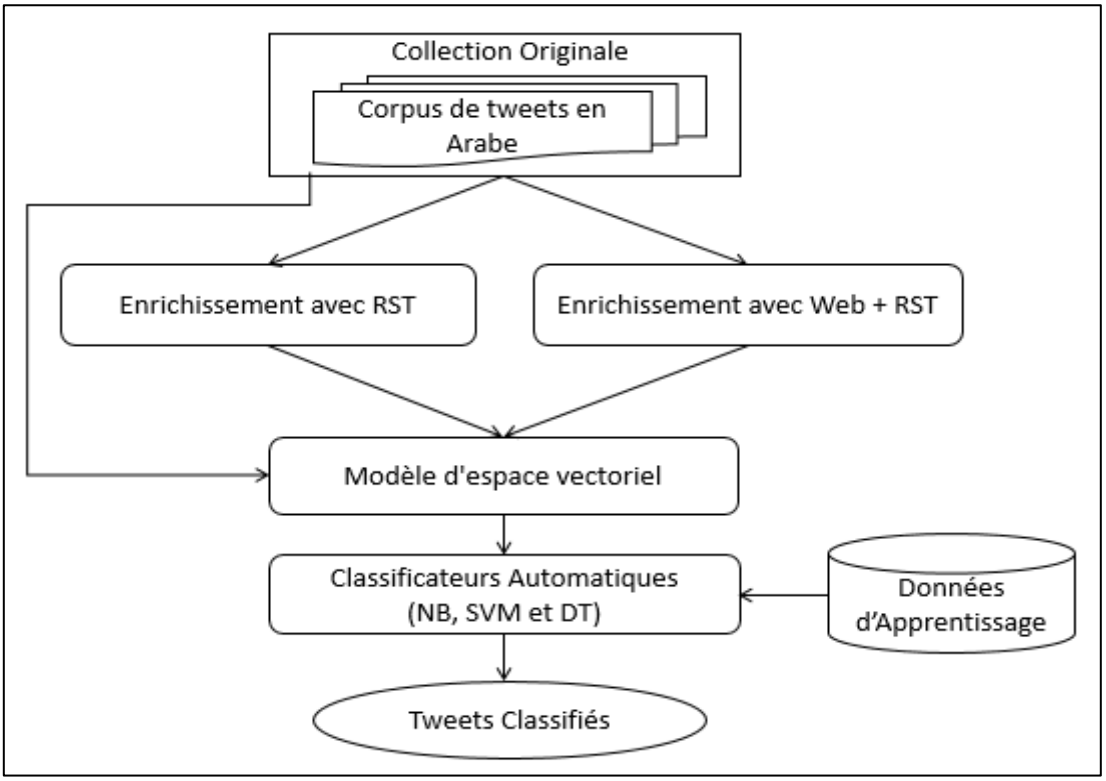


Figure 18 . Description des expérimentations de l'enrichissement basé sur le Web comparé avec l'enrichissement basé sur RST

La figure 18 décrit les étapes que nous avons effectuées dans nos expérimentations. Premièrement, la catégorisation se fait avec une représentation ordinaire des tweets (sans enrichissement) ; ensuite nous avons généré deux corpus enrichis, le premier en utilisant la RST et le second est basé sur le Web.

4.2 Corpus Utilisés

Les tweets utilisés sont collectés toujours avec l'outil *NodeXL Excel Template*. Nous avons construit deux corpus classés manuellement chacun en huit catégories (tableau 18).

L'efficacité de notre système a été évaluée et comparée en termes de la mesure f1-mesure à l'aide de trois classificateurs NB [87], SVM [88] [89] et les arbres de décision (DT) [98].

Le corpus a été divisé en deux parties : un corpus d'apprentissage (avec 75% des tweets par catégorie) et un autre de test (avec 25% des tweets par catégorie).

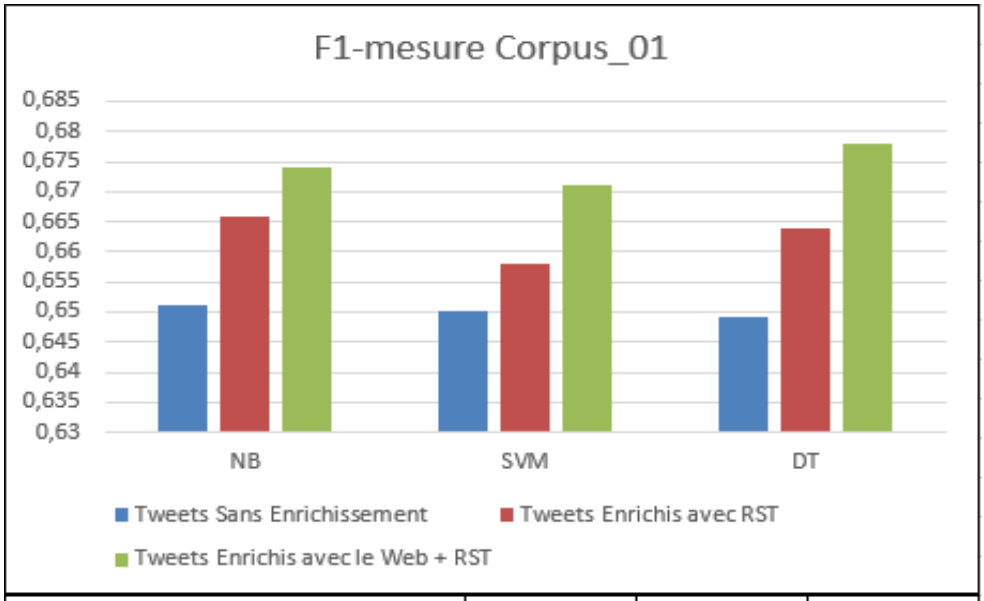
Tableau 18 . Description des deux corpus

	Catégories	**Nombre de tweets**
Corpus_1	Cinéma / سينما	240
	Documentaire / وثائقي	188
	Education / التعليم	147
	Economie / اقتصاد	131
	Santé / الصحة	320
	Actualité / أخبار	120
	Sport / الرياضة	139
	Tourisme / السياحة	196
	Total	**1481**
Corpus_2	Economie / اقتصاد	682
	Electronique / إلكترونيك	160
	Santé / الصحة	580
	Politique / السياسة	154
	Religion / الدين	628
	Science / العلوم	540
	Sociologie / علم الاجتماع	223
	Sport / الرياضة	508
	Total	**3475**

4.3 Résultats & Discussion

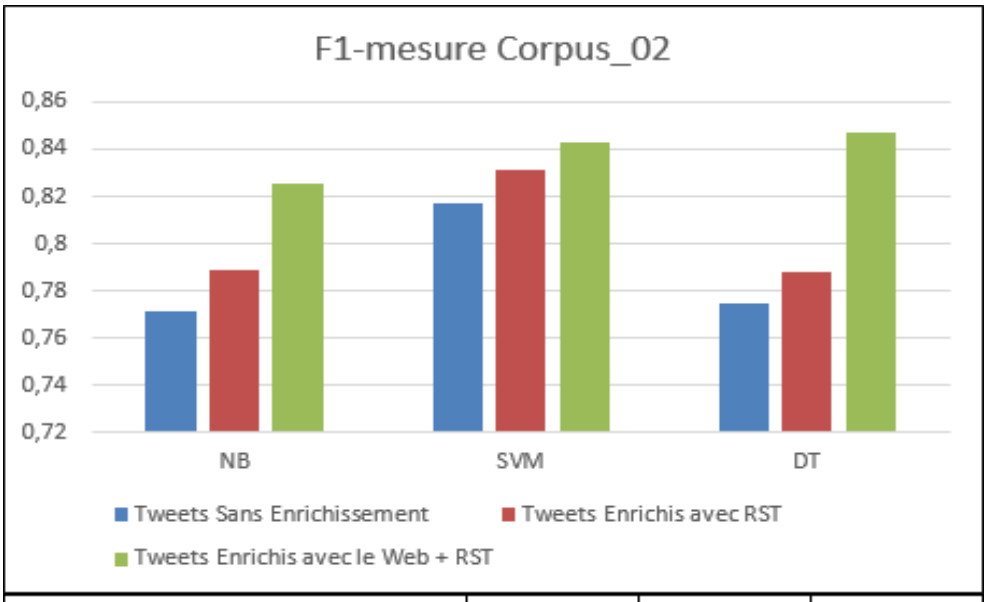
Figure 19 et 20 présente les résultats obtenus de la F1-mesure pour le premier et le second corpus respectivement.

En utilisant le classificateur NB par exemple, la f1-mesure obtenue des tweets sans enrichissement dans le corpus_01 est 0,651 (0,771 dans le corpus_02) ; lorsque nous appliquons l'enrichissement des tweets avec RST seulement, la f1-measure obtenue est 0,666 (0,789 dans le corpus_02) ; par contre, le moment où les tweets seront enrichis avec le Web, la f1-mesure sera 0,674 (0,825 dans le corpus_02). Nous pouvons remarquer une amélioration significative de la performance du système de catégorisation grâce à de l'enrichissement de la représentation des tweets.



F1-mesure	NB	SVM	DT
Tweets Sans Enrichissement	0,651	0,65	0,649
Tweets Enrichis avec RST	0,666	0,658	0,664
Tweets Enrichis avec le Web + RST	0,674	0,671	0,678

Figure 19 . F1-mesure de l'enrichissement basé sur le Web comparé avec l'enrichissement basé sur RST pour le premier corpus



F1-mesure	NB	SVM	DT
Tweets Sans Enrichissement	0,771	0,817	0,775
Tweets Enrichis avec RST	0,789	0,831	0,788
Tweets Enrichis avec le Web + RST	0,825	0,843	0,847

Figure 20 . F1-mesure de l'enrichissement basé sur le Web comparé avec l'enrichissement basé sur RST pour le deuxième corpus

Les résultats obtenus montrent qu'enrichir les tweets avec des termes supplémentaires provenant du Web conduit à une meilleur amélioration de la représentation des tweets plutôt que d'enrichir leur contenu en se basant uniquement sur les termes existant dans le même corpus (c'est-à-dire un enrichissement avec RST). En plus, les résultats obtenus avec les deux autres classificateurs (SVM et DT) montrent l'intérêt de notre proposition.

5 Conclusion

Dans ce chapitre, nous avons proposé une deuxième contribution afin d'enrichir la représentation de texte court en utilisant le Web comme étant un grand corpus ouvert et dynamique combiné avec la théorie des ensembles approximatifs comme étant un outil mathématique pour traiter les données vagues et imprécises. Notre proposition a été testée et évaluée dans un système de catégorisation des tweets en arabe avec trois classificateurs NB, SVM et DT; les résultats obtenues ont monté l'intérêt de notre proposition.

Partie 3 : Enrichissement de Texte Court basé sur les Concepts (Concept-based)

- **Chapitre 5 : Web-RST Mesure de Similarité dédiée au Texte Court**

- **Chapitre 6 : Représentation Conceptuelle de Texte Court basé sur LDA**

Introduction de la partie 3

Avec l'utilisation croissante d'Internet, un grand nombre de textes courts ont été générés sur le Web et les applications mobiles. Contrairement aux documents traditionnels comportant un grand nombre de mots, les textes courts, qui ne contiennent que quelques mots, représentent un défi pour plusieurs applications de fouille de texte. La plupart des techniques existantes représentent les textes comme des vecteurs de termes (sac de mot (*Bag of Word – BoW*)) qui ignorent toutes les informations sémantiques ou conceptuelles contenues dans le texte. Les textes courts rendent ces problèmes encore plus sérieux, en raison de leur brièveté et le manque de l'information contextuelle. Afin de résoudre les problèmes précédents de la représentation basée sur les termes, des ressources sémantiques externes telles que des thésaurus ou des ontologies peuvent être utilisées pour remplacer la représentation BoW par une autre représentation basée sur les concepts (*Bag of Concept – BoC*) ; qui transforme la représentation des données en une représentation plus courte, plus compacte et plus prédictive.

La conceptualisation consiste à remplacer les termes dans un texte par leurs concepts correspondants dans une ressource sémantique [99]. En général, la conceptualisation du texte est réalisée en deux étapes : premièrement, on transforme chaque document en texte brut ; puis on supprime les mots vides qui sont considérés comme des mots non pertinents (prépositions et particules) ; et enfin, un algorithme de racinisation ou de lemmatisation peut être appliqué pour générer la racine ou le lemme afin d'éliminer la redondance. Deuxièmement, une liste de termes candidats est créé pour lesquels nous allons chercher leurs concepts correspondants, en faisant le mapping (dit aussi la projection) de ces termes dans une ressource sémantique, qui seront par la suite intégré dans le texte conceptualisé final. Trois stratégies différentes peuvent être utilisées pour la conceptualisation du texte [99] :

- Ajout de Concepts (***Add Concept***)
  Le texte est enrichi par les concepts correspondants des termes tout en gardant le contenu initial (c'est-à-dire les termes originaux du texte).
- Conceptualisation Partielle (***Partial Conceptualization***)
  Chaque terme est remplacé par son concept correspondant et les termes sans concepts sont préservés dans le texte.

- Conceptualisation Complète (***Complete Conceptualization***)

  Les termes sont remplacés par leurs concepts correspondants, tandis que les termes n'ayant pas de concept sont éliminés.

Les concepts ne sont pas ambigus : chaque concept représente un sens unique. Les synonymes sont mappés (projetés) vers le même concept, tandis que les termes ayant plusieurs significations sont mappés vers différents concepts correspondant à leurs significations. Cependant, lors de la recherche des concepts correspondant aux termes polysémiques, plusieurs correspondances seront détectées et par conséquent, ils introduisent des ambiguïtés dans la représentation du texte conceptualisé final. En effet, l'attribution d'un terme à un concept correspondant est une étape cruciale dans le processus de conceptualisation. Pour surmonter ce problème, trois stratégies de désambiguïsation lexicale (***Word Sense Disambiguation - WSD***) peuvent être utilisées [99] :

- Tous les Concepts (***All Concepts***)

  Tous les concepts correspondants seront ajoutés pour représenter le terme mappé dans le texte conceptualisé final.
- Premier Concept (***First Concept***)

  Le terme mappé sera remplacé par le concept le plus utilisé en se basant sur des statistiques linguistiques.
- Concept de Contexte (***Context Concept***)

  Le terme mappé sera remplacé par le concept le plus similaire parmi les concepts correspondants en se basant sur le contexte du terme dans le document.

Durant notre étude des travaux concernant la conceptualisation de texte court, nous avons pu soulever deux problèmes majeurs dans le processus de conceptualisation. Le problème de WSD et celui du taux de couverture pour le cas de la langue arabe dans les ressources sémantiques existantes.

Cette troisième partie est composée des deux chapitres 5 et 6, elle présente respectivement deux contributions concernant la représentation BoC pour le texte court en arabe. Dans le cinquième chapitre, nous présentons une solution efficace pour surmonter le problème de WSD dans le processus de conceptualisation en se basant sur une nouvelle mesure de similarité dédiée au texte court. Dans le sixième chapitre, nous proposons de résoudre le problème de couverture de la langue arabe dans les ressources sémantiques en intégrant la modélisation par thèmes (*Topic Modeling – TM*) dans le processus de mapping terme / concept.

CHAPITRE 5

Web-RST Mesure de Similarité dédiée au Texte Court

1 Introduction 91
2 Web-RST pour Mesurer la Similarité de Textes Court 91
3 Evaluation Subjective de Web-RST 94
4 Intégration de la Mesure WEB-RST pour Résoudre le Problème d'Ambiguïté Lexicale 96
5 Expérimentations et Résultats 98
5.1 Description des Expérimentations 98
5.2 Corpus Utilisés 99
5.3 Résultats & Discussion 101
6 Conclusion 103

1 Introduction

La désambiguïsation lexicale est le processus qui consiste à déterminer le sens d'un mot utilisé dans un contexte bien déterminé. Dans ce chapitre, nous proposons une méthode très performante pour résoudre le problème de désambiguïsation. Pour cette raison, nous réintroduisons la fonction *Web-Kernel* pour mesurer la similarité sémantique entre une paire de texte court en langue arabe. Cette nouvelle mesure sera intégrée par la suite dans le processus de WSD. La méthode proposée a été testée, évaluée et comparée en utilisant un système de catégorisation de texte court en langue arabe; les résultats obtenus permettent de mettre en évidence la performance de la méthode proposée.

2 Web-RST pour Mesurer la Similarité de Textes Court

Dans cette section, nous présentons en détail notre méthode nommée *Web-RST* pour la mesure de similarité entre deux textes courts en arabe (Figure 21).

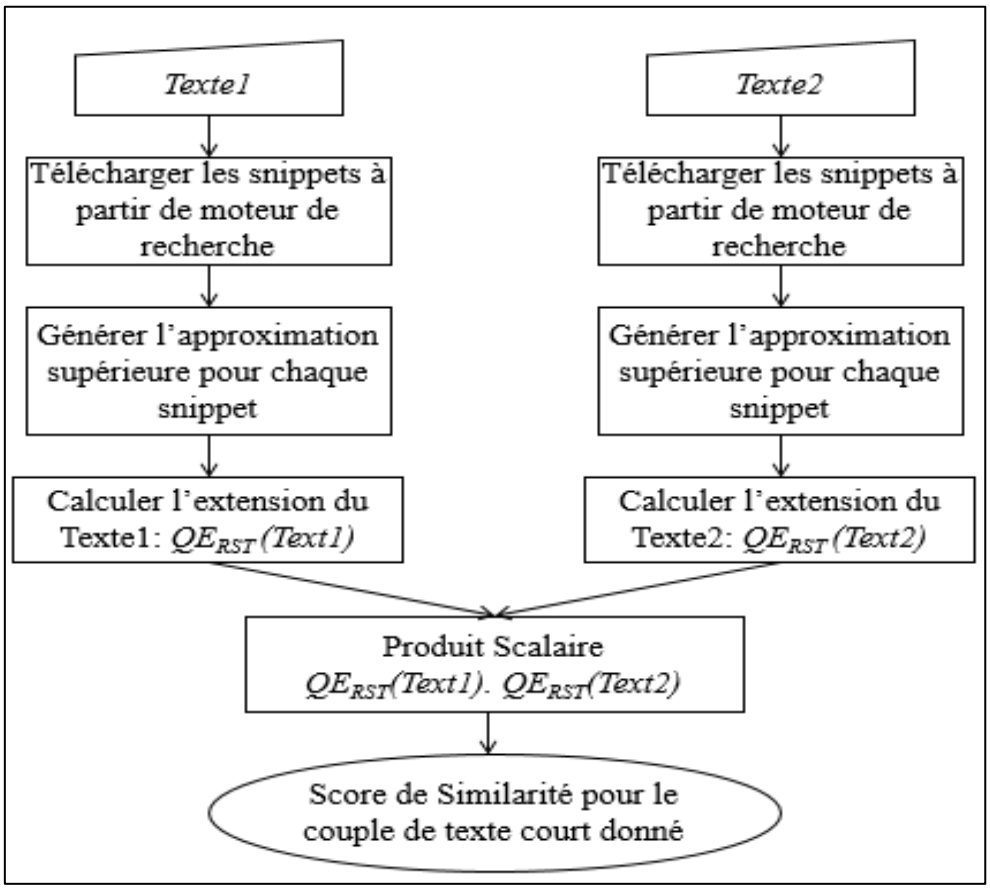


Figure 21 . La méthode Web-RST pour calculer le score de similarité entre deux textes courts

Notre mesure est une extension de la fonction *Web-based Kernel* proposée par Sahami et Heilman [69] et que nous avons déjà présentée dans le premier chapitre d'Etat d'Art. Ils traitent

chaque texte court comme une requête à envoyer vers un moteur de recherche Web, puis en se basant sur les snippets renvoyés par le moteur de recherche, ils créent pour chaque requête un vecteur de mots (nommé *vecteur de contexte*) qui apparient dans le même contexte que les termes originaux en utilisant TF*IDF comme un système de pondération. Enfin, la similarité est définie comme étant le produit scalaire des deux vecteurs de contexte.

Dans notre proposition, nous avons apporté deux changements majeurs par rapport à l'algorithme original, nous pouvons les résumés de la façon suivante :

1. Étant donné que les snippets (documents) renvoyés du moteur de recherche Web sont considérés comme du texte court, nous proposons d'enrichir ces snippets et de renforcer les liens sémantiques entre eux en utilisant la théorie des ensembles approximatifs (RST). Dans ce cas, chaque document sera remplacé par son approximation supérieure à l'aide de la formule (15).
2. Chaque document sera représenté par un vecteur de poids où la formule originale de TF-IDF est remplacée par la formule (16). Cette formule permet de faire la différence entre les termes qui existent seulement dans l'approximation supérieure et non pas dans le texte lui-même.

Les étapes détaillées de la nouvelle méthode d'extension de requête sont décrites ci-dessous :

- Soit D_n (x) l'ensemble des *n* premiers snippets renvoyés par un moteur de recherche lors de l'utilisation de x comme une requête (nous utilisons les services proposés par Google[33] et Bing[34]). Notons que nous avons établi un mécanisme pour éliminer la redondance en fonction de l'URL de chaque snippet retourné par les deux moteurs de recherche utilisé.
- Chaque snippet $d_i \in D_n$ (x) sera nettoyé en supprimant les caractères latins, les caractères spéciaux comme (/, #, $, etc…) et les mots vides de la langue arabe. Ensuite, Nous générons l'approximation supérieure U (d_i).
- Construire le vecteur des poids v_i pour chaque approximation U (d_i), où chaque élément de v_i est le poids $w_{i,j}$ du terme t_j dans l'approximation supérieure de d_i en utilisant le nouveau système de pondération.
- Calculer C(x) le centre des vecteurs v_i selon la norme L_2 défini de la façon suivante :
  $C(x) = 1/n * \sum (vi / || vi ||)$

[33] https://developers.google.com/custom-search/docs/start
[34] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44

- Calculer QE_{RST} (x) l'extension de la requête x et qui est défini comme la normalisation du centre C(x) selon la norme L_2 en utilisant la formule : QE_{RST} (x) = C (x) / || C (x) ||

Étant donné une paire de texte courts x et y, la mesure de similarité Web-RST est définie comme le produit scalaire des deux extensions de x et y respectivement :
QERST (x) .QERST (y).

L'algorithme de calcul de la similarité Web-RST peut être résumé et présenté comme ci-dessous (algorithme 2).

Algorithme 2 : WebRST pour calculer la similarité entre deux textes court

```
Entrée : Un couple de texte court en langue arabe (textI, text2)
Sortie : Le score de similarité entre textI et text2
Début
    slI ← N top résultats retournés par le moteur de recherche en utilisant textI comme une requête
    sl2 ← N top résultats retournés par le moteur de recherche en utilisant text2 comme une requête
    genererApproximationSuperieur (slI)      //Calculer l'approximation supérieure des document de slI
    genererApproximationSuperieur (sl2)      //Calculer l'approximation supérieure des document de sl2
    calculer le vecteur des poids vi= pour chaque approximation U(di) dans slI
    calculer le vecteur des poids wi pour chaque approximation U(di) dans sl2
    CI ← (I / N) * ∑ vi / ||vi||2            // CI est le centre des vecteurs vi
    QEI ← CI / ||CI||2                       // QEI est l'extension de la requête textI
    C2 ← (I / N) * ∑ wi / ||wi||2            // C2 est le centre des vecteurs wi
    QE2 ← C2 / ||C2||2                       // QE2 est l'extension de la requête text2
    sim ← QEI . QE2                          // . est le produit scalaire
    return sim
end

void genererApproximationSuperieur (u)
Début fonction
    Prétaitement du corpus u
    Initialiser l'ensemble des termes s
    Initialiser la matrice des co-occurrences
    Pour chaque terme ti dans s do
        Pour chaque terme tj dans s do
            Si le nombre de co-occurrence de ti et tj > θ alors
                ajouter tj à la classe de tolérance de ti
            fin si
        Fin Pour
    Fin Pour
    Pour chaque docuemnt d dans u do
        Pour chaque terme tj dans s do
            tct = la classe de tolerance de tj
            cof = |tct ∩ d| / |tct|
            Si cof > 0 alors
                Ajouter tj à l'approximation supérieur de d
            fin si
        Fin Pour
    Fin Pour
Fin fonction
```

3 Evaluation Subjective de Web-RST

Nous avons examiné plusieurs paires de textes courts en arabe pour déterminer le score de similarité donné par l'algorithme originale présenté par Sahami et Heilman [69] et le score donné par notre algorithme amélioré. Nous avons donné aussi le score de similarité obtenu aussi par la mesure cosinus en calculant l'angle entre les deux textes et qui est défini par la formule (6).

Nous avons examiné en particulier trois types de correspondance de texte court : les acronymes, les individus et leurs positions, et les termes à multiples facettes (termes pouvant avoir plusieurs sens dans divers contextes [49]). Le tableau 19 présente quelques exemples des paires de texte court pour lesquelles nous avons compté le score de similarité.

La première partie du tableau traite l'identification des acronymes. Nous remarquons que notre méthode est efficace et performante. En effet, elle permet de capturer la similarité sémantique entre un acronyme et son nom complet et donne de meilleurs résultats par rapport à l'algorithme d'origine. Cela est dû à l'utilisation de l'approximation supérieure des snippets renvoyés par le moteur de recherche. D'autre part, ce n'est pas surprenant que l'utilisation de la similarité cosinus soit totalement inappropriée pour une telle tâche (puisque le nom complet d'un acronyme ne contient pratiquement jamais l'acronyme lui-même).

De plus, nous avons testé l'utilisation de notre mesure de similarité pour identifier les noms d'un certain nombre d'individus avec leurs positions qu'ils occupent. Les résultats de ces exemples sont présentés dans la deuxième section du tableau 19. Nous avons calculé le score de similarité aux paires correctes de position et individu, ainsi qu'aux paires impliquant un individu et une position proche mais incorrecte. Par exemple, « 'Ban Ki-moon' / 'بان كي مون' » et « 'Yasser Arafat' / 'عرفات ياسر' » sont tous les deux des personnalités politiques, notre algorithme est efficace pour déterminer les bons rôles et leur attribuer des scores élevés.

Enfin, nous avons également examiné l'efficacité de notre mesure lorsqu'elle est appliquée à des textes contenant des termes de multiples facettes. Notre algorithme permet de définir les différentes facettes des termes, par exemple en identifiant « 'Invasion spatiale' / 'الفضاء غزو' » avec « 'Espace collectif pour l'éducation' / 'الفضاء الجمعوي للتربية' ».

Tableau 19 . Exemples des mesures de similarité obtenus pour le texte court

Texte 1	Texte 2	Web-RST	Web-based	Cosine
Acronymes				
فقدان المناعة المكتسبة Immunodéficience acquise	الأيدز SIDA	0.719	0.701	0
فقدان المناعة المكتسبة Immunodéficience acquise	السيدا SIDA	0.7	0.678	0
رابطة الاتحاد الدولي لكرة القدم Fédération Internationale de Football Association	فيفا FIFA	0.635	0.599	0
الحمض النووي Acide nucléique	الدنا DNA	0.845	0.812	0
منظمة الأغذية والزراعة Organisation pour l'alimentation et l'agriculture	الفاو FAO	0.719	0.692	0
منظمة الأمم المتحدة للتربية والعلوم و الثقافة Organisation des Nations Unies pour l'éducation, la science et la culture	اليونسكو UNESCO	0.745	0.727	0
المركز الوطني للملاحة الجوية وإدارة الفضاء Le Centre National d'Aéronautique et d'Administration Spatiale	ناسا NASA	0.646	0.594	0
Individus and leur positions				
أمير الأطباء Prince des Médecins	ابن سينا Ibn Sina	0.725	0.709	0
رئيس السلطة الوطنية الفلسطينية Président de l'Autorité Nationale Palestinienne	ياسر عرفات Yasser Arafat	0.761	0.712	0
النظرية النسبية Théorie de la relativité	ألبرت اينشتاين Albert Einstein	0.918	0.912	0
الأمين العام للأمم المتحدة Secrétaire général des Nations Unies	بان كي مون Ban Ki-moon	0.748	0.668	0
رئيس السلطة الوطنية الفلسطينية Président de l'Autorité Nationale Palestinienne	بان كي مون Ban Ki-moon	0.442	0.428	0
Termes à multiple facettes				
صلاة العشاء Ishaa Prière	حفل العشاء Diner de Gala	0.523	0.518	0.5
صلاة المغرب Maghreb Prière	المغرب العربي Grand Maghreb	0.264	0.266	0.5
الفضاء الجمعوي للتربية Espace collectif pour l'éducation	غزو الفضاء Invasion de l'espace	0.129	0.127	0.408

Une mesure de similarité peut être testée et évaluée avec la précision globale à l'aide d'un corpus contenant des paires de texte court. Ces corpus sont généralement annotés à la main avec une information de similarité binaire déterminant si une paire de texte court est similaire ou non. En raison de l'absence d'un corpus dédié au test et à l'évaluation de la similarité entre les textes courts en arabe, nous avons opté pour une autre méthode pour confirmer notre contribution et nous croyons que cette mesure peut s'avérer très utile dans la conceptualisation de texte court, et précisément pour améliorer le processus de désambiguïsation lexicale.

4 Intégration de la Mesure WEB-RST pour Résoudre le Problème d'Ambiguïté Lexicale

Dans cette section, nous présentons en détail notre méthode proposée pour la conceptualisation de texte court en arabe. Pour obtenir une représentation basée sur les concepts (Bag of Concept (BoC)), il faut tout d'abord passer par la représentation basée sur les termes (Bag of Word (BoW)) afin de trouver les termes candidats à projeter dans la ressource sémantique utilisée.

Chaque texte sera transformé en texte brut, nettoyé en supprimant les mots latins, les caractères spéciaux tels que (/, #, $, etc…) et les mots vides arabes considérés comme non pertinents (prépositions et particules). L'étape suivante consiste à extraire la racine (root) / radical (stem) du mot afin d'effectuer une certaine forme de normalisation morphologique pour créer des caractéristiques plus générales. Différents algorithmes ont été proposés dans ce sujet, tels que Khoja [79] et Larkey Stemmer [100], qui sont les algorithmes les plus connus pour le texte arabe.

Pour surmonter les faiblesses de la représentation BoW, une représentation BoC en utilisant BabelNet [107] comme étant une ressource sémantique largement utilisée. La figure 22 présente l'organigramme illustrant les étapes de notre méthode proposée pour la conceptualisation de texte court en arabe. Au cours de la conceptualisation, différentes stratégies peuvent être mises en œuvre comme il est décrit précédemment (ajout de concepts, conceptualisation partielle et conceptualisation complète). La mise en œuvre de la représentation BoC repose sur deux étapes principales :

- Mapping des Termes (la projection)

  Cette étape consiste à projeter les termes dans BabelNet afin de sélectionner les concepts candidats ainsi que le choix de stratégie de conceptualisation utilisée.

- La Désambiguïsation Lexicale (WSD)

  Rappelons que l'utilisation d'une mesure de similarité de texte court adéquate peut améliorer le processus d'identification du sens utilisé d'un mot. Nous avons été inspirés par la mesure de similarité basée sur le Web introduite par Sahami et Heilman [69], nous avons apporté des modifications cruciales pour l'exploiter dans le processus de WSD comme expliqué par la suite.

Le but de la deuxième étape est d'identifier le concept correspondant à un terme dans BabelNet. Pour ce faire, nous avons appliqué notre mesure de similarité *Web-RST* pour une paire de texte court où les deux textes seront formés de la manière suivante : le premier sera composé par le terme à chercher dans BabelNet avec ses termes voisins pour considérer son contexte ; et le second sera la définition d'un concept candidat. Nous répétons ce traitement pour tous les concepts candidats et le concept à choisir est celui avec la similarité la plus grande. L'algorithme de désambiguïsation peut être résumé et présenté comme ci-dessous (algorithme 3).

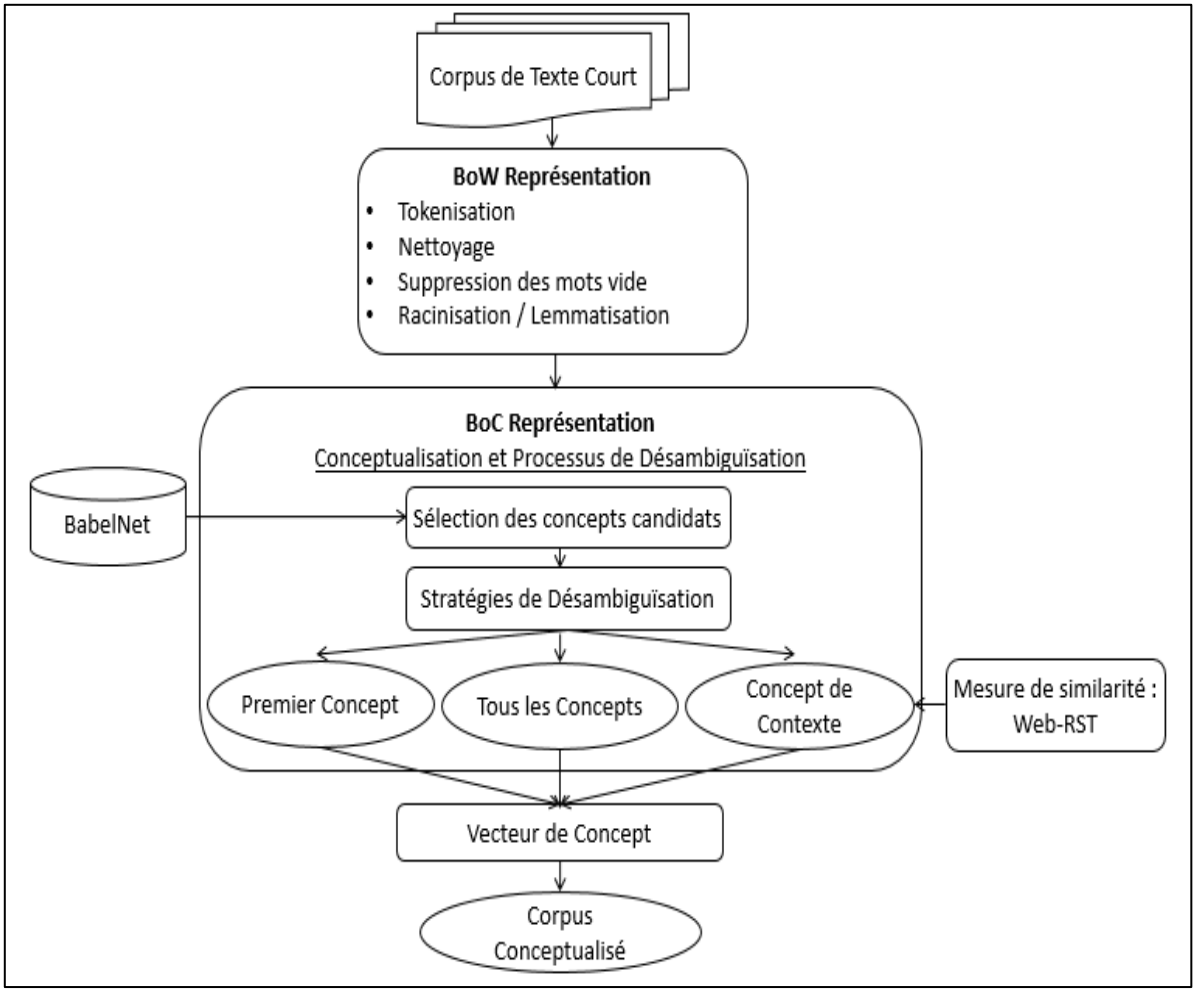


Figure 22 . Méthode proposé pour la conceptualisation de texte court en arabe basée sur la mesure Web-RST

Algorithme 3 : Processus de WSD

```
Entrée : un terme ambiguë t
Sortie : le concept correspondant à t
Début
res ← sera le concept à retourner pour le terme t
cl ← sera la liste des concepts candidats retournés à partir de BabelNet pour le terme t
qt ← le terme t avec ses termes voisins
Pour chaque concept c dans cl faire
   qc ← la définition du concept c
   sim ← WebRSTSim(qt, qc)
   Si sim > maxSim alors
      maxSim ←  sim
      res ← c
   Fin Si
Fin Pour
retourner res
Fin
```

Afin de tester l'efficacité de notre méthode proposée pour la conceptualisation de texte en arabe, nous avons construit un système de catégorisation. Nous avons utilisé deux des algorithmes d'apprentissage automatique les plus populaires pour la catégorisation de textes : NB [87] et SVM [88] [89].

5 Expérimentations et Résultats

Dans cette section, nous présentons une description détaillée des expérimentations que nous avons menées afin de valider notre contribution, ainsi que les corpus utilisés et les résultats obtenus.

5.1 Description des Expérimentations

Pour illustrer que notre méthode améliore la représentation des textes courts en arabe et par conséquent les performances de notre système de catégorisation, une série d'expérimentations a été menée (Figure 23). Au cours du processus de conceptualisation, nous avons utilisé les trois stratégies différentes décrites précédemment (ajout de concepts, conceptualisation partielle et conceptualisation complète). De plus, tout en recherchant des concepts correspondants à des termes polysémiques, nous avons adopté les trois stratégies de désambiguïsation décrites précédemment (premier concept, tous les concepts, et contexte concept). Notez que, la troisième stratégie de contexte est basée sur notre mesure de similarité *Web-RST*.

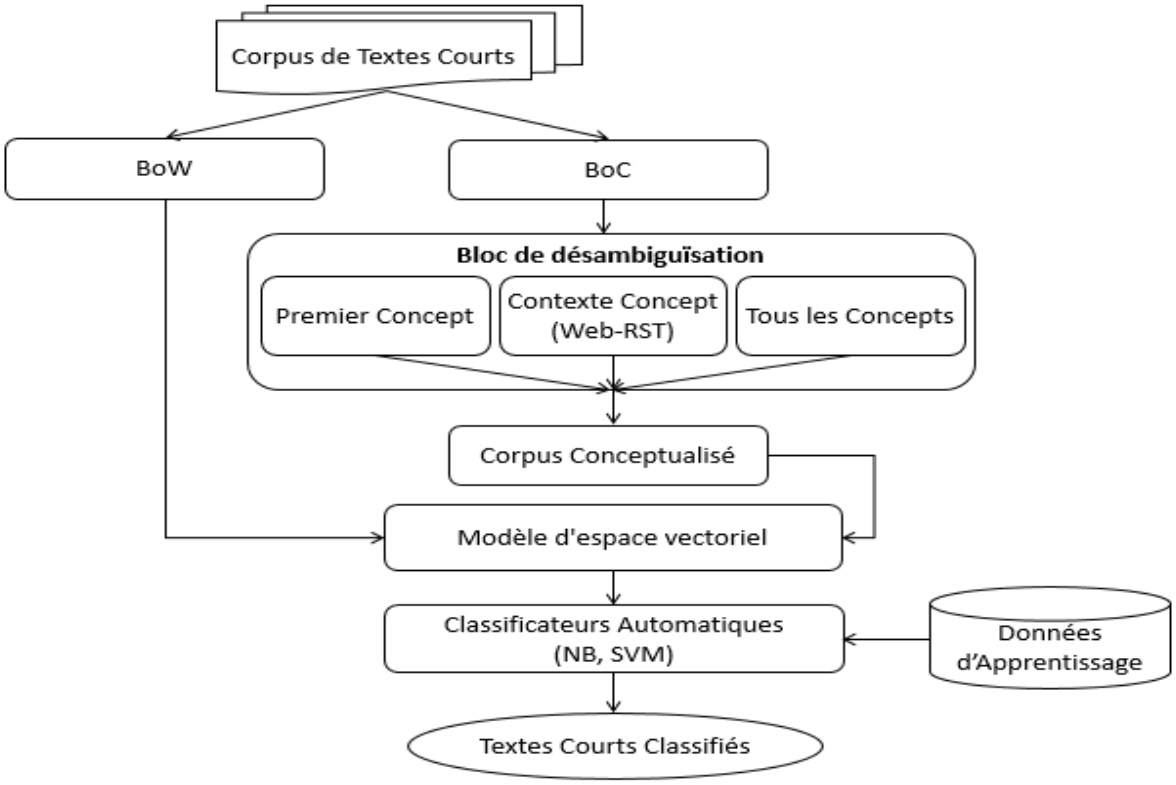


Figure 23 . Description des expérimentations de la conceptualisation de texte court en arabe

La section suivante présente les expérimentations qui ont été menées à l'aide de trois corpus différents et variés afin de valider notre contribution.

5.2 Corpus Utilisés

Dans nos expérimentations, nous avons utilisé trois différents corpus (tableau 20) afin de bien valider notre proposition.

- Le premier corpus utilisé dans nos expérimentations est un sous-ensemble du répertoire *« open Directory Project (ODP)*[35] *»* qui est un annuaire multi-langue, composé de quelques millions de pages Web pré-classées et organisées en arborescence. Pour la langue arabe, l'ODP inclut 4781 textes courts sous forme des snippets pré-classés en 459 catégories par un groupe d'experts. Le sous-ensemble que nous avons utilisé est composé de sept catégories : commerce / 'التجارة', famille / 'الاسرة', santé / 'الصحة', économie / 'اقتصاد', actualité / 'أخبار', sport / 'الرياضة' et étude / 'الدراسة'.
- Le second corpus est un ensemble de tweets en arabe collecté à partir de Twitter toujours en utilisant l'outil *NodeXL Excel Template*[36]. Le corpus est classé manuellement en huit

[35] www.dmoz.org
[36] http://social-dynamics.org/twitter-network-data/

catégories : économie / 'اقتصاد', électronique / 'إلكترونيات', santé / 'الصحة', politique / 'السياسة', religion / 'الدين', science / 'العلوم', sociologie / 'علم الاجتماع' et sport / 'الرياضة'.

- Le dernier corpus utilisé dans est un sous-ensemble du corpus *« the Large-Scale Arabic Book Reviews Dataset (LARB dataset) »* [101]. Ce corpus contient plus de 63 000 opinions concernant des livres en arabe. Les opinions ont été tirées du site Web *«www.goodreads.com»* au cours du mois de mars 2013. Le sous-ensemble que nous avons utilisé est composé de trois catégories : positive / 'ايجابي', négative / 'سلبي' et neutre / 'محايد'.

Tableau 20 . Description des trois corpus

Corpus	Catégories	Nombre de textes
ODP_Corpus	Commerce / التجارة	242
	Famille / الاسرة	69
	Santé / الصحة	145
	Economie / اقتصاد	251
	Actualité / أخبار	252
	Sport / الرياضة	105
	Etude / الدراسة	153
	Total	**1217**
Tweet_Corpus	Economie / اقتصاد	682
	Electronique / إلكترونيات	160
	Santé / الصحة	580
	Politique / السياسة	154
	Religion / الدين	628
	Science / العلوم	540
	Sociologie / علم الاجتماع	223
	Sport / الرياضة	508
	Total	**3475**
LARB_ Corpus	Positive / ايجابي	725
	Négative / سلبي	628
	Neutre / محايد	495
	Total	**1848**

5.3 Résultats & Discussion

L'efficacité de notre système a été évaluée et comparée en termes de la mesure f1-mesure à l'aide des classificateurs NB et SVM en utilisant une technique de validation croisée sur *10 Fold*. Le tableau 21 montre les résultats obtenus des trois corpus utilisés. Nous avons testé la catégorisation avec une représentation BoW et en utilisant chacune des stratégies de conceptualisation présentées précédemment.

La représentation de BoC utilisant la stratégie de conceptualisation partielle et le concept de contexte basé sur notre mesure de similarité *Web-RST* comme méthode de WSD obtient la f1-mesure la plus élevée par rapport à la représentation de base (BoW). L'amélioration est significative car BoC supprime les problèmes introduits par la synonymie et la polysémie et augmente donc les performances du classificateur. Grâce à cette stratégie, nous pouvons enrichir la représentation en texte court sans perdre aucune information. La f1-mesure la plus basse a été obtenue par la stratégie de conceptualisation complète combinée à tous les concepts comme une méthode de résolution de WSD. Ceci est dû à l'augmentation de l'espace de représentation.

Comme mentionné précédemment, la mesure de similarité pour le texte court s'est avérée être très utile pour le processus de désambiguïsation lexicale. Nous pouvons remarquer que les meilleurs résultats pour la f1-mesure sont obtenus en utilisant notre méthode proposée pour calculer la mesure de similarité entre une paire de textes courts avec toutes les stratégies de conceptualisation (ajout de concepts, conceptualisation partielle et conceptualisation complète). En plus, les résultats obtenus avec l'autre classificateur SVM montre l'intérêt de notre proposition.

Tableau 21 . F1-mesure pour la conceptualisation de texte court

			NB	SVM
			F1-measure (Baseline Gap)	**F1-measure (Baseline Gap)**
ODP_Corpus	BoW (Baseline)		0,781	0,612
	Tous les Concepts	Ajout de Concepts	0,703 (-7,80%)	0,561 (-5,10%)
		Conceptualisation Partielle	0,712 (-6,90%)	0,567 (-4,50%)
		Conceptualisation Complète	0,693 (-8,80%)	0,548 (-6,40%)
	Premier Concept	Ajout de Concepts	0.783 (+0,20%)	0,613 (-0,10%)
		Conceptualisation Partielle	0.792 (+1,10%)	0,619 (+0,70%)
		Conceptualisation Complète	0.768 (-1,30%)	0,563 (-4,90%)
	Concept de Contexte	Ajout de Concepts	0,834 (+5,30%)	0,645 (+3,30%)
		Conceptualisation Partielle	0,857 (+7,60%)	0,679 (+6,70%)
		Conceptualisation Complète	0,806 (+2,50%)	0,638 (+2,60%)
Tweet_Corpus	BoW (Baseline)		0,812	0,724
	Tous les Concepts	Ajout de Concepts	0,778 (-3,40%)	0,695 (-2,90%)
		Conceptualisation Partielle	0,79 (-2,20%)	0,712 (-1,20%)
		Conceptualisation Complète	0,754 (-5,80%)	0,687 (-3,70%)
	Premier Concept	Ajout de Concepts	0,809 (-0,30%)	0,718 (-0,60%)
		Conceptualisation Partielle	0,814 (+0,20%)	0,725 (+0,10%)
		Conceptualisation Complète	0,801 (-1,10%)	0,702 (-2,20%)
	Concept de Contexte	Ajout de Concepts	0,859 (+4,70%)	0,766 (+4,20%)
		Conceptualisation Partielle	0,891 (+7,90%)	0,801 (+7,70%)
		Conceptualisation Complète	0,847 (+3,50%)	0,749 (+2,50%)
LARB_Corpus	BoW (Baseline)		0,714	0,652
	Tous les Concepts	Ajout de Concepts	0,681 (-3,30%)	0.623 (-2.90%)
		Conceptualisation Partielle	0,697 (-1,70%)	0.637 (-1.50%)
		Conceptualisation Complète	0,668 (-4,60%)	0.610 (-4.20%)
	Premier Concept	Ajout de Concepts	0,705 (-0,90%)	0,640 (-1,20%)
		Conceptualisation Partielle	0,718 (+0.40%)	0,654 (+0,20%)
		Conceptualisation Complète	0,697 (-1,70%)	0,631 (-2.10%)
	Concept de Contexte	Ajout de Concepts	0,749 (+3,50%)	0.684 (+3.20%)
		Conceptualisation Partielle	0,762 (+4,80%)	0.701 (+4.90%)
		Conceptualisation Complète	0,736 (+2,20%)	0.676 (+2.40%)

6 Conclusion

Dans ce chapitre, nous avons proposé une mesure de similarité dédiée au texte court, cette dernière permet de résoudre le problème de désambiguïsation lexicale. En effet, nous avons réintroduit la fonction *Web-Kernel* pour mesurer la similarité sémantique entre une paire de texte court en langue arabe. Cette nouvelle mesure a été intégrée par la suite dans le processus de WSD. La méthode proposée a été testée, évaluée et comparée en utilisant un système de catégorisation de texte court en arabe avec deux classificateurs NB et SVM ; les résultats obtenus ont montré l'intérêt de notre proposition.

CHAPITRE 6

Représentation Conceptuelle de Texte Court basé sur LDA

1 Introduction 105
2 Problème de Couverture de la Langue Arabe dans les Ressources Sémantiques 105
3 L'Allocation Dirichlet Latente 106
3.1 Définition 106
3.2 Le Modèle 107
3.3 Processus de Génération 108
3.4 Loi de Dirichlet 108
3.5 Probabilité Jointe et Loi Marginale 109
3.6 Inférence et Estimation de Paramètres 110
3.7 Inférence 110
4 Intégration de LDA dans la Conceptualisation de Texte Court 112
5 Expérimentations et Résultats 114
5.1 Description des Expérimentations 114
5.2 Corpus utilisé 116
5.3 Résultats & Discussion 116
6 Conclusion 118

1 Introduction

Dans ce chapitre, nous présenterons notre deuxième contribution pour enrichir la représentation du texte court en arabe en utilisant une représentation basée sur les concepts plutôt que sur les termes, ce qui transforme la représentation des données en une structure plus courte, plus compacte et plus prédictive. Cependant, pour la langue arabe, la majorité des ressources sémantiques sont des projets incomplets ; cela pose un sérieux problème de taux couverture de la langue arabe par rapport aux autres langues. Partant de l'hypothèse que les termes appartenant au même topic (sujet) partagent de nombreux liens sémantiques dans un corpus, leurs concepts correspondants partageront presque les mêmes liens sémantiques dans le même corpus. De ce fait, nous proposons d'intégrer la modélisation par thèmes (Topic Modeling) comme un moyen permettant de rassembler les termes ayant les mêmes liens sémantiques. Pour montrer l'efficacité de notre méthode, nous avons choisi l'une des tâches les plus importantes récemment : l'analyse des sentiments ou bien fouille d'opinions. Les résultats obtenus montrent l'intérêt de notre proposition.

2 Problème de Couverture de la Langue Arabe dans les Ressources Sémantiques

Dans le processus de conceptualisation, la ressource sémantique utilisée joue un rôle très important dans le résultat de conceptualisation, elle peut influencer positivement ou négativement sur la performance de toute application de fouille de texte. Ce n'est peut-être pas un problème pour l'anglais ou certaines autres langues européennes, mais pour la langue arabe, c'est un problème très sérieux qui nécessite un peu de réflexion pour surmonter son influence sur les résultats de la conceptualisation.

En effet, pour la langue arabe, la plupart des ressources sémantiques existantes sont des projets incomplets. Par exemple, Arabic WordNet couvre environ 10 198 concepts alors que dans English WordNet, nous en avons trouvé 117 659. Le tableau 22 présente des statistiques sur la couverture de la langue arabe par rapport à l'anglais dans WordNet [103] et BabelNet [104].

Avec ce taux de couverture de la langue arabe dans WordNet, BabelNet ou bien d'autres ressources sémantiques, il est évident que ce problème de couverture influence négativement sur la qualité des résultats de la conceptualisation de texte court en langue arabe. Pour faire face à ce problème dans le cadre de ce travail de thèse, nous proposons une deuxième contribution permettant d'améliorer la représentation conceptuelle de texte court en arabe.

Tableau 22 . Comparaison de la couverture de la langue Arabe et Anglaise dans WordNet et BabelNet

	WordNet		BabelNet	
	Synsets	Word-Senses	Synsets	Word-Senses
Anglais	117, 659	206, 941	8, 855, 224	22, 922, 522
Arabe	10, 198	37, 463	2, 942, 886	4, 267, 782

Rappelons que les termes appartenant au même topic (sujet) partagent de nombreux liens sémantiques dans un corpus, leurs concepts correspondants partageront presque les mêmes liens sémantiques dans le même corpus. De ce fait, un outil permettant d'identifier les termes avec les mêmes liens sémantiques peut être d'une grande utilité pour surmonter les limitations précédentes concernant la couverture de la langue arabe dans les ressources sémantiques. La modélisation par thèmes *« Topic Modeling (TM) »* pourrait être un bon moyen pour surmonter ce problème. En effet, TM a été utilisé avec succès pour diverses tâches du traitement des langues naturelles, c'est un modèle probabiliste permettant de déterminer des sujets ou topics abstraits dans une collection de documents. La plus populaire est l'allocation de Dirichlet latente (LDA) [64], dans laquelle chaque topic est représenté comme une distribution multinomiale sur des mots ; et chaque document est représenté comme une distribution multinomiale sur des topics [105]. Les concepts de base du modèle seront présentés dans la section suivante.

3 L'Allocation Dirichlet Latente

3.1 Définition

L'Allocation Dirichlet Latente (En anglais : Latent Dirichlet Allocation (LDA)) [64] est un modèle probabiliste génératif permettant de décrire des collections de documents de texte ou d'autres types de données discrètes. LDA fait partie d'une catégorie de modèles appelés "topic model", qui cherche à découvrir des structures thématiques cachées dans des documents. Ceci permet d'obtenir des méthodes efficaces pour le traitement et l'organisation des documents de ces archives : organisation automatique des documents par sujet, recherche, compréhension et analyse du texte, ou même résumé des textes.

Aujourd'hui, ce genre de méthodes s'utilisent fréquemment dans le web, par exemple pour analyser des ensemble d'articles d'actualité, les regrouper par sujet, faire de la recommandation d'articles, etc. Des modèles de ce type peuvent également s'utiliser pour l'analyse des sentiments afin de regrouper les documents ou les tweets par sujet.

LDA est un modèle Bayésien hiérarchique à 3 couches (voir la Figure 24 pour une représentation graphique) : chaque document est modélisé par un mélange de topics (thèmes) qui génère ensuite chaque mot du document. La structure des documents du corpus peut être déterminée par des techniques d'inférence approchée basées sur des méthodes variationnelles ou des méthodes de Gibbs sampling [110]. Les paramètres des distributions peuvent être estimés par l'algorithme Expectation Maximisation (EM) [111].

3.2 Le Modèle

La figure 24 représente le modèle graphique de LDA et la figure 25 présente un schéma décrivant LDA. Commençons par expliciter les différents termes et paramètres du modèle :

- Un mot w est la donnée discrète, correspondant à l'indice d'un mot dans un vocabulaire fixe de taille V. On peut considérer que w est un vecteur de taille V de composantes toutes nulles sauf pour la composante i où i est l'indice du mot choisi ($w_i = 1$).
- Un document est un N-uplet de mots, $w = (w_1, \ldots, w_N)$.
- Un corpus est une collection de D documents, $D = (d_1, \ldots, d_D)$.
- Les variables $z_{d,n}$ représentent le topic choisi pour le mot $w_{d,n}$.
- Les paramètres θ_d représentent la distribution de topics du document d.
- α et η définissent les distributions à priori sur θ et β respectivement, où β_k décrit la distribution du topic k.

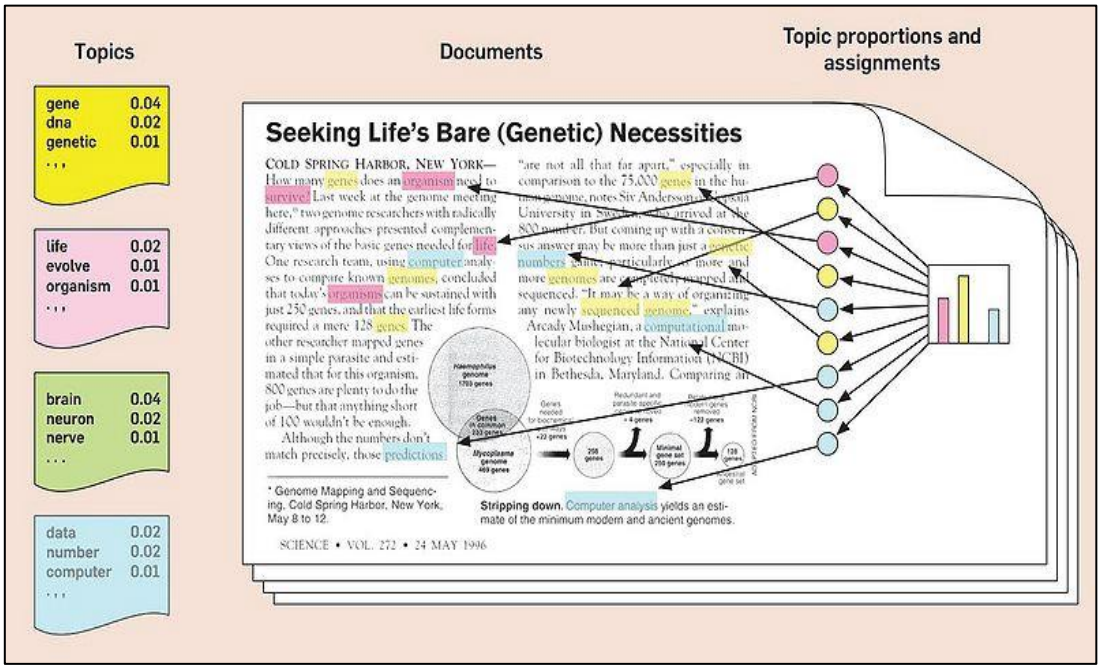


Figure 24 . Schéma Décrivant LDA

A gauche, on peut voir la structure de chaque topic, donnant une probabilité à chaque mot d'un vocabulaire fixe. Pour un document donné, l'histogramme à droite décrit la distribution de topics dans ce document. Pour chaque mot du document, on choisit d'abord un sujet depuis cette distribution (les bulles), puis on tire un mot depuis le sujet choisi [102].

3.3 Processus de Génération

Le processus génératif suivi par LDA pour un document w est le suivant (voir le modèle graphique de la Figure 25) :

1. Choisir $\theta \sim$ Dirichlet(α).
2. Pour chaque mot w_n :
   a. Choisir un topic $z_n \sim$ Multinomial(θ)
   b. Choisir un mot $w_n \sim$ Multinomial (β_k), avec $k = z_n$

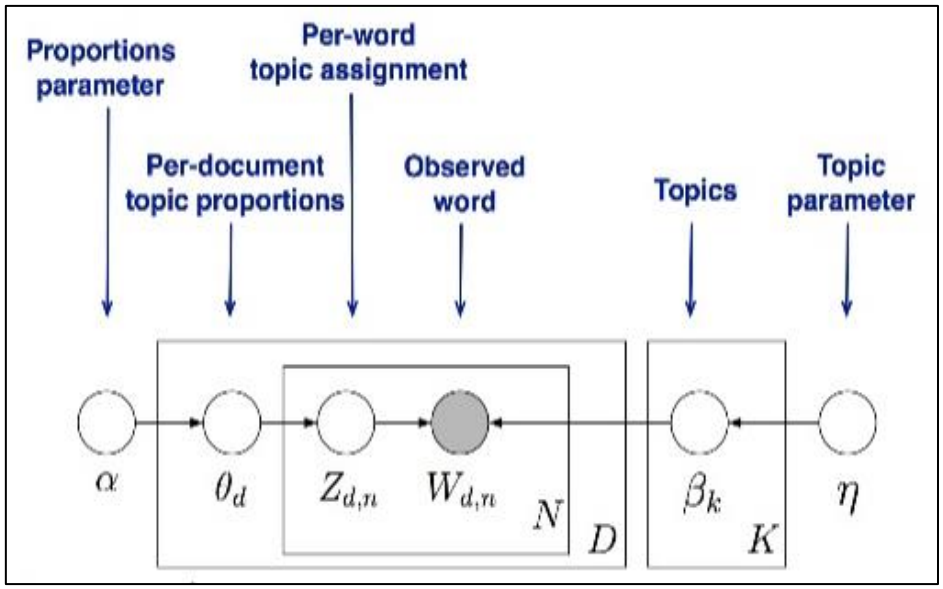


Figure 25 . Représentation de LDA sous forme de modèle graphique

Les boîtes représentent des répliques du modèle qu'elles contiennent (par exemple, il y a une boîte N pour chaque document D) [102].

3.4 Loi de Dirichlet

La loi de Dirichlet permet de tirer une variable θ telle que $\forall i, \theta_i \geq 0$ et $\Sigma_{i=1,k}\, \theta_i = 1$ (θ est dans le $(k-1)$-simplexe). Sa densité est de la forme :

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \qquad (21)$$

Avec $\alpha \in R^k$, $\alpha_i > 0$ et $\Gamma(x)$ la fonction Gamma.

Cette distribution permet donc d'obtenir une distribution multinomiale de paramètre θ, correspondant pour LDA au mélange de topics d'un document w. La Figure 26 montre la densité d'une telle distribution pour 3 topics. Chaque sommet du triangle correspond à un topic, et chaque point du triangle représente donc une pondération des 3 topics. Le paramètre $\alpha = \alpha_1 + \cdots + \alpha_k$ contrôle l'homogénéité des θ_i : lorsque α est grand, les θ_i sont proches et homogènes, lorsque α est petit, la plupart des θ_i sont proches de 0 sauf quelques-uns. Dans les cas extrêmes, tous les θ_i sont égaux ($\alpha \to \infty$) ou tous les θ_i sont nuls sauf un ($\alpha \to 0$).

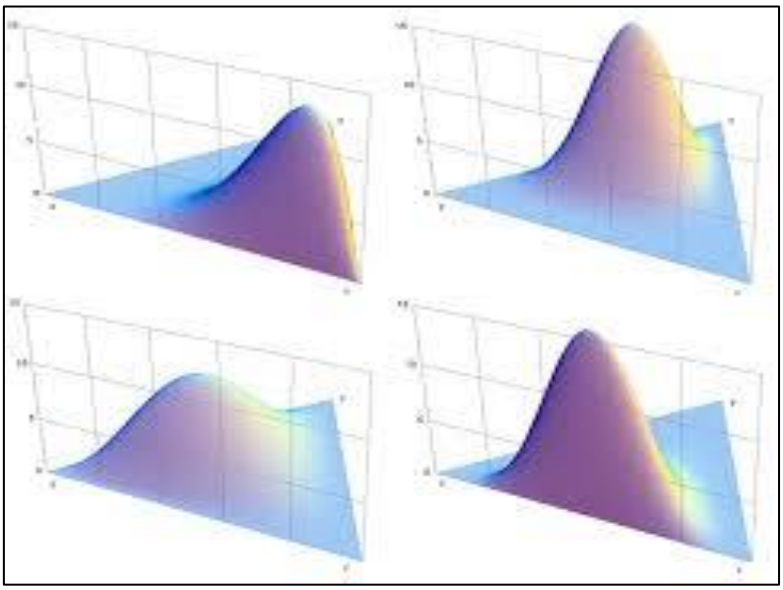


Figure 26 . Fonction de densité de loi de Dirichlet à 2 dimensions (k = 3 topics) sur le triangle

L'un des avantages de la loi de Dirichlet est qu'elle est conjuguée à la loi multinomiale, c'est à dire que si $z_1, \ldots z_N$ sont des variables multinomiales de paramètre θ, alors la variable $\theta|z_1, \ldots z_N$ donnée par :

$$p(\theta|z_1, \ldots z_N) \propto p(z_1, \ldots, z_n \mid \theta)\, p(\theta|\alpha) \qquad (22)$$

suit également une loi de Dirichlet. Ceci permettra de simplifier les calculs au moment de l'inférence.

3.5 Probabilité Jointe et Loi Marginale

Etant donnés les paramètres α et β, la probabilité jointe du mélange de topics θ, des N topics z et de N mots w est donnée par :

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1,N} p(z_n|\theta)\, p(w_n|\beta_{zn}) \qquad (23)$$

La loi marginale d'un document w est alors :

$$P(w|\alpha, \beta) = \int(\theta|\alpha) \left(\Pi_{n=1,N} \Sigma\, p(z_n|\theta)\, p(w_n|\beta z_n) \right) d\theta \quad (24)$$

Il suffit de prendre le produit de cette quantité pour chaque document w du corpus pour obtenir la probabilité de ce corpus.

3.6 Inférence et Estimation de Paramètres

On a décrit le modèle du LDA au paragraphe précédent et on a montré son utilité pour l'analyse de corpus de tweets. Mais les variables et les paramètres du modèle ne sont pas connus initialement, et il faut essayer de les apprendre à partir des données observables, c'est à dire les termes qui forment les tweets. Dans la représentation graphique de la Figure 25, on peut voir que les seules variables observées sont les mots $w_{d,n}$, alors que toutes les autres variables sont cachées. Etant donnés les paramètres α et β, le rôle de l'inférence est de déterminer les variables cachées θ et z_n d'un document w, étant donnée la liste des mots w_n du document. Les principales méthodes d'inférence (approchée) pour LDA sont les méthodes de sampling (notamment le collapsed Gibbs sampling) et les méthodes variationnelles (particulièrement les méthodes mean-field, qui peuvent se faire en batch ou en ligne), on peut ensuite faire recours à ce procédé d'inférence pour estimer les paramètres α, β et η du modèle grâce à l'algorithme Expectation Maximisation (EM).

3.7 Inférence

Le problème principal de l'inférence pour LDA est celui de déterminer la distribution à posteriori des variables cachées étant donné le document (et les paramètres α et ß) :

$$p(\theta, z|w,\alpha, \beta) = p(\theta, z, w|\alpha, \beta) / p(w|\alpha, \beta) \quad (25)$$

Cette distribution est malheureusement très difficile à calculer. Une inférence exacte utilisant la distribution à posteriori des paramètres est donc inenvisageable. Mais il existe plusieurs méthodes d'inférence approchée qui peuvent être utilisées pour LDA, par exemple l'approximation variationnelle ou le Markov Chain Monte Carlo [112].

Markov Chain Monte Carlo et Gibbs Sampling

Les méthodes Markov Chain Monte Carlo (MCMC) [112] consistent à construire une chaîne de Markov sur les variables cachées dont la loi stationnaire est la distribution à posteriori cherchée. Rappelons que, une chaîne de Markov est définie par une loi de transition $P(\theta^{t+1}, z^{t+1}|\theta^t, z^t)$ d'un état (θ^t, z^t) au suivant (θ^{t+1}, z^{t+1}), et peut converger vers une loi stationnaire qui

est laissée invariante par la transition. Une fois l'état stationnaire atteint, les échantillons donnés par la chaîne de Markov suivent cette loi stationnaire qui est la distribution voulue.

Le Gibbs Sampling est une méthode MCMC où la transition de la chaîne est donnée par la loi conditionnelle d'une variable cachée étant données les observations et l'état courant des autres variables cachées. Pour LDA, on peut successivement obtenir des échantillons $\theta|w, z$ puis $z_n|\theta$ et w.

Le collapsed Gibbs Sampling échantillonne sur les topics seulement, depuis $z_n|z_{-n}$, w, en intégrant les valeurs θ.

Inférence Variationnelle

L'inférence variationnelle utilise l'optimisation plutôt que l'échantillonnage. L'idée est de commencer par établir une distribution sur les variables cachées avec des paramètres libres (les paramètres variationnels), puis d'optimiser les paramètres variationnels pour que la distribution converge vers la distribution à posteriori souhaitée.

Estimation des Paramètres

Pour estimer les paramètres α et β, on peut utiliser la méthode Empirical Bayes (ou maximum de vraisemblance marginale), qui consiste à chercher des paramètres α et β qui maximisent la log-vraisemblance (marginale) des données :

$$\ell(\alpha, \beta) = \Sigma_{d=1, D} \log p(w_d|\alpha,\beta) \quad (26)$$

On a vu que le calcul de $p(w_d|\alpha, \beta)$ est impossible en pratique, mais l'inférence variationnelle permet d'obtenir la borne inférieure de la log-vraisemblance qui peut être exploité par un algorithme Espérance-Maximisation appelé variational EM, qui utilise en plus les paramètres variationnels de la méthode décrite au paragraphe précédente. L'algorithme itère les deux étapes suivantes : l'étape E qui effectue une inférence variationnelle avec les paramètres α et β courants pour calculer la log-vraisemblance, et l'étape M qui maximise en α et β la borne inférieure du log vraisemblance calculée.

Après avoir donné les concepts de base de LDA, dans les sous-sections suivantes nous présentons en détail son intégration dans le processus de conceptualisation pour améliorer la représentation de texte court en langue arabe dans le cadre d'une application d'analyse des sentiments.

4 Intégration de LDA dans la Conceptualisation de Texte Court

Dans cette section, nous présentons en détail notre deuxième méthode proposée pour la conceptualisation de texte court en arabe et qui tend principalement à résoudre le problème de couverture de la langue arabe dans les ressources sémantiques. Pour obtenir une représentation basée sur les concepts (BoC), nous avons déjà noté qu'il faut tout d'abord passer par la représentation basée sur les termes (BoW) afin de trouver les termes candidats à projeter dans la ressource sémantique utilisée.

Dans cette première étape, chaque texte sera transformé en texte brut, nettoyé en supprimant les mots latins, les caractères spéciaux tels que (/, #, $, etc…) et les mots vides arabes considérés comme non pertinents (prépositions et particules). L'étape suivante consiste à extraire la racine (root) / radical (stem) du mot afin d'effectuer une certaine forme de normalisation morphologique pour créer des caractéristiques plus générales.

Pour surmonter les faiblesses de la représentation BoW, nous proposons d'utiliser une représentation BoC en utilisant BabelNet comme étant une ressource sémantique largement utilisée. La figure 27 présente l'organigramme illustrant les étapes de notre méthode proposée pour la conceptualisation. Au cours de conceptualisation, différentes stratégies peuvent être mises en œuvre comme il a été décrit précédemment dans le chapitre 5 (ajout de concepts, conceptualisation partielle et conceptualisation complète). La mise en œuvre de la représentation BoC repose sur deux étapes :

La première étape consiste à mapper (projeter) les termes dans BabelNet afin de trouver les concepts candidats pour représenter chacun des termes. Cependant, il se peut que la recherche des concepts correspondants à un terme donné ne retourne aucun concept vu le problème de couverture discuté auparavant. Pour surmonter ce problème, nous suggérons d'intégrer la modélisation des topics en tant qu'outil permettant d'identifier les termes ayant les mêmes liens sémantiques en utilisant le model LDA. Pour cette raison, si ce terme n'existe pas dans la ressource sémantique, nous proposons de rechercher le concept correspondant au terme le plus proche dans le même topic et ainsi de suite jusqu'à ce que le terme soit associé à un concept.

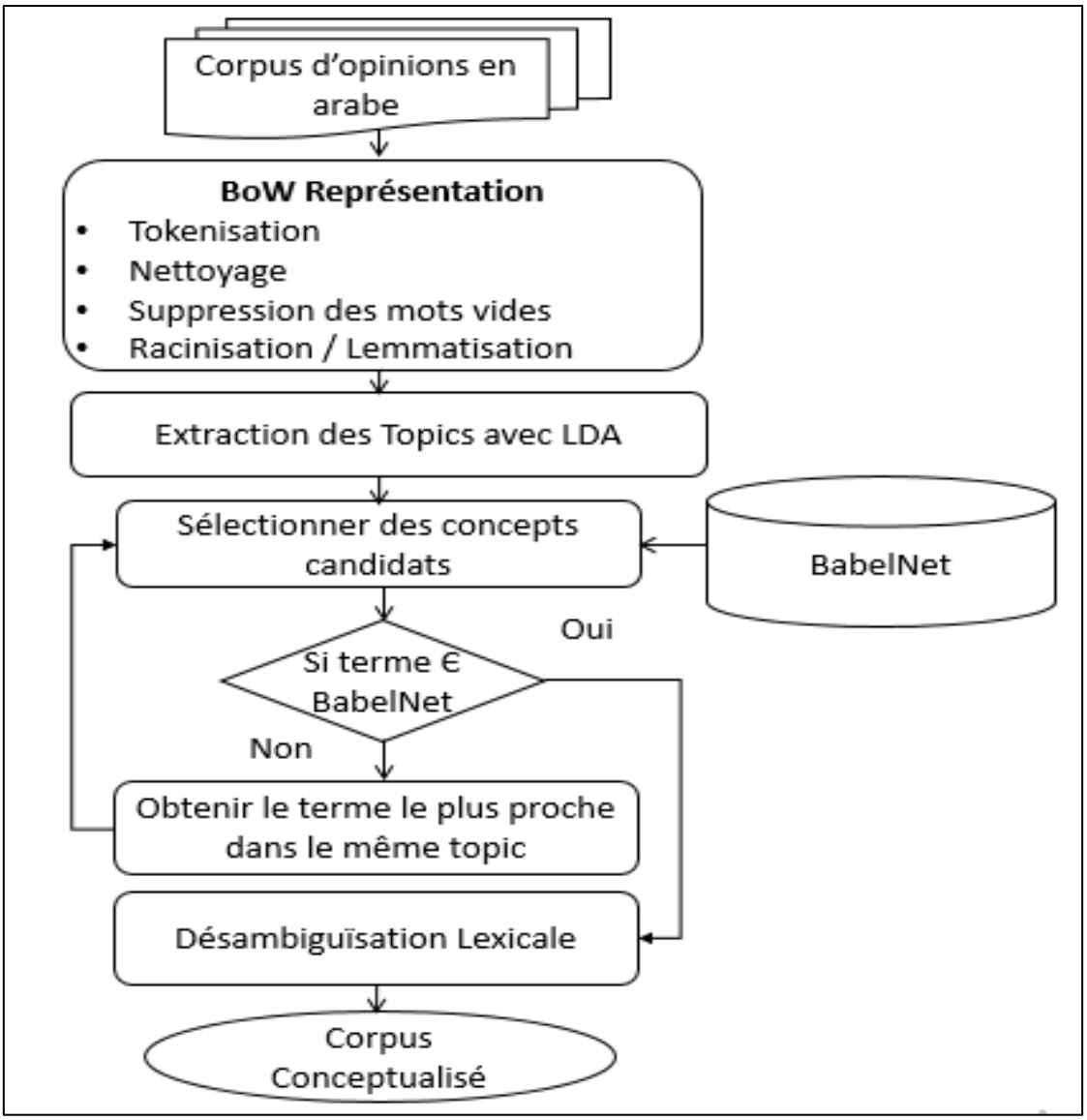


Figure 27 . Méthode de conceptualisation de texte court pour l'analyse des sentiments en langue arabe

En outre, lors de la recherche des concepts correspondants à des termes polysémiques, de multiples correspondances peuvent être détectées, et par conséquent, elles peuvent introduire des ambiguïtés dans la représentation du document final. Par la suite, attribuer un terme à un concept est une étape cruciale lors du processus de conceptualisation. Trois stratégies de désambiguïsation peuvent être utilisées telles que nous avons introduit dans ce rapport (premier concept, tous les concepts, et le concept de contexte).

Dans notre proposition, nous avons adopté la stratégie basée sur le contexte où nous calculons la similarité entre le terme à mapper avec leurs termes voisins et la définition du concept. Finalement le concept avec la similarité maximale est le concept représentant du terme. L'algorithme de conceptualisation peut être résumé et présenté ci-dessous (algorithme 4).

Afin de tester l'efficacité de notre méthode proposée et confirmer que l'utilisation de LDA dans le processus de conceptualisation permet de résoudre le problème de la couverture en langue arabe. Une série d'expérimentations ont été menées à l'aide du corpus LARB [101], qui est le plus grand ensemble de données d'analyse de sentiment pour la langue arabe. Nous avons utilisé trois des algorithmes d'apprentissage automatique les plus populaires pour la catégorisation de texte : NB [87], SVM [88] [89] et DT [98]. Les corpus utilisés avec les résultats obtenus seront présentés et discutés dans la section suivante.

Algorithme 4 : Conceptualization Process

```
Entrée (Corpus d'opinions wd, Liste des Topics tl)
Sortie Corpus Conceptualisé
Début
Pour chaque document w dans wd faire
  Pour chauque term t dans w faire
    cc ← la liste des concepts candidats retournés à partir de BabelNet pour le terme t
    Si cc est Null Alors
      wt ← le terme le plus proche de t dans le même topic
      Tant que wt contient un terme tp Et cc est Null
        cc ← la liste des concepts retournés à partir de BabelNet pour le terme tp
    Fin Tant que
    Fin Si

    // Processus de WSD
    res ← le concept retourné pour le terme t
    txt1 ← les termes voisins du terme t
    Pour chauqe concept c dans cc faire
      txt2 ← la définition du concept c
      sim ← Similarity(txt1, txt2)
      Si sim > maxSim Alors
        maxSim ← sim
        res ← c
      Fin Si
    Fin Pour
  Fin Pour
Fin Pour
Fin
```

5 Expérimentations et Résultats

Dans cette section, nous présentons une description détaillée des expérimentations que nous avons mené afin de valide notre contribution, ainsi que le corpus utilisé et les résultats obtenus.

5.1 Description des Expérimentations

Afin de mettre en évidence l'intérêt de notre méthode proposée pour améliorer la représentation de texte court en arabe à l'aide de BoC et améliorer ainsi la performance de notre système d'analyse des sentiments en arabe, une série d'expérimentations a été bien menée (Figure 28).

Au cours du processus de conceptualisation, nous avons utilisé les trois différentes stratégies décrites précédemment (ajout de concepts, conceptualisation partielle et conceptualisation complète). De plus, lors de la phase de la désambiguïsation lexicale (WSD), nous avons adopté la stratégie de concept de contexte. Par ailleurs, pour valider que notre contribution basée sur LDA conduit à la résolution du problème de couverture de la langue arabe, nous avons réalisé la conceptualisation sans et avec l'intégration du LDA.

Dans la section suivante, nous décrivons le corpus utilisé dans nos expérimentations.

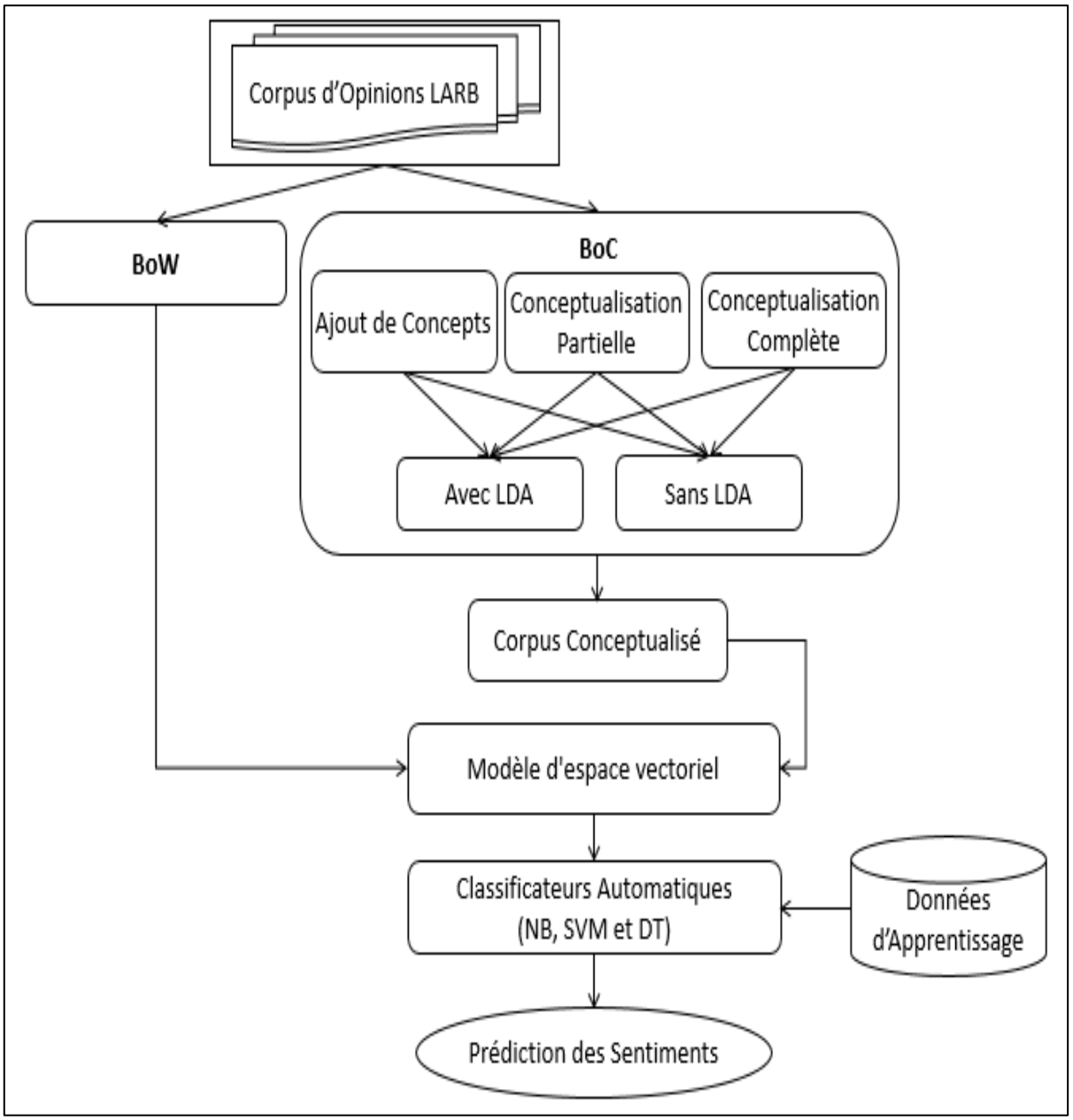


Figure 28 . Description des expérimentations de l'analyse des sentiments en langue arabe

5.2 Corpus utilisé

Le corpus utilisé dans nos expérimentations est un sous-ensemble du corpus *« the Large-Scale Arabic Book Reviews Dataset (LARB dataset) »* [101]. Ce dernier contient plus de 63 000 opinions concernant des livres en arabe. Les opinions ont été tirées du site Web *«www.goodreads.com»* au cours du mois de mars 2013. Le corpus est classé en trois catégories : positive, négative et neutre. Tableau 23 décrit le corpus utilisé.

Tableau 23 . Description du corpus utilisé pour l'analyse des sentiments

Catégories	Nombre de textes
Positive / ايجابي	5275
Négative / سلبي	2110
Neutre / محايد	823
Total	**8208**

5.3 Résultats & Discussion

L'efficacité de notre système a été évaluée et comparée en termes de la mesure f1-mesure à l'aide des classificateurs NB, SVM et DT en utilisant une technique de validation croisée sur *10 Fold*. Le tableau 24 montre les résultats obtenus en utilisant les trois classificateurs. Nous avons testé la catégorisation avec une représentation BoW et en utilisant chacune des stratégies de conceptualisation présentées précédemment.

La stratégie de conceptualisation partielle combinée avec LDA comme étant un outil efficace pour surmonter le problème de couverture de la langue arabe dans BabelNet permettant d'obtenir la f1-mesure la plus élevée par rapport à la représentation de base (BoW). L'amélioration est significative parce que d'un côté, BoC supprime les problèmes introduits par la synonymie et la polysémie et augmente donc les performances du classificateur. De l'autre côté, LDA permet d'associer les termes n'ont pas de concepts correspondants dans BabelNet avec les concepts des termes ayant avec eux une forte liaison sémantique dans le même corpus. Grâce à cette stratégie, nous pouvons enrichir la représentation conceptuelle des opinions et par la suite améliorer la performance de notre système d'analyse des sentiments en arabe. La f1-mesure la plus basse a été obtenue par la stratégie de conceptualisation complète, même si l'utilisation de la LDA augmente relativement les résultats de cette stratégie. Ceci est dû à l'augmentation de l'espace de représentation.

Tableau 24 . F1-mesure pour l'analyse des sentiments en langue arabe

	NB		DT		SVM	
	F1-measure	Baseline Gap	F1-measure	Baseline Gap	F1-measure	Baseline Gap
BoW (Baseline)	0,754		0,728		0,742	
Conceptualisation Complète sans LDA	0,705	-4,90%	0,677	-5,10%	0,708	-3,40%
Conceptualisation Complète avec LDA	0,738	-1,60%	0,709	-1,90%	0,729	-1,30%
Conceptualisation Partielle sans LDA	0,778	+2,40%	0,747	+1,90%	0,768	+2,60%
Conceptualisation Partielle avec LDA	0,817	+6,30%	0,779	+5,10%	0,803	+6,10%
Conceptualisation avec Ajout de Concept sans LDA	0,769	+1,50%	0,739	+1,10%	0,759	+1,70%
Conceptualisation avec Ajout de Concept avec LDA	0,793	+3,90%	0,756	+2,80%	0,786	+4,40%

Rappelons que le nombre de topics est un paramètre très décisif de LDA. Pour étudier l'impact de ce paramètre sur la performance des résultats de notre système d'analyse des sentiments. Dans le cadre de ce travail de thèse, nous avons également mené plusieurs expérimentations pour tester l'impact de la modification de ce nombre sur les résultats de la conceptualisation. Dans ce sens, nous avons effectué la conceptualisation avec des valeurs différentes de nombre de topics (5, 10, 15 et 20 topics).

La figure 29 présente les résultats obtenus tout en modifiant le nombre de topics. Nous pouvons noter que généralement le changement de nombre de topics n'influence pas sur les résultats de la catégorisation. Parce que dans notre contribution nous n'apportons aucune modification sur l'algorithme LDA. Si nous ne trouvons pas de terme dans BabelNet, nous prenons les termes les plus proches dans le même topic un par un jusqu'à trouver un concept pour le représenter. Le nombre de topics n'est pas important dans notre cas ; ce qui nous intéresse, c'est que les termes d'un même topic partagent de nombreux liens sémantiques dans le même corpus.

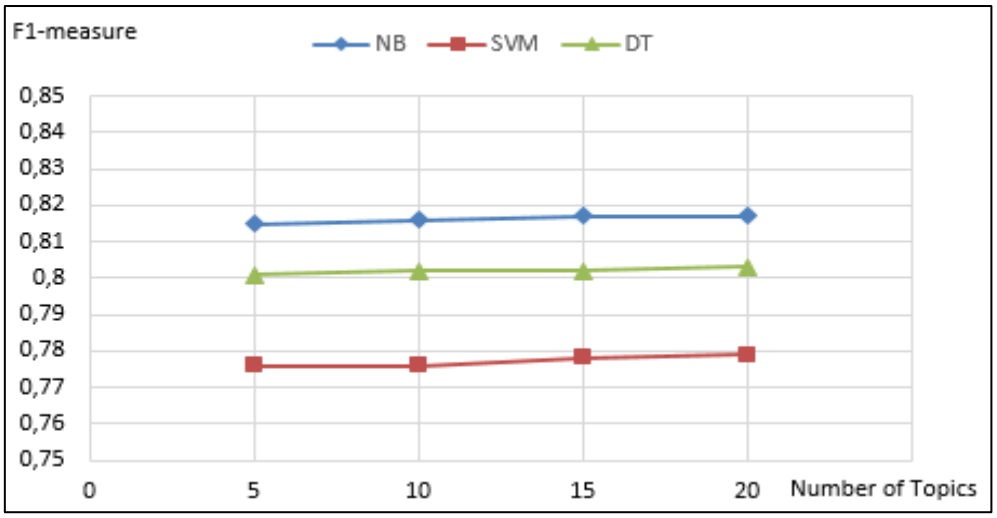


Figure 29 . F1-mesure avec des valeurs différentes de Topics (5, 10, 15 et 20)

En plus, les résultats obtenus avec les trois classificateurs (NB, SVM et DT) montrent l'intérêt de notre proposition.

6 Conclusion

Dans ce chapitre, nous avons présenté une quatrième contribution pour enrichir la représentation du texte court en arabe en se basant sur une représentation conceptuelle, ce qui transforme la représentation des données textuelles en une structure plus courte, plus compacte et plus prédictive. Cependant, pour la langue arabe, la majorité des ressources sémantiques sont des projets incomplets ; cela pose un sérieux problème de taux de couverture pour le cas de la langue arabe par rapport aux autres langues. Partant de l'hypothèse que les termes appartenant au même topic (sujet) partagent de nombreux liens sémantiques dans un corpus, leurs concepts correspondants partageront presque les mêmes liens sémantiques dans le même corpus. De ce fait, nous avons proposé d'intégrer la modélisation par thèmes comme un moyen permettant de rassembler les termes ayant les mêmes liens sémantiques. Pour montrer l'efficacité de notre méthode, nous avons choisi l'une des tâches les plus importantes récemment : l'analyse des sentiments ou bien fouille d'opinions. Les résultats obtenus ont montré l'intérêt de notre proposition.

Conclusion Générale et Perspectives

De nos jours, environ un milliard d'utilisateurs interagissent quotidiennement sur les réseaux sociaux en ligne, où ils partagent des informations et discutent d'une grande variété de sujets. Ces réseaux sociaux permettent aux utilisateurs de publier généralement du texte court qui se distinguent des documents traditionnels par leur brièveté et le manque d'informations contextuelles, en outre la langue arabe complique encore plus ces problèmes vu la richesse de sa morphologie. Par conséquent, les techniques traditionnelles de représentation de texte seront inadéquates et peuvent avoir un impact négatif sur la performance de toute application de fouille de texte. L'enrichissement du texte court semble être un bon moyen pour surmonter ces problèmes et améliorer la représentation de texte court.

Cette thèse s'inscrit dans un le cadre d'un projet qui vise à améliorer le rendement des applications d'analyse du contenu des réseaux sociaux en langue arabe. Pour cette raison, nous nous sommes intéressés, dans notre travail, par l'enrichissement de texte court à travers deux modèles de représentation : la représentation basée sur les termes et celle basée sur les concepts.

Concernant la représentation basée sur les termes, nous avons pu enrichir le texte court à travers la théorie des ensembles approximatifs (RST) permettant de renforcer les liens sémantiques entre les documents d'une même collection de donnée. Ensuite, nous avons utilisé le Web comme étant un corpus large et dynamique afin de prendre en considération le contexte du texte court et de l'enrichir avec des termes d'une forte liaison sémantique avec le contenu initiale. Par ailleurs et pour la représentation conceptuelle, nous avons proposé une mesure de similarité (*Web-RST*) entre une paire de texte court dédiée à la langue arabe afin de surmonter le problème de désambiguïsation lexicale (WSD) qui reste une étape primordiale dans le processus de conceptualisation et qui tend à attribuer un terme au concept correspondant. Finalement, pour pallier au problème de taux de couverture de la langue arabe dans les ressources sémantique et partant de l'hypothèse que les termes appartenant au même sujet (*topic*) partagent de nombreux liens sémantiques dans un corpus, leurs concepts correspondants partagent aussi les mêmes liens sémantiques dans le même corpus. Nous avons proposé d'intégrer la modélisation par thème (*Topic Modeling*) à travers la méthode LDA.

Afin de tester et valider nos contributions, nous avons opté vers deux parmi les applications les plus répondues : la catégorisation de texte court et l'analyse des sentiments. Les résultats obtenus montrent l'intérêt de nos contributions.

Perspectives :

Les différentes voies explorées dans le cadre de cette thèse débouchent sur plusieurs perspectives. Nous présentons ci-dessous celles qui nous paraissent les plus prometteuses :

Comme premier perspective, vu que une grande partie du contenu des réseaux sociaux est présenté sous forme d'un langage dialectale, nous visons de mener quelques investigations afin d'intégrer ainsi dans nos propositions l'étude de texte dialectale.

Comme deuxième perspective, nous envisageons de concrétiser nos propositions d'amélioration de la représentation de texte court en arabe et utiliser ces résultats pour aborder d'autres axes de recherche, voir les systèmes de recommandation, détection des communautés.

Comme troisième perspective, en raison que la majorité de nos contributions fait référence à la langue arabe seulement au bloc de prétraitement, nous planifions de mettre en place nos contributions avec d'autre langues dans l'intention de proposer des systèmes multilingue.

Bibliographie

[1] R.Feldman and I.Dagan, Kdt : Knowledge discovry in texts: In Proc. of the First Int, Conf. onKnowledge Discovry KDD, Pages 112_117, 1995.

[2] V.BUS, 'As we may think: The Atlantic Monthly, Trad. Ch. Monnatte, juillet 1945.

[3] A. F. Smeaton, “Progress In The Application Of Natural Language Processing To Information Retrieval Tasks,” Comput. J., Vol. 35, No. 3, PP. 268–278, 1992.

[4] K.Darwish, Building a shallow Arabic morphological analyzer in one day: Proc. Of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02) (pp. 1–8). Philadelphia, Pennsylvania, United States, 2002.

[5] G.Salton and M.J.McGill, Introduction to modern information retrieval: Singapore: McGraw-Hill International Book Co, 1987.

[6] G.Salton and C.Bukley, Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management: 24(5), pp. 513-523, 1988.

[7] Jaillet S, Catégorisation automatique de documents: LIRMM.

[8] B.Piwowarski, Technique d'apprentissage pour le traitement d'informations structurées: application à la recherche d'information: Thèses de Doctorat; Université de Paris; France, le 17 juillet 2003.

[9] D.Hawking, Overview of the TREC-8 Web Track. Nov, 1999.

[10] J.Ying, Analysis and Comparison of Existent Information Extraction Methods: Diploma Thesis, Darmstadt University of Technology (TUD), Darmstadt, September 2006.

[11] O. Jin, N. N. Liu, K. Zhao, Y. Yu, Q. Yang, Transferring topical knowledge from auxiliary long texts for short text clustering, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 775-784.

[12] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi, Q.-T. Ha, A hidden topic-based framework toward building applications with short web documents, IEEE Transactions on Knowledge and Data Engineering 23 (7) (2011) 961-976.

[13] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, H. Hao, Semantic expansion using word embedding clustering and convolutional neural network for improving short text classi_cation, Neurocomputing 174 (2016) 806-814.

[14] El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2018). Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. Journal of Information Science,

[15] Kuzi S, Shtok A and Kurland O. Query expansion using word embeddings. In Proceedings of the 25th ACM international on conference on information and knowledge management, Indianapolis, IN, 24–28 October 2016, pp. 1929–1932. New York: ACM.

[16] Mahgoub AY, Rashwan MA, Raafat H et al. (2014). Semantic query expansion for Arabic information retrieval. In: Proceedings of the Arabic natural language processing workshop, pp. 87–92.

[17] W. Guo, H. Li, H. Ji, and M. Diab, (3013) "Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 239–249

[18] S. R. Yerva, Z. Mikl´os, and K. Aberer, (2011), "What have fruits to do with technology? : the case of orange, blackberry and apple", in Proceedings of the International Conference on Web Intelligence, Mining and Semantics.

[19] Y. S. Yegin Genc and J. V. Nickerson, (2011), "Discovering context: Classifying tweets through a semantic transform based on wikipedia," in Proceedings of HCI International,

[20] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. (2010). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 841–842. ACM

[21] Neelam Choudhary and Ankit Kumar Jain. 2017. Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique. D. Singh et al. (Eds.): ICAICR 2017, CCIS 712, pp. 18–30, 2017.

[22] Ezpeleta, E., Zurutuza, U., & Hidalgo, J. M. G. 2016. Short messages spam filtering using personality recognition. Proceedings of the 4th Spanish Conference on Information Retrieval, CERI '16, ACM, New York, NY, USA, pp. 7:1–7:7.

[23] Jialin, M., Zhang, Y., Liu, J., Yu, K., Wang, X.: Intelligent SMS spam filtering using topic model. In: International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 380–383. IEEE (2016). doi:10.1109/INCoS.2016.47

[24] Jiaying Song, Yu He, Guohong Fu, (2015), Polarity Classification of Short Product Reviews via Multiple Cluster-based SVM Classifiers. 29th Pacific Asia Conference on Language, Information and Computation, pp 267 - 274

[25] Kang Liu, Liheng Xu and Jun Zhao. (2014). Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp 314–324

[26] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web, pages 519–528. ACM, 2003.

[27] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics, 2002.

[28] Alayba, A., Palade, V., England, M., & Iqbal, R. (2018). A Combined CNN and LSTM Model for Arabic Sentiment Analysis. In Cross Domain Conference for Machine Learning and Knowledge Extraction: CD-MAKE

[29] Tartir, S., Abdul-Nabi, I. (2017). A Semantic Sentiment Analysis in Arabic Social Media. Journal of King Saud University - Computer and Information Sciences 29(2), 229– 233.

[30] M. El-Masri, N. Altrabsheh, H. Mansour and A. Ramsay, (2017). A web-based tool for Arabic sentiment analysis. Procedia Computer Science, Vol 117, pp 38-45

[31] A. A. Al Sallab, R. Baly, and H. Hajj, (2015) "Deep Learning Models for Sentiment Analysis in Arabic," pp. 9–17.

[32] Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language, 28(1), 20–37.

[33] Kundu, S., Srijith, P. ., & Desarkar, M. S. (2018). Classification of Short-Texts Generated During Disasters: A Deep Neural Network Based Approach. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[34] Liu, Y., & Li, D. (2018). Research on User Gender Prediction of Chinese Microblog Based on Short Text Analysis. 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC).

[35] Hucko, M., Gaspar, P., Pikuliak, M., Triglianos, V., Pautasso, C., & Bielikova, M. (2018). Short Texts Analysis for Teacher Assistance During Live Interactive Classroom Presentations. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA).

[36] Chen, D., & Wang, H. (2018). Research on Short Text Classification Algorithm Based on Neural Network. 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC).

[37] Sotthisopha, N., & Vateekul, P. (2018). Improving Short Text Classification Using Fast Semantic Expansion on Multichannel Convolutional Neural Network. 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

[38] W. E. Winkler, (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage"

[39] Huang, Y.-F., & Huang, C.-T. (2015). Mining Domain Information from Social Contents Based on News Categories. Proceedings of the 19th International Database Engineering & Applications Symposium on - IDEAS '15.

[40] Fang Wang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. (2014). Concept-based short text classi_cation and ranking. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp1069-1078

[41] Walid Magdy, Hassan Sajjad, Tarek El-Ganainy, and Fabrizio Sebastiani. (2015). Distant supervision for tweet classification using YouTube labels. In Ninth International AAAI Conference on Web and Social Media

[42] Ramakrishna B. Bairi, Raghavendra Udupa, and Ganesh Ramakrishnan. (2016). A framework for task-specific short document expansion. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM'16), pp 791-800

[43] Miles Efron, Peter Organisciak, and Katrina Fenlon. (2012). Improving retrieval of short texts through document expansion. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, pp 911-920.

[44] L. Sang, F. Xie, X. Liu, and X. Wu. (2016). Wefest: Word embedding feature extension for short text classi_cation. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp 677-683.

[45] Flisar, J., & Podgorelec, V. (2018). Document Enrichment using DBPedia Ontology for Short Text Classification. Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics - WIMS ’18.

[46] Yu, Z., Wang, H., Lin, X., & Wang, M. (2016). Understanding Short Texts through Semantic Enrichment and Hashing. IEEE Transactions on Knowledge and Data Engineering, 28(2), 566–579.

[47] G. McDonald, R. Deveaud, R. Mccreadie, C.Macdonald, and I. Ounis. (2015). Tweet enrichment for effective dimensions classification in online reputation management, International AAAI Conference on Web and Social Media, North America

[48] W. Meng, L. Lanfen, W. Jing, Y. Penghua, L. Jiaolong, and X. Fei. (2013). Improving short text classification using public search engines. In International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, pages 157-166. Springer

[49] J. Tang, X. Wang, H. Gao, X. Hu, and H. Liu. (2012). Enriching short text representation in microblog for clustering. Frontiers of Computer Science, 6(1), pp 88-101

[50] Z. Dai, A. Sun, and X. Y. Liu. (2013). Crest: Cluster-based representation enrichment for short text classification. In Advances in Knowledge Discovery and Data Mining, pages 256-267. Springer

[51] F. Abel, Q. Gao, G. J. Houben, and K. Tao. (2011). Semantic enrichment of twitter posts for user pro_le construction on the social web. In The Semanic Web: Research and Applications, pp 375-389. Springer

[52] M. CALISAN, C. O. SAKAR. (2017). Classification of Short-Texts by Utilizing an External Knowledge Source. Journal of Science and Engineering. Vol 19

[53] W. Li, & J. Zhao. (2016). TextRank Algorithm by Exploiting Wikipedia for Short Text Keywords Extraction. 2016 3rd International Conference on Information Science and Control Engineering (ICISCE)

[54] P.n Nand, R. Perera, and G. Klette. (2015). A tweet classification model based on dynamic and static component topic vectors. In Australasian Joint Conference on Artificial Intelligence, pages 424{430. Springer

[55] A. M. Raez, E. M. Camara, M. T. M. Valdivia, and L. A. U. Lopez. (2014). A knowledge-based approach for polarity classification in twitter. Journal of the Association for Information Science and Technology, 65(2):414-425

[56] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. (2016). Topic modeling for short texts with auxiliary word embeddings. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 165-174

[57] Hu X, Sun N, Zhang C, Chua TS. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. Proceedings of the 18th ACM Conference on Information and Knowledge Management; New York: ACM. pp. 919–28

[58] E, Gabrilovich. S, Markovitch. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. Twenty-First National Conference on Artificial Intelligence, vol 2, pp. 1301-1306

[59] Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. Cognitive Science Proceedings (LEA), 660-665.

[60] Wen-tau Yih and Christopher Meek. (2007). Improving Similarity Measures for Short Segments of Text. Proceeding AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence, V 2, pp. 1489-1494

[61] Jones, R.; Rey, B.; Madani, O.; and Greiner, W. (2006). Generating query substitutions. Proc. of WWW conference

[62] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. (1990). Indexing by Latent Semantic Analysis. Journal of the Society for Information Science, vol. 41, no 6, pp 391-407

[63] Thomas Hofmann. (1999). Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)

[64] D. M. Blei, Andrew Y. Ng, Michael I. Jordan. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3. pp 993-1022

[65] Adel, N., Crockett, K., Crispin, A., Chandran, D., & Carvalho, J. P. (2018). FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).

[66] E. M. B. Nagoudi. D. Schwab. 2017. Semantic Similarity of Arabic Sentences with Word Embeddings. Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), Valencia, Spain, pp. 18–24

[67] Nakamura, T., Shirakawa, M., Hara, T., & Nishio, S. (2014). Semantic Similarity Measurements for Multi-lingual Short Texts Using Wikipedia. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).

[68] Shirakawa, M., Nakayama, K., Hara, T., & Nishio, S. (2013). Probabilistic semantic similarity measurements for noisy short texts using Wikipedia entities. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM

[69] Sahami, M., and Heilman, T. 2006. "A web-based kernel function for measuring the similarity of short text snippets". In Proc. of WWW 2006.

[70] M. H. Maâloul, "Approche Hybride Pour Le Résumé Automatique De Textes. Application A La Langue arabe," Phd Thesis, Université De Provence - Aix-Marseille I, 2012.

[71] S. Boulaknadel, "Traitement Automatique Des Langues Et Recherche D'information En Langue arabe Dans Un Domaine De Spécialité: Apport Des Connaissances Morphologiques Et Syntaxiques Pour L'indexation," Phd Thesis, Université De Nantes, 2008.

[72] F. Chaar, "Traitement Automatique De La Langue arabe : La Modération En Arabe," Institut National Des Langues Et Civilisations Orientales (I.N.A.L.C.O), 2009.

[73] F. S. Douzidia, "Résumé Automatique De Texte Arabe," Master Thesis, Université De Montréal, 2004.

[74] S. Baloul, M. Alissali, M. Baudry, And M. Boula De, " Interface Syntaxe-Prosodie Dans Un Système De Synthèse De La Parole A Partir Du Texte En Arabe," In Journées D'etudes Sur La Parole (Jep'02), 2002, PP. 329–332.

[75] K. S. Motaz, "The Impact Of Text Preprocessing And Term Weighting On Arabic Text Classification," Master Thesis, The Islamic University – Gaza, Palestine, 2010.

[76] L. H. Belguith, C. Aloulou, And A. Ben Hamadou, "Maspar : De La Segmentation A L'analyse Syntaxique De Textes Arabes," Rev. Inf. Interact. Intell. I3, Vol. 7, PP. 9–36, 2001.

[77] R. Ouersighni, "A Major Offshoot Of The Diinar-Mbc Project: Araparse, A Morphosyntactic Analyzer For Unvowelled Arabic Texts," In Acl/Eacl 2001 Workshop On Arabic Language Processing, 2001, PP. 9–16.

[78] Lovins, J. B. "Development of a Stemming Algorithm." Mechanical Translation and Computational Linguistics 11, 1968, 22—31.

[79] R.Garside S.Khoja, Stemming Arabic text: Computer Science Department, Lancaster University, Lancaster, UK, 1999.

[80] G.Kanaan, H.Al-Serhan R.Al-Shalabi, New approach for extracting Arabic roots: In the Int. Arab Conf. on Information Technology ACIT'2003, Egypt, 2003.

[81] Boudchiche, M.; Mazroui, A.; Ould Abdallahi Ould Bebah, M.; Lakhouaja, A.; Boudlal, A.; 2017. "AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer", Journal of King Saud University – Computer and Information Sciences. 29(2). pp. 141-146.

[82] A.Anil.Kumar, Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering: International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 5, July – 2012.

[83] Tim-Buckwalter, Issues in Arabic orthography and morphology analysis: In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Semitic '04, pages 31–34, Stroudsburg, PA, USA, Association for Computational Linguistics, 2004.

[84] Brihaye, P. (2003). AraMorph. Retrieved April 23, 2015

[85] Ramzi Abbes, « la conception et la realization d'un concordancier électronique pour l'arabe », Doctoral Thesis, Institut Nationale des Sciences Appliquées, Lyon, 2004

[86] Rehab Nasser, Al-Wehaibi, Muhammad Badruddin, Khan "Understanding the Content of Arabic Tweets by Data and Text Mining Techniques", Symposium on Data Mining and Applications (SDMA2014)

[87] Rish, Irina. (2001). "An empirical study of the naive Bayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.

[88] Vapnik V. (1995). The Nature of Statistical Learning Theory, chapter 5. Springer-Verlag, New York.

[89] Joachims T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In Proceedings of the European Conference on Machine Learning (ECML), 1998, pp.173-142

[90] Adrzej Skowron, J. S. Tolerance approximation spaces. Fundamenta Informaticae 27, 2-3 (1996), 245-253.

[91] Jan Komorowski, Zdzislaw Pawlak, L. P. A. S. Rough sets: a tutorial, 1998.

[92] Pawlak, Z. Rough sets: Theoretical aspects of reasoning about data. Kluwer Dordrecht, 1991.

[93] Saori Kawasaki, Ngoc Binh Nguyen, T. B. H. Hierarchical document clustering based on tolerance rough set model. In Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000, Proceedings (2000), D. A. Zighed, H. J. Komorowski, and J. M. Zytkow, Eds., vol. 1910 of Lecture Notes in Computer Science, Springer.

[94] Tu Bao Ho, N. B. N. Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems 17, 2 (2002), 199{212.

[95] Zdzislaw Pawlak, A. S. Rough membership functions. In Advances in the Dempster-Shafer Theory of Evidence (New York, 1994), J. K. M. Fedrizzi and R. Yager, Eds., John Wiley & Sons, pp. 251-271.

[96] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259. (2003)

[97] Rehab Nasser, Al-Wehaibi*, Muhammad Badruddin, Khan "Understanding the Content of Arabic Tweets by Data and Text Mining Techniques", Symposium on Data Mining and Applications (SDMA2014)

[98] Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

[99] Shereen Albitar, Sébastien Fournier, and Bernard Espinasse. 2012. The Impact of Conceptualization on Text Classification" WISE 2012, LNCS 7651, pp. 326–339

[100] Larkey L., L. Ballesteros, and M. E. Connell. 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. Proceedings of SIGIR'02, pp. 275–282.

[101] Mohamed Aly and Amir Atiya. 2013. LABR: Large-scale Arabic Book Reviews Dataset. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pp. 494–498

[102] D. Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

[103] L. Abouenour, K. Bouzoubaa, P. Rosso. On the evaluation and improvement of Arabic WordNet coverage and usability. Lang. Resour. Eval. (2013), pp. 1-27

[104] https://babelnet.org/stats (accédé le 05/09/218)

[105] J. Wang, M. Bansal, K. Gimpel, B. Ziebart, and T. Clement. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. TACL, 3:59–71.

[106] George A. Miller. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[107] R. Navigli and S. Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.

[108] G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975

[109] Ngo Chi Lang.A tolerance rough set approach to clustering web search results. Poland: Warsaw University. (2003)

[110] S. Geman and D. Geman, IEE Transactions on Pattern Analysis and Machine Intelligence 6, 721 (1984)

[111] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B. 39 (1): 1–38.

[112] Walsh, Bruce. 2004. Markov Chain Monte Carlo and Gibbs Sampling. Lecture Notes for EEB 581, University of Arizona. http://nitro.biosci.arizona.edu/courses/EEB581-2004/handouts/Gibbs.pdf.

[113] Karen Spärck Jones, « A statistical interpretation of term specificity and its application in retrieval », Journal of Documentation, vol. 28, no 1, 1972, p. 11–21 (DOI 10.1108/eb026526

[114] Hanane Froud. Contributions au Text Minning sur la langue Arabe: application au clustering des Documents Texte arabe. Thèse de doctorat. Soutenue à la FST-Fès, le 22 février 2014.