Université Echahid Cheikh Larbi
Tébessi – Tébessa
Faculté des Sciences Exactes, Science
de la Nature et de la Vie
Département des Mathématiques et Informatique

LAMIS
Laboratoire de Mathématiques,
Informatique et Systèmes

THÈSE

Pour obtenir le grade de docteur
3ème cycle L.M.D. en informatique
Option : Systèmes d'Information

Thème

Deep Learning pour un système d'intégrité et d'authentification des textes numériques du Saint Coran

Présentée par : **TOUATI HAMAD Zineb**

Soutenue le : ../../ 2024, devant le jury composé de :

Pr. Makhlouf Derdour (Professeur), *Université Larbi Ben Mhidi, Oum El Bouaghi,*	Président ;
Pr. LAOUAR Mohamed Ridda (Professeur), *Université Larbi Tébessi, Tébessa,*	Encadreur ;
Dr. BENDIB Issam (MCA), *Université Larbi Tébessi, Tébessa,*	Co-Encadreur ;
Pr. BENDJENNA Hakim (Professeur), *Université Larbi Tébessi, Tébessa,*	Examinateur ;
Dr. TALBI Hichem (MCA), *U niversité Abdelhamid Mehri - Constantine 2, Constantine*	Examinateur.

Année universitaire : 2023/2024

Deep Learning pour un système d'intégrité et d'authentification des textes numériques du saint coran

TOUATI HAMAD Zineb
zineb.touati-hamad@univ@tebessa.dz

Laboratoire de Mathématiques, Informatique et Systèmes (LAMIS),
Université Larbi Tébessi - Tébessa, Algerie.

LAMIS

Laboratoire de Mathématiques,
Informatique et Systèmes

بسم الله الرحمن الرحيم

وَقُل رَّبِّ زِدْنِي عِلْمًا

Et dis: "Ô mon Seigneur, accroît mes connaissances!"

Coran : Sourate 20, Verset 114

Cette thèse est dédiée à ma mère, figure inspirante de persévérance et de foi en l'accomplissement ;

Remerciements

Pendant la préparation de cette thèse, j'ai été en contact avec de nombreux chercheurs, académiciens et praticiens, dont les contributions ont enrichi ma compréhension et mes idées. Je tiens particulièrement à remercier mon directeur de thèse, le Pr. Mohamed Ridda Laouar, pour son encouragement, ses conseils avisés, ses critiques constructives et son amitié. Mon immense gratitude va également à mon co-directeur de thèse, le Dr. Issam Bendib, pour ses orientations, ses conseils et son soutien constant. Sans leur accompagnement attentif et leur soutien continu, cette thèse ne serait pas ce qu'elle est aujourd'hui.

Je tiens également à exprimer ma sincère reconnaissance au jury pour avoir accepté d'évaluer cette étude. Mes remerciements vont à mon ancien enseignat, le Pr. Dardour Mekhlouf, au directeur du laboratoire, le Pr. Benjanna Hakim, ainsi qu'au l'enseignat invité, le Dr. Talbi Hichem, pour leur expertise et leurs précieuses contributions. Vos expériences et vos avis sont essentiels pour améliorer ce travail.

Résumé :

Le domaine de l'analyse du contenu en ligne sensible évolue rapidement, représentant un vaste champ de recherche qui exige une attention considérable. Avec la prolifération du contenu numérique sur diverses plates-formes en ligne, les questions de l'authenticité et de l'intégrité sont devenues de plus en plus prégnantes. Ces facteurs soulignent le besoin critique d'efforts de recherche complets visant à évaluer et à améliorer les performances et l'efficacité des méthodes utilisées pour évaluer le contenu numérique, en particulier sous forme de texte numérique. Au milieu de la profusion de contenu numérique disponible sur Internet, le cas du Saint Coran constitue un point focal convaincant pour la recherche. Alors que les algorithmes dominent tous les aspects du contenu numérique, il est ironique de constater que le contenu coranique arabe reste largement sous-exploité en termes de linguistique informatique, en particulier avec l'émergence des algorithmes d'intelligence artificielle. Malheureusement, ce contenu manque de surveillance et est rarement égalé en sophistication. Il est extrêmement difficile, en particulier pour les non-arabophones, de distinguer et de vérifier l'authenticité des versets coraniques présentés dans des textes arabes non coraniques. Les techniques de traitement de texte classées en dehors du domaine du traitement du langage naturel (NLP) donnent des résultats moins efficaces, en particulier avec les textes arabes.

Pour aborder les problèmes mentionnés, cette thèse présente différentes méthodes pour authentifier le contenu numérique sensible dans le but d'améliorer les étapes de recherche et de récupération. La première méthode repose sur l'identification des versets coraniques intégrés ou référencés dans les textes arabes, qui sont difficiles à distinguer du langage arabe non coranique. L'objectif est atteint en utilisant le Word Embeddings (WE) avec des techniques de Deep Learning (DL). Le travail proposé a été évalué en utilisant douze modèles différents d'incorporation de mots avec deux des classificateurs binaires courants, à savoir : le réseau neuronal convolutif (CNN) et la mémoire à long terme (LSTM). Les résultats expérimentaux ont montré que l'approche proposée surpasse les méthodes traditionnelles dans la distinction entre les versets coraniques et le texte arabe avec une précision de 98,33 %.

Alors que les méthodes traditionnelles d'authentification se sont principalement concentrées sur les mots ou les phrases individuelles, l'authentification des versets coraniques et la préservation de leur ordre en tenant compte de la corrélation de la séquence des mots/versets représentent des aspects pivotaux qui nécessitent une attention spécialisée. Par conséquent, la deuxième méthode repose sur l'authentification des versets/citations spécifiques de la méthode précédente en termes de séquence de mots/versets. L'objectif est atteint en appliquant des algorithmes d'apprentissage profond pour vérifier automatiquement l'intégrité de l'ordre du contenu coranique. L'algorithme LSTM a été choisi pour ce travail, en raison de sa distinction par rapport aux autres réseaux dans le traitement des données séquentielles. Les résultats ont montré une précision de test de 99,98 % sur l'ensemble de données que nous avons créé en utilisant les données du site Tanzil.

Mots-clés:

Contenu en ligne sensible, Authentification du Coran, Identification du Coran, Intégrité du contenu, Deep Learning, Word Embeddings.

Abstract:

The field of analyzing sensitive online content is rapidly evolving, representing a vast area of research that demands considerable attention. With the proliferation of digital content across various online platforms, issues of authenticity and integrity have become increasingly prominent. These factors underscore the critical need for comprehensive research efforts aimed at evaluating and improving the performance and efficiency of methods used to assess digital content, particularly in the form of digital text. Amidst the abundance of digital content available on the internet, the case of the Holy Quran serves as a compelling focal point for research. Despite the dominance of algorithms in all aspects of digital content, it is ironic that Arabic Quranic content remains largely underutilized in terms of computational linguistics, especially with the emergence of artificial intelligence algorithms. Unfortunately, this content lacks surveillance and is rarely matched in sophistication. It is extremely challenging, especially for non-Arabic speakers, to distinguish and verify the authenticity of Quranic verses presented in non-Quranic Arabic texts. Text processing techniques classified outside the realm of natural language processing (NLP) yield less effective results, particularly with Arabic texts.

To address the mentioned issues, this thesis presents various methods for authenticating sensitive digital content with the aim of improving the search and retrieval stages. The first method relies on identifying embedded or referenced Quranic verses in Arabic texts, which are difficult to distinguish from non-Quranic Arabic language. The objective is achieved using Word Embeddings (WE) with Deep Learning (DL) techniques. The proposed work was evaluated using twelve different word embedding models with two common binary classifiers, namely Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). Experimental results showed that the proposed approach outperforms traditional methods in distinguishing between Quranic verses and Arabic text with an accuracy of 98.33%.

While traditional authentication methods have primarily focused on individual words or phrases, authenticating Quranic verses and preserving their order while considering the correlation of word/verse sequences are pivotal aspects that require specialized attention. Therefore, the second method focuses on authenticating specific verses/quotations from the previous method in terms of word/verse sequence. The objective is achieved by applying deep learning algorithms to automatically verify the integrity of the Quranic content order. The LSTM algorithm was chosen for this work due to its distinction from other networks in handling sequential data. The results showed a test accuracy of 99.98% on the dataset created using Tanzil website data.

Keywords:

Sensitive Online Content, Quran Authentication, Quran Identication, Content Integrity, Deep learning, Word Embeddings.

ملخص:

يتطور مجال تحليل المحتوى الحساس عبر الإنترنت بسرعة، ويمثل مجالًا كبيرًا من الأبحاث التي تتطلب اهتمامًا كبيرًا. مع انتشار المحتوى الرقمي عبر مختلف المنصات عبر الإنترنت، أصبحت أسئلة الأصالة والنزاهة بارزة بشكل متزايد. تسلط هذه العوامل الضوء على الحاجة الماسة لجهود بحثية شاملة تهدف إلى تقييم وتحسين أداء وفعالية الأساليب المستخدمة لتقييم المحتوى الرقمي، وخاصة في شكل نص رقمي. وسط وفرة المحتوى الرقمي المتاح على شبكة الإنترنت، فإن حالة القرآن الكريم توفر نقطة محورية مقنعة للبحث. وفي حين تهيمن الخوارزميات على جميع جوانب المحتوى الرقمي، فمن المثير للسخرية أن المحتوى القرآني العربي لا يزال غير مستغل إلى حد كبير من حيث اللغويات الحاسوبية، خاصة مع ظهور خوارزميات الذكاء الاصطناعي. لسوء الحظ، يفتقر هذا المحتوى إلى الرقابة ونادرًا ما يكون مواكبًا للتطور. من الصعب للغاية، وخاصة بالنسبة لغير الناطقين باللغة العربية، التمييز والتحقق من صحة الآيات القرآنية الواردة في النصوص العربية غير القرآنية. إن تقنيات معالجة النصوص المصنفة خارج مجال معالجة اللغة الطبيعية (NLP) تعطي نتائج أقل فعالية، خاصة مع النصوص العربية.

ولمعالجة القضايا المذكورة، تقدم هذه الأطروحة طرقًا مختلفة لتوثيق المحتوى الرقمي الحساس بهدف تحسين مراحل البحث والاسترجاع. تعتمد الطريقة الأولى على تحديد الآيات القرآنية المضمنة أو المشار إليها في النصوص العربية، والتي يصعب تمييزها عن اللغة العربية غير القرآنية. يتم تحقيق الهدف باستخدام تضمين الكلمات (WE) مع تقنيات التعلم العميق (DL). تم تقييم العمل المقترح باستخدام اثني عشر نموذجًا مختلفًا لتضمين الكلمات مع اثنين من المصنفات الثنائية الشائعة، وهما: الشبكة العصبية التلافيفية (CNN) والذاكرة الطويلة قصيرة المدى (LSTM). أظهرت النتائج التجريبية أن المنهج المقترح يتفوق على الطرق التقليدية في التمييز بين الآيات القرآنية والنص العربي بدقة بلغت 98.33%.

وفي حين ركزت طرق التوثيق التقليدية بشكل رئيسي على الكلمات أو العبارات الفردية، فإن توثيق الآيات القرآنية والحفاظ على ترتيبها مع مراعاة ترابط تسلسل الكلمات/الآيات يمثل جوانب محورية تتطلب اهتماما متخصصا. ولذلك فإن الطريقة الثانية تعتمد على توثيق آيات/ اقتباسات محددة من الطريقة السابقة من حيث تسلسل الكلمات/ الآيات. يتم تحقيق الهدف من خلال تطبيق خوارزميات التعلم العميق للتحقق تلقائيًا من سلامة ترتيب المحتوى القرآني. تم اختيار خوارزمية LSTM لهذا العمل، نظراً لتميزها عن الشبكات الأخرى في معالجة البيانات المتسلسلة. أظهرت النتائج دقة اختبار بنسبة 99.98% على مجموعة البيانات التي أنشأناها باستخدام بيانات من موقع Tanzil.

الكلمات المفتاحية:

المحتوى الحساس عبر الإنترنت، توثيق القرآن، تحديد القرآن، سلامة المحتوى، التعلم العميق، تضمين الكلمات.

Table de matières

CHAPITRE 2

Représentation du contenu numérique du Saint Coran dans le domaine de NLP

CHAPITRE 3

Identification du contenu coranique arabe à l'aide de Deep Learning et Word Embeddings

CHAPITRE 4

Authentification des séquences de contenu coranique à l'aide de Deep Learning

Liste des figures

Liste des Tableaux

Liste des Abréviations

AAC	Advanced Audio Coding
AES	Advanced Encryption Standard
AIFF	Audio Interchange File Format
ALC	Arabic Learner Corpus
ARAVEC	Arabic Word Embeddings
ASCII	American Standard Code for Information Interchange
AVI	Audio Video Interleave
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DL	Deep Learning
DOC2VEC	Document to Vector
FN	False Negative
FNNs	Feedforward Neural Networks
FP	False Positive
HMAC	Hash-Based Message Authentication Code
HTML	Hypertext Markup Language
LSTM	Long Short-Term Memory
MD	Message Digest
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
PDF	Portable Document Format
RELU	Rectified Linear Unit
RNN	Recurrent Neural Network
SG	Skip-Gram
SGD	Stochastic Gradient Descent
SHA	Secure Hash Algorithm
SQL	Structured Query Language
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UTF16	Unicode Transformation Format 16-bit
UTF8	Unicode Transformation Format 8-bit
WAV	Waveform Audio File Format
WE	Word Embeddings
WEBM	WebM Video File Format
WORD2VEC	Word to Vector
WOS	Web of Science
XML	Extensible Markup Language

Introduction générale

INTRODUCTION GENERALE

Cette introduction présente le contexte de cette étude, met en évidence le problème de recherche et la motivation de la réalisation de cette recherche, et formule l'énoncé du problème ainsi que les questions de recherche. Elle inclut également la portée et la signification de la recherche, ainsi que la méthodologie de recherche.

1. Introduction

L'authenticité/l'authentification constitue un ensemble de politiques et de procédures nécessaires pour valider l'entrée fournie [1]. Deux principales procédures pour authentifier le contenu numérique comprennent la phase de recherche/récupération et la vérité de référence vérifiée (base de données). Sans une stratégie appropriée de recherche/récupération, le processus d'authentification prendra plus de temps pour traiter l'entrée fournie (qui peut être n'importe quel support numérique) [2]. Actuellement, la problématique de l'authentification est en hausse en raison de l'avènement des gadgets modernes. Au cours des dernières années, on observe une augmentation constante de l'utilisation des médias numériques mis en ligne et téléchargés sur Internet. Cette augmentation constante des médias numériques sur Internet constitue l'un des principaux défis auxquels les chercheurs sont confrontés aujourd'hui en termes de détermination de l'authenticité. D'autres problèmes surviennent en raison de la disponibilité des médias numériques sous différents formats tels que l'image, le texte, l'audio et la vidéo. Cette disponibilité des médias numériques sous différents formats rend l'identification des approches appropriées pour authentifier un format particulier plus complexe.

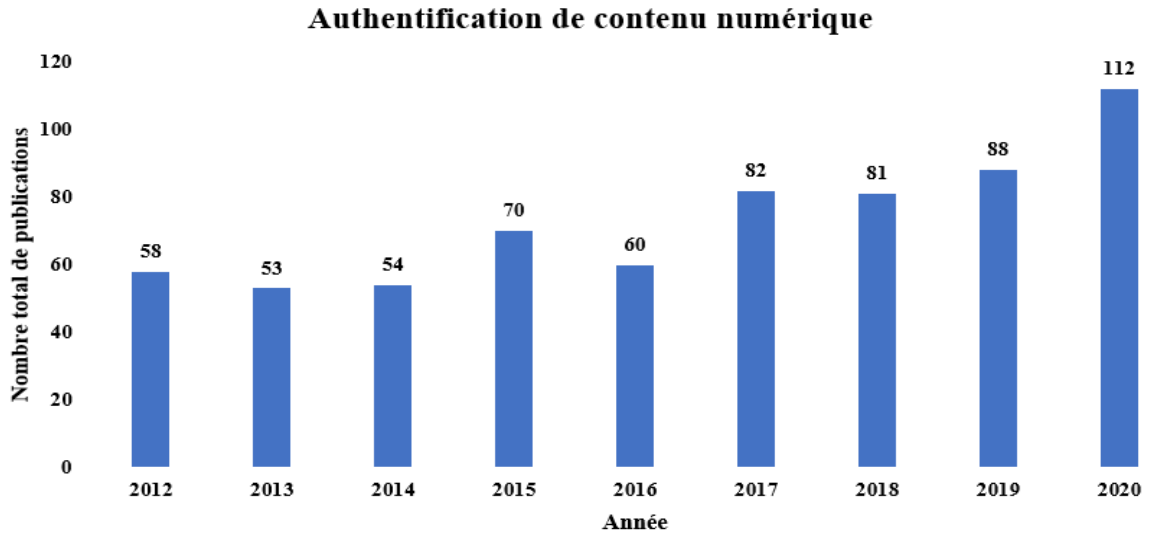


Figure 0. 1:Recherche dans le domaine de l'authenticité/intégrité des données sensibles en ligne. (WOS 2020)

De plus, l'utilisation du contenu numérique a considérablement augmenté les cas de violations de droits d'auteur, incitant les chercheurs à étudier les problèmes liés à l'intégrité, à l'authenticité du contenu numérique et à la vulnérabilité des données [3] [4].C'est la raison pour laquelle une quantité substantielle de recherche est en cours dans

le domaine de l'intégrité des données, de l'authentification et de la sécurité. Les données liées à la recherche en matière d'authentification du contenu numérique sont présentées dans la Figure 0.1.

La dépendance excessive à l'internet et l'augmentation du nombre d'utilisateurs ont exacerbé le problème d'intégrité et d'authenticité. Selon les informations fournies par les statistiques mondiales sur l'internet (Internet World Stats[1], 2020), le nombre d'utilisateurs d'internet dans le monde ne cesse d'augmenter à un rythme remarquable (comme le montre la Figure 0.2).

Face à cette tendance inquiétante, le taux de publication de contenus numériques sensibles en ligne est nécessairement également en hausse. Une grande quantité de contenu numérique de Saint Coran est disponible en ligne et peut être accessible et publier à partir de différentes sources, telles que des sites web religieux, des sites de réseaux sociaux et d'autres blogs en ligne.

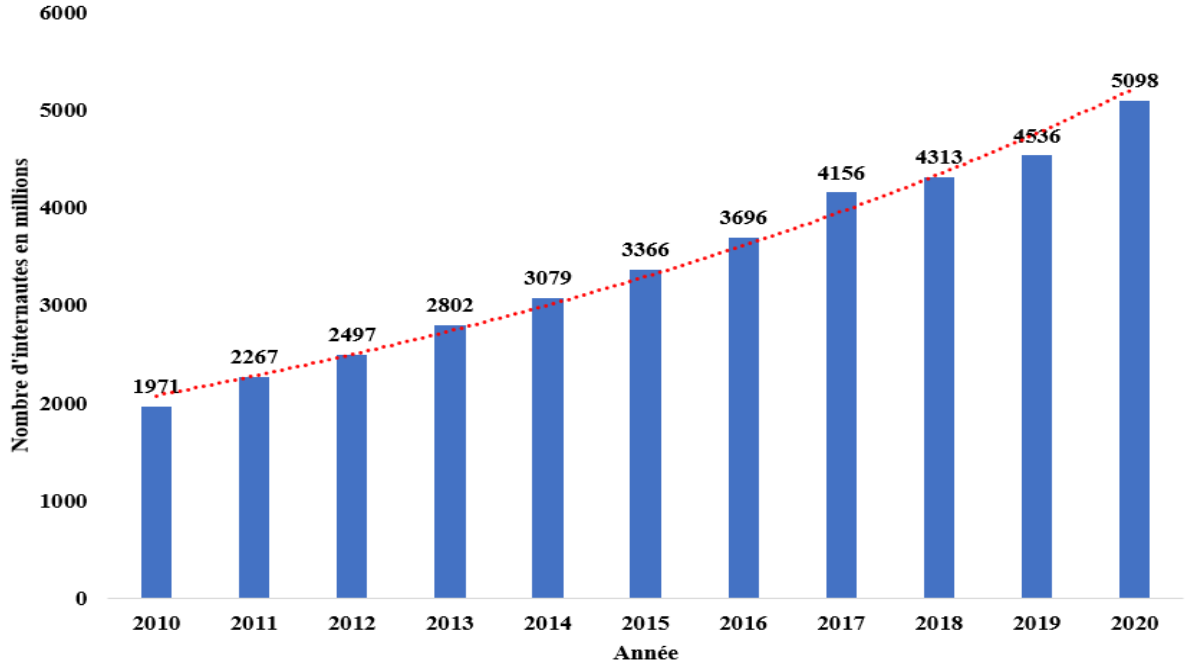


Figure 0. 2:Nombre d'utilisateurs d'Internet par an (Internet World Stats,, 2020)

2. Motivation

Le domaine de l'analyse du contenu en ligne sensible évolue rapidement, présentant un champ de recherche en pleine expansion qui demande une attention considérable. Alors que le contenu numérique continue de proliférer sur diverses plateformes en ligne, les questions relatives au droit d'auteur, à l'authenticité et à l'intégrité sont devenues de plus en plus prévalentes. Ces facteurs soulignent le besoin critique d'efforts de recherche complets visant à évaluer et à améliorer les performances et l'efficacité des méthodes utilisées pour évaluer le contenu numérique, en particulier sous forme de texte

[1] Internet World Stats est un site Web international qui fournit des statistiques et des informations sur l'utilisation d'Internet dans le monde entier : https://www.internetworldstats.com/

numérique.

Au milieu de l'immensité du contenu numérique disponible sur Internet, le cas du Saint Coran constitue un point focal convaincant pour la recherche. Le Coran, représenté numériquement dans divers styles de script, y compris les styles bien établis d'Othmani et de Coran simple (le plus couramment utilisé), présente des défis et des opportunités uniques pour la recherche. Alors que les méthodes d'authentification traditionnelles se sont principalement concentrées sur les mots ou les versets individuels, l'authentification des versets coraniques et la préservation de leur ordre, en considérant la corrélation de la séquence des mots/versets, représentent des aspects pivotaux qui nécessitent une attention spécialisée.

Particulièrement remarquable est la difficulté à discerner les fragments de versets intégrés dans des publications textuelles étendues. Par exemple, considérons l'exemple fourni dans la figure 0.3 Identifier le verset intégré (surligné en jaune) au sein de cette publication longue s'avère être une tâche redoutable en raison de la complexité et de la verbosité du texte.

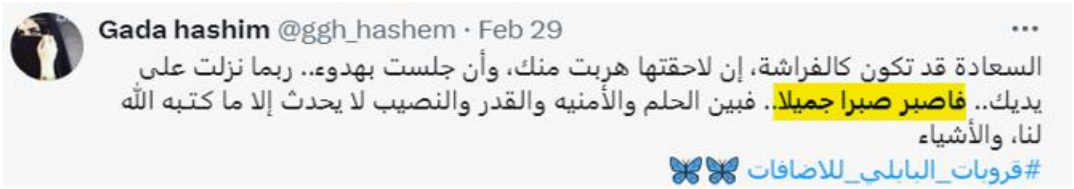


Figure 0. 3:Exemple d'un verset intégré dans une publication

De plus, le défi s'étend à garantir le bon ordre des mots ou des versets, un aspect crucial pour maintenir l'intégrité et l'authenticité des textes coraniques. Considérons un autre scénario où un verset authentique avec le bon ordre est présenté dans la figure 0.4.a. Cependant, dans la figure 0.4.b, le même verset a été altéré en modifiant la séquence de deux mots, entraînant un changement profond de sens.

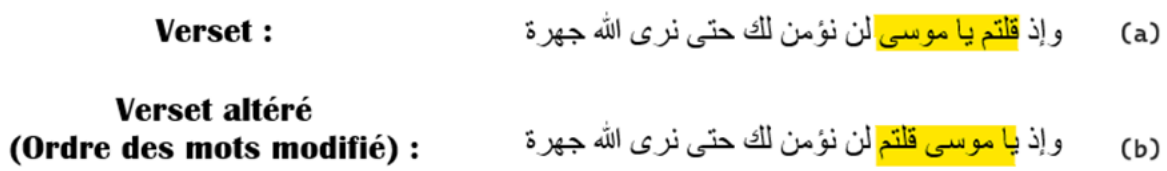


Figure 0. 4: Altération d'un verset par réorganisation des mots ;
(a) Verset authentique avec le bon ordre, (b) Verset altéré par réorganisation des mots

Les exemples fournis ci-dessus soulignent l'importance critique d'identifier et d'authentifier avec précision l'ordre des versets coraniques au sein du contenu en ligne en arabe. Cependant, notre motivation réside principalement dans l'identification et l'authentification de l'ordre des versets au sein des textes coraniques, plutôt que de se concentrer sur les diacritiques. Ainsi, bien que des études antérieures [5] aient mis l'accent sur l'impact des diacritiques sur la récupération, notre recherche est motivée par la nécessité de relever les

défis associés à l'identification et à l'authentification de l'ordre des versets dans le contenu en ligne en arabe.

Ce faisant, nous visons à garantir l'intégrité et l'authenticité du contenu textuel sensible, en maintenant ainsi sa signification et sa fiabilité dans les contextes numériques.

3. Problématique

L'intégration de textes sensibles, tels que le Coran, au sein de contenus en ligne en arabe ou de statuts sur les réseaux sociaux, présente des défis significatifs en termes d'authentification, notamment en ce qui concerne l'ordre des versets. Alors que les méthodes traditionnelles d'authentification des versets individuels peuvent suffire dans certains contextes, l'authentification de plusieurs versets séquencés pose des complexités uniques.

La complexité de la représentation d'une seule lettre ou d'un seul caractère dans le texte arabe, notamment en tenant compte des structures linguistiques complexes, pose des défis inhérents en raison de la nature de la langue. Chaque caractère arabe nécessite un nombre important de bits pour sa représentation, dépassant souvent les schémas de codage standard. Cela se traduit par une complexité computationnelle accrue et des besoins de stockage plus importants, en particulier dans les systèmes traitant de grands volumes de texte.

De plus, les approches d'authentification existantes se concentrent principalement sur l'authentification au niveau des mots plutôt que sur l'identification et l'authentification au niveau des phrases ou des versets. Alors que l'authentification au niveau des mots fournit des informations précieuses sur l'intégrité des mots individuels, elle néglige la nature cohésive du texte arabe, en particulier dans les écritures religieuses comme le Coran. Les versets du Coran ne sont pas simplement une collection de mots, mais plutôt des unités cohésives avec des relations linguistiques et sémantiques complexes.

En conséquence, les méthodes d'authentification actuelles peuvent ne pas réussir à capturer les variations nuancées et les subtilités présentes dans les versets coraniques. Cette limitation compromet l'efficacité des processus d'authentification, car elle peut négliger les altérations ou les divergences potentielles dans l'ordre ou l'arrangement des versets, qui sont cruciales pour préserver l'intégrité du texte.

Pour relever ces défis, une mutation des méthodologies d'authentification est nécessaire, mettant l'accent sur l'importance de l'identification et de l'authentification au niveau des versets. En adoptant des approches qui tiennent compte de la structure holistique et de la sémantique des versets coraniques, il est possible de développer des méthodes d'authentification plus robustes et précises. Ces méthodes devraient comprendre des techniques d'analyse linguistique sophistiquées, notamment le traitement du langage naturel et la modélisation sémantique, pour garantir l'identification et l'authentification précises des versets coraniques dans le texte arabe en ligne.

En résumé, la complexité de la représentation des caractères arabes individuels, associée à la focalisation prédominante sur l'authentification au niveau des mots, souligne la nécessité d'approches innovantes qui privilégient l'identification et l'authentification au niveau des phrases ou des versets. En relevant ces défis, il devient possible d'améliorer l'authenticité/l'intégrité du texte coranique dans le contenu en ligne en arabe, préservant ainsi sa signification et sa fiabilité dans les contextes numériques.

4. Questions de recherche

Cette étude aborde les questions de recherche suivantes qui ont été formulées en correspondance avec les objectifs de recherche afin de délimiter le champ de cette recherche :

a) Quels sont les défis spécifiques rencontrés dans l'identification des versets coraniques intégrés ou cités dans des textes arabes non coraniques sur Internet, en tenant compte de la diversité des styles de citations des mots / part of verse / verse or multiple versets ?

b) Quelles sont les techniques les plus appropriées pour représenter efficacement les textes coraniques en prenant en considération les particularités linguistiques et structurelles du Coran ?

c) Quelle est la meilleure technique peuvent-elles être adaptées et appliquées de manière optimale pour identifier et authentifier l'ordre des séquences de mots/versets dans les textes coraniques numériques, en tenant compte de la complexité et de la diversité linguistique des versets ?

5. Objectifs de recherche

L'objectif de cette thèse est de déterminer l'intégrité des textes coraniques numériques en se concentrant sur l'identification et l'authentification de l'ordre des séquences de mots/versets. Pour atteindre cet objectif, nous avons formulé les objectifs suivants :

- Analyser en profondeur les défis spécifiques liés à l'authentification de l'intégrité des textes coraniques numériques, en mettant en évidence les aspects linguistiques et techniques impliqués dans l'identification des versets coraniques dans les textes non coraniques ;

- Examiner une représentation efficace des données du contenu coranique. L'objectif est de disposer d'une représentation des données qui capture fidèlement les spécificités du texte coranique et qui puisse être utilisée efficacement dans les étapes ultérieures de vérification et d'authentification.

- Développer une approche avancée pour l'identification du contenu coranique dans des textes arabes numériques. Le but de cet objectif est de concevoir une méthode efficace capable d'identifier avec précision les versets coraniques, y compris ceux intégrés dans des textes arabes simples (sans diacritiques).

- Développer une méthode d'authentification du texte coranique. L'objectif est de fournir une solution robuste et fiable permettant de vérifier l'intégrité du texte coranique, en prenant en compte non seulement l'identification des mots/versets individuels, mais également leur ordre de présentation.

6. Signification de l'Intégrité et de l'Authentification dans le Multimédia Coranique

Assurer l'intégrité et l'authentification du texte coranique numérique est primordial pour les musulmans et l'humanité dans son ensemble. Il est impératif d'utiliser les technologies de l'information de manière responsable, surtout lorsqu'il s'agit de servir le Saint Coran, afin de prévenir toute forme de manipulation ou de diffusion de fausses versions du texte sacré.

Allah a assuré dans le Coran :

﴿ إِنَّا نَحْنُ نَزَّلْنَا ٱلذِّكْرَ وَإِنَّا لَهُۥ لَحَٰفِظُونَ ۝٩ ﴾
[سُورَةُ الحِجۡرِ:٩]

En vérité, c'est Nous qui avons révélé le Dhikr (c'est-à-dire le Coran) et c'est Nous qui en assurerons la préservation (contre la corruption).
Traduction du Complexe d'Imprimerie du Coran Glorieux du Roi Fahd

Par conséquent, l'objectif principal de cette recherche est de répondre au besoin critique d'assurer l'authentification et l'intégrité des textes coraniques numériques. Cela implique de sensibiliser à l'importance de protéger le contenu original du Coran et de mettre en œuvre des mesures de protection robustes pour empêcher les modifications ou les altérations non autorisées.

En mettant l'accent sur l'importance de maintenir l'authenticité du Coran, ce travail vise à empêcher toute altération délibérée ou involontaire du texte de passer inaperçue. De plus, il cherche à établir des mécanismes de détection de telles modifications et à mettre en œuvre des actions correctives promptement.

En essence, l'objectif est d'utiliser les technologies de l'information modernes pour préserver la sainteté et l'exactitude du Coran, garantissant ainsi qu'il reste préservé et protégé contre toute forme de corruption ou de falsification.

6.1. Exemples de problèmes d'intégrité et d'authentification associés au contenu multimédia numérique du Coran

Les figures suivantes mettent en lumière quelques statues sur la plateforme X, illustrant la citation et la diffusion de versets altérés/erronés. Le partage sur ces plateformes est le moyen le plus répandu de diffusion, en raison de leur facilité d'utilisation.

La Figure 0.5 illustre la citation du verset 85 de la Sourate Al-Imran, où le mot "الإسلام" a été remplacé par "السلام", entraînant un changement de sens du verset. De même, la Figure 0.6 montre le verset 8 de la Sourate An-Nahl, où l'ajout du mot "للحياة" a conduit à une altération et un changement de sens du verset.

En outre, la Figure 0.7 montre le verset 7 de la Sourate Al-Isra, où le simple remplacement d'une lettre (le mot "مرة" par "مره") a entraîné une distorsion du sens du verset.

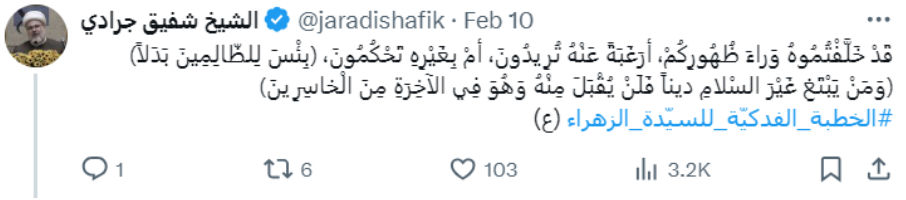


Figure 0. 5:Exemple de problème d'intégrité avec remplacement d'un mot

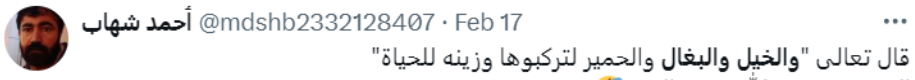


Figure 0. 6:Exemple de problème d'intégrité avec l'ajout d'un mot

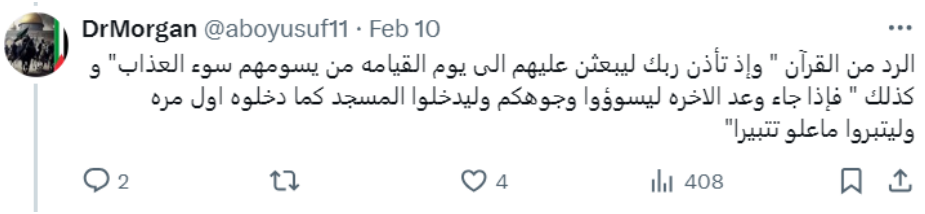


Figure 0. 7:Exemple de problème d'intégrité avec remplacement d'une lettre

Une recherche sur internet concernant les applications informatiques au service du Saint Coran a révélé des résultats inattendus. Il est particulièrement remarquable de constater que plusieurs applications diffusent des versets inexacts ou omis du Saint Coran. À titre d'exemple, la Figure 0.8 illustre une application mobile affichant un verset erroné dans le chapitre Al-Baqara. Même lorsque le texte est présenté sous forme d'image, il reste vulnérable aux risques de réécriture ou de copie.

Figure 0. 8:Exemple d'un verset coranique falsifié affiché sur une application mobile

La Figure 0.9, en revanche, mentionne la sourate Al-Baqara, verset 255, qui est Ayat al-Kursi, a été mentionnée sans l'inclusion de la verse "لا إِكْرَاهَ فِي الدِّينِ" qui est le numéro 206, mais immédiatement suivie de la verse "اللَّهُ وَلِيُّ الَّذِينَ آمَنُو" qui est le numéro 207.

Figure 0. 9: Exemple d'une sourate coranique falsifié affiché sur une application mobile

Le Figure 0.10 illustre également un exemple d'un site de Tafsir du Coran, où des versets coraniques erronés sont cités. Plus précisément, la Figure montre le verset 66 de la Sourate Al-Anfal , où certains mots ont été supprimés et d'autres ont été manipulés dans leur ordre.

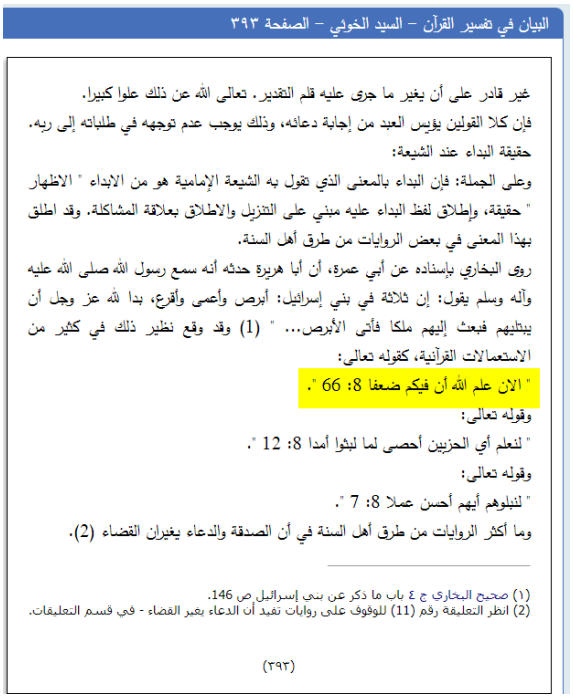


Figure 0. 10: Exemple d'un verset coranique falsifié affiché sur un site web

La Figure 0.11 illustre un autre exemple de script fabriqué, spécifiquement le livre du Vrai Furqan, disponible à l'achat via la boutique en ligne d'Amazon. Ce livre, connu pour son contenu altéré et sa présentation erronée du Coran, soulève des préoccupations quant à la diffusion de fausses interprétations de textes sacrés sur des plateformes commerciales populaires telles qu'Amazon.

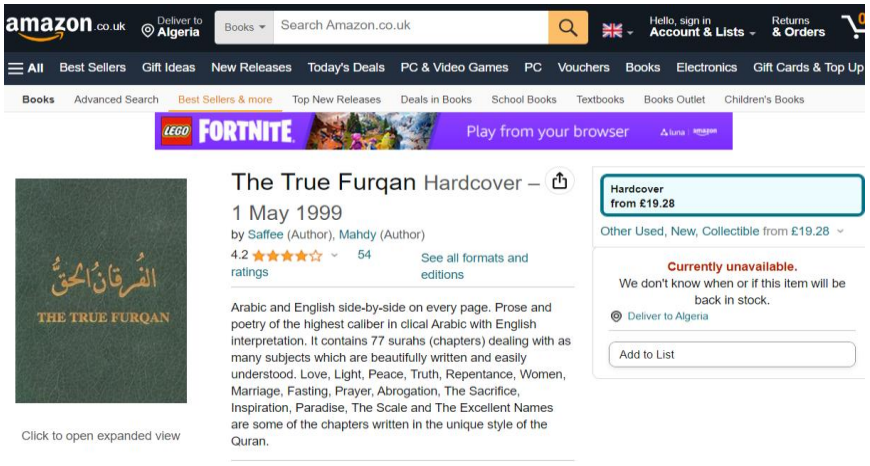


Figure 0. 11:Exemple d'un Coran falsifié sous forme de livre

Par conséquent, il est impératif pour les musulmans de donner la priorité à la mise en œuvre de mesures visant à identifier et à corriger toute présentation trompeuse du Saint Coran, que ces altérations soient intentionnelles ou non. Cela est crucial pour garantir que les individus accédant au contenu coranique en ligne ne reçoivent que du matériel précis et authentique, sans compromis.

7. Méthodologie de recherche

Le projet de recherche est divisé en deux phases d'étude : la phase 1 consiste en une revue de la littérature, tandis que la phase 2 porte sur l'authentification (comprenant l'identification et la vérification d'authenticité).

La première phase de la recherche couvre la revue de la littérature, qui présente les concepts de base et les problématiques liés au champ d'étude couvrant le premier objectif. Les travaux connexes sur la préservation de l'intégrité du contenu de textes sensibles, avec l'étude de cas du saint Coran numérique, sont étudiés, et de nouvelles taxonomies pouvant ouvrir la voie à de futures directions de recherche sont proposées. Pour relever le deuxième objectif, une méthode basée sur lle DL est proposée. Pour la phase d'identification impliquant la récupération de versets coraniques, un jeu de données est constitué et une nouvelle représentation des données est proposée. Pour l'authentification de l'ordre de séquence des versets/sourates, une méthode basée sur le DL est proposée afin d'améliorer la précision et la robustesse du traitement du texte.

La représentation des données du Coran est abordée au Chapitre 2, et la méthodologie de recherche individuelle des méthodes mentionnées ci-dessus est décrite et présentée séparément. Les Chapitres 3 et 4 sont organisés en introduction, travaux connexes, méthodologie de recherche, résultats et conclusion. Le flux de recherche et les méthodes de recherche correspondantes sont résumés dans la Figure 0.12.

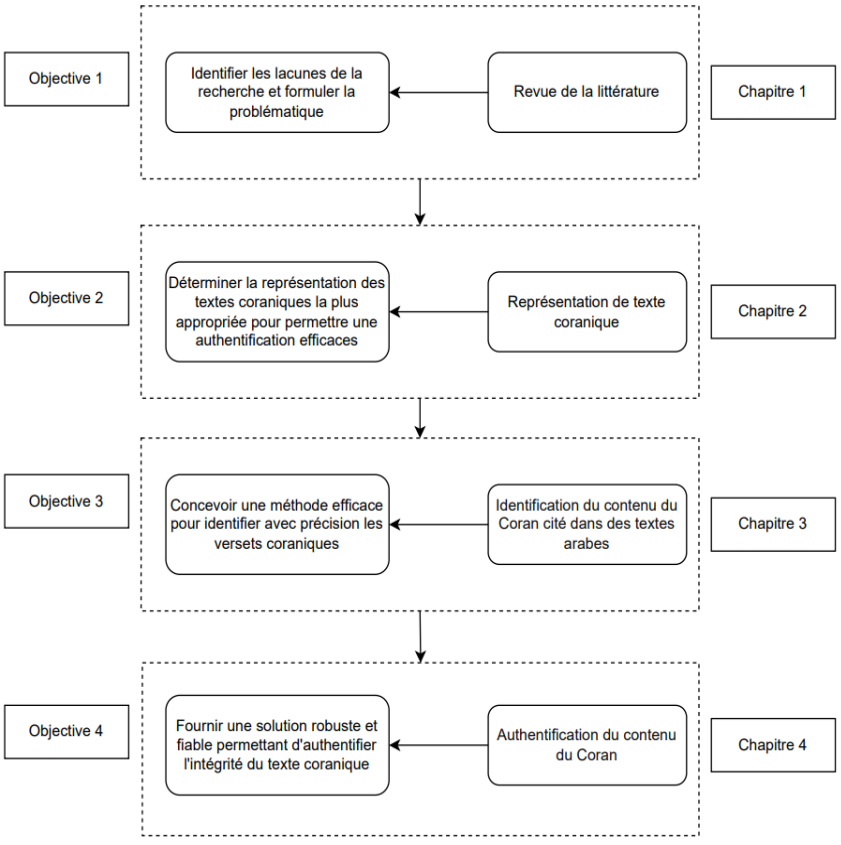


Figure 0. 12:Flux de recherche proposées.

8. Contributions

Les contributions de cette recherche comprennent :

- L'identification d'approches pouvant être utilisées pour préserver l'intégrité du contenu du texte coranique à travers des taxonomies ;
- La constitution d'un ensemble de données étiqueté comprenant à la fois des versets coraniques et des textes arabes simples ;
- Une méthode de représentation efficace des données pour authentifier les versets cités dans des textes arabes ;
- Des modèles pour identifier et authentifier des textes arabes coraniques ;
- Des recommandations pour les futures recherches dans le domaine du texte coranique.

La faisabilité des approches techniques proposées a été démontrée en prenant le Coran numérique comme étude de cas. Cette étude constitue également la première contribution réalisée sur les différentes approches utilisant des modèles d'intelligence artificielle appliqués pour authentifier l'intégrité du Coran numérique.

9. Organisation de la thèse

Notre thèse est structurée en quatre chapitres organisés comme suit :

- L'introduction générale présente le contexte de travail, les motivations de la recherche, la problématique, une esquisse générale de la méthodologie utilisée, et les objectifs souhaités.
- Le premier chapitre introduit une revue approfondie de la littérature. Il passe en revue les différentes approches existantes pour l'authentification du contenu coranique, leurs forces et leurs limites. Une taxonomie des techniques d'authentification est également élaborée.
- Le second chapitre est consacré à la représentation du contenu coranique en NLP. Il étudie en détail la structure et les caractéristiques statistiques du texte coranique. Ensuite, il compare et évalue différentes méthodes de représentation des données, comme les techniques de Word Embeddings , afin d'identifier la meilleure approche pour capturer fidèlement les spécificités du contenu coranique.
- Le troisième chapitre couvre notre première contribution, qui concerne l'identification du contenu coranique cité dans des textes arabes. Cette approche se base

essentiellement sur des techniques de deep learning et de Word2vec. Les performances de cette méthode sont évaluées de manière approfondie.

- Le quatrième chapitre présente notre deuxième contribution, qui porte sur l'authentification du contenu textuel coranique en se basant sur l'ordre des mots/versets. Des expérimentations sont menées sur des cas d'étude impliquant les versets les plus utilisés en ligne.
- Enfin, la conclusion générale résume l'ensemble de nos travaux et décrit les perspectives envisagées pour de futures recherches.

Chapitre 1

Revue de la littérature

1. Introduction

Le Saint Coran numérique est l'un des domaines de contenu les plus sensibles qui ont été identifiés et authentifiés par les chercheurs dans des travaux antérieurs connexes. Le contenu sur le Coran peut être centré sur des images, du texte et de l'audio/vidéo. Cette étude s'intéresse aux seules approches textuelles ainsi qu'à leurs limites face à l'énorme volume de contenus numériques disponibles sur Internet. De plus, ce chapitre fournit une brève explication de certaines techniques de base utilisées par les chercheurs pour authentifier et protéger les autres format de contenu avec certains avantages et inconvénients associés à chacune d'entre elles. Enfin, des questions de recherche ouvertes ainsi que des conclusions sont tirées.

2. Le contenue sensible

Comme le stipule la Loi sur la sécurité informatique de 1987, le contenu en ligne sensible peut être définie comme celui dont l'exactitude, la sécurité et la véracité sont d'une importance cruciale et critique pour les utilisateurs [6].

2.1. Intégrité de contenue sensible

‘‘ *Intégrité : Propriété d'exactitude et de complétude d'une information.*’’ (Source : ISO/CEI 27000:2018[2])

‘‘ *Intégrité : Propriété assurant qu'une information n'a pas été modifiée ou détruite de façon non autorisée.*’’ (Source : Instruction Générale Interministérielle n°1300[3])

‘‘*Intégrité : Qualité d'un document ou d'une donnée qui n'a pas été altéré. Dans le monde numérique, un document ou une donnée est réputé intègre si son empreinte à un temps t+1 est identique à l'empreinte prise à un temps t.*’’ (Source : Référentiel General de Gestion des Archives[4])

Dans le cadre de ce travaille , l'intégrité de contenue sensible est une propriété assurant que le contenu sensible , n'a pas été altéré, modifié, ou détruit de manière non autorisée depuis sa révélation originale, garantissant ainsi son exactitude, sa complétude et son authenticité.

3. Classification de contenue sensible

[2] ISO/CEI 27000:2018 - Technologies de l'information - Techniques de sécurité - Systèmes de management de la sécurité de l'information - Vue d'ensemble et vocabulaire.

[3] Instruction générale interministérielle n° 1300 sur la protection du secret de la défense nationale, approuvée par arrêté du Premier Ministre du 9 août 2021

[4] Référentiel General de Gestion des Archives R2GA - octobre 2013 : Référentiel général de gestion des archives (R2GA) (FranceArchives)

Le contenu sensible en ligne peut être classé en quatre types différents en fonction de leur format respectif (comme illustré dans la Figure 1.1). Une telle classification permet à la recherche d'identifier et d'analyser différentes techniques pour protéger et vérifier l'intégrité du contenu sensible en ligne.

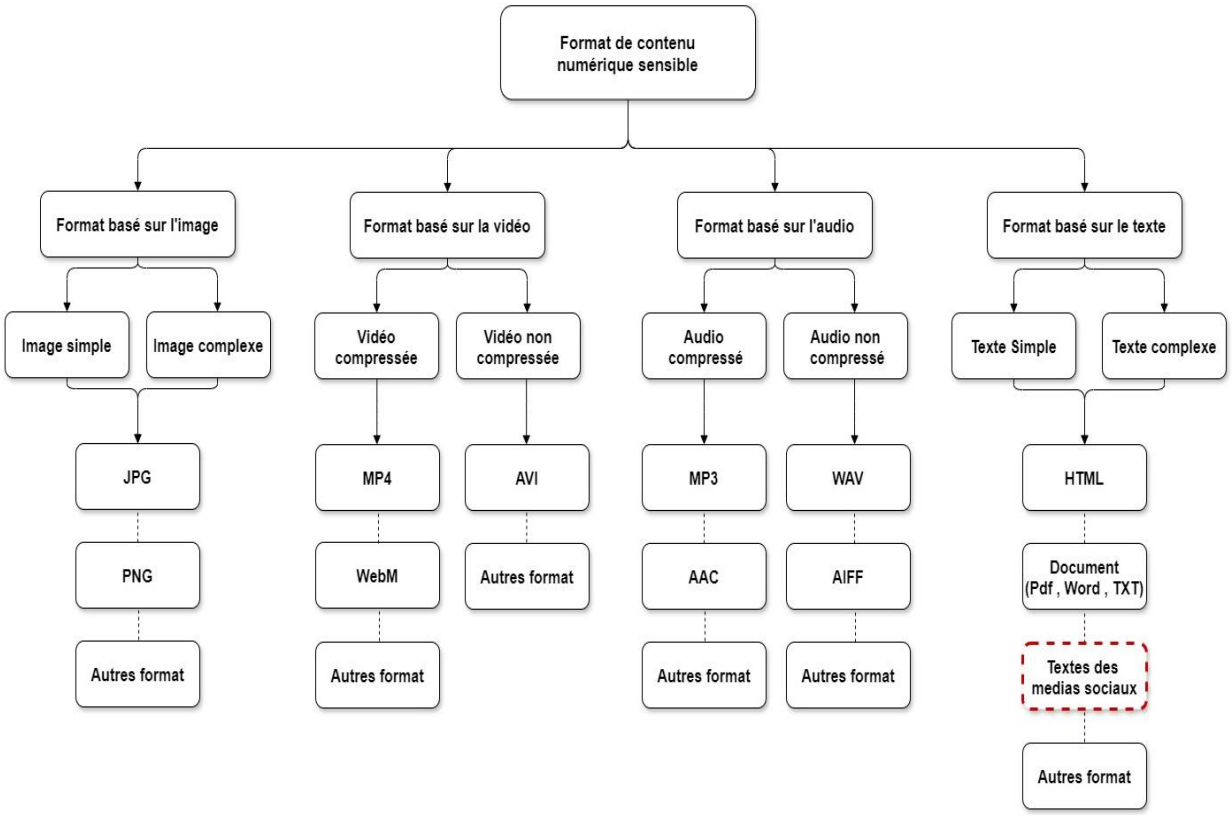


Figure 1. 1:Format de contenue numérique sensible

3.1.Format basé sur l'image

Internet regorge de contenus basés sur des images, facilement accessibles lors de la navigation sur différents sites web. Par exemple, les livres, les captures sur les réseaux sociaux sont souvent présentés sous forme d'images. Nous avons classé les images en deux sous-catégories : les images simples et les images complexes. Les images simples comprennent des photos basiques avec des détails de couleur et de clarté moyens, tandis que les images complexes présentent des détails de couleur plus riches mais avec moins de clarté. Les images simples et complexes peuvent être affichées dans différents formats tels que PNG, JPG et d'autres encore. Ce domaine relève du traitement d'image, et il existe différentes techniques utilisées pour vérifier leur intégrité.

3.2. Format basé sur le fichier audio/vidéo

Une grande quantité de contenu est disponible en ligne sous forme de fichiers audio et vidéo. Pour les fichiers audio, ils peuvent être classés en deux catégories principales : les fichiers audio compressés comme MP3 et AAC, qui offrent une bonne qualité sonore mais avec une compression de données, et les fichiers audio non compressés tels que WAV et AIFF, qui préservent la qualité audio maximale mais nécessitent plus d'espace de stockage. De même,

pour les vidéos, elles peuvent être divisées en deux catégories : les vidéos compressées comme MP4 et WebM, qui offrent un équilibre entre qualité et taille de fichier, et les vidéos non compressées comme AVI, qui préservent la qualité maximale mais nécessitent plus d'espace de stockage. Pour vérifier l'authenticité de ce contenu, différentes techniques peuvent être appliquées dans les deux cas.

3.3. Format basé sur le texte

Le contenu en ligne est également riche en texte, qu'il soit simple ou complexe. Le texte simple est généralement en anglais basique utilisant le format ASCII, tandis que le texte complexe peut inclure diverses langues comme la langue Arabe et caractères spéciaux, utilisant des encodages tels qu'UTF-8 et UTF-16. Les formats pour le texte incluent HTML pour les sites web, formats de document tels que Word, PDF et TXT ainsi que les textes des médias sociaux.

Les textes de media sociaux peuvent prendre de nombreux formats, des messages courts et rapides aux publications plus longues et réfléchies, en passant par les hashtags, les citations et les messages privés. La nature du texte dépend de la plateforme, du contexte et des intentions de l'utilisateur.

Le texte arabe des médias sociaux présente des caractéristiques particulières. Un tel texte peut être un mélange de variétés d'arabe standard moderne, d'arabe classique et également d'arabe dialectal, et en outre, il peut contenir des mots non arabes, des échantillons, des notations et des caractéristiques orthographiques telles que des fautes d'orthographe, des lettres répétées ou exprimer des mots émotifs.

Comparativement à d'autres formats, la recherche sur l'intégrité du contenu textuel a été un peu limitée (comme indiqué dans la figure 1.2) en raison de la complexité des mécanismes d'authentification requis.

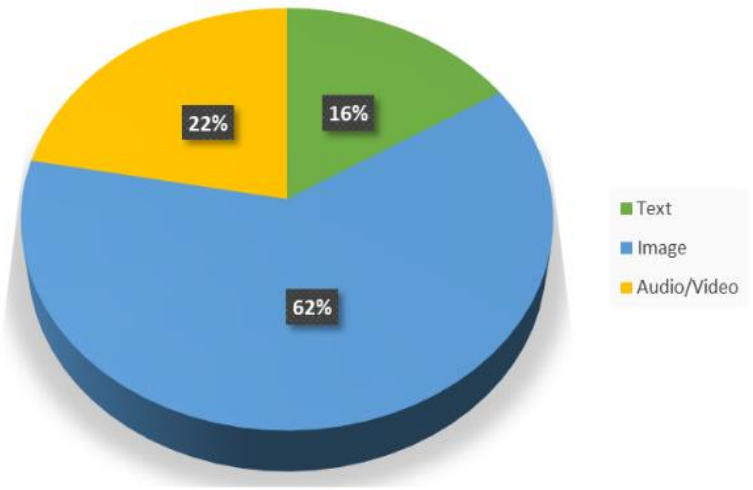


Figure 1. 2: Répartition des recherches sur l'intégrité du contenu des différents formats (WoS , 2020)

Nous avons plongé dans une multitude de formats de ressources en ligne, fournissant aux chercheurs un cadre robuste pour classifier les techniques adaptées à des formats spécifiques.

Cette classification facilite l'identification et l'évaluation des avantages et des limites distincts de chaque format. Dans les formats d'image et d'audio, une variété de méthodologies sont actuellement employées pour garantir l'authenticité et l'intégrité du contenu du Saint Coran numérique. Ces approches incluent des techniques différentes d'authentification, ainsi que des méthodes de protection pour préserver la sacralité de ce texte vénéré dans sa manifestation numérique.

4. Les niveaux d'authentification et de préservation de l'intégrité du contenu coranique

Dans cette section, nous abordons les différents niveaux d'authentification et de préservation de l'intégrité du contenu coranique. Cela inclut l'authentification au niveau du document entier, ainsi que des aspects plus spécifiques tels que la vérification de la validité des versets ou des ayas complets ou incomplets, la préservation des signes diacritiques, et même la conservation de l'ordre des mots ou des versets. Chaque niveau d'authentification représente un défi unique dans la préservation de l'intégrité du contenu coranique, et les différentes approches développées visent à répondre à ces défis de manière efficace et précise.

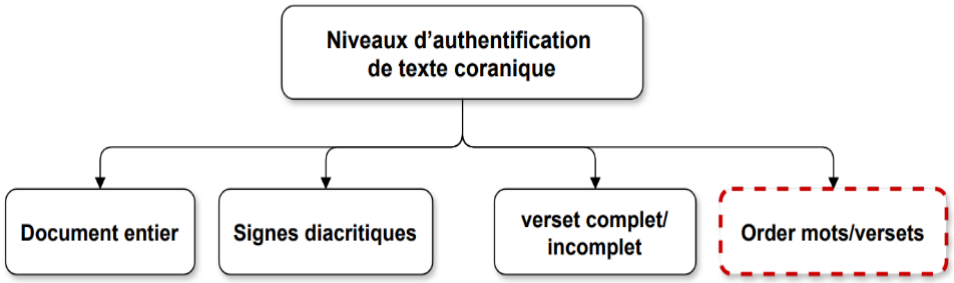


Figure 1. 3: les niveaux d'authentification et de préservation de l'intégrité de contenue coranique

4.1. Authentification au niveau du document entier

L'authentification du document entier du Coran est un processus complexe qui vise à garantir que l'ensemble du texte coranique est préservé dans sa forme originale, sans altération ni modification. Cela implique la vérification de l'authenticité des manuscrits du Coran, en remontant aux premiers manuscrits compilés après la mort du Prophète Muhammad (que la paix soit sur lui). Les érudits coraniques examinent attentivement les chaînes de transmission (isnad) des manuscrits, en vérifiant leur continuité et leur authenticité à travers les générations.

De plus, ils comparent les différentes versions manuscrites du Coran pour détecter toute variation ou divergence. Les normes d'authentification exigent que les manuscrits du Coran soient en accord avec les principes de transmission orale établis et qu'ils correspondent aux manuscrits les plus anciens disponibles. Tout écart par rapport à ces normes pourrait compromettre l'intégrité du texte sacré.

4.2. Vérification de la validité des versets ou des ayas complets ou incomplets

Chaque verset (aya) du Coran est considéré comme une unité de sens complète et sacrée. Par conséquent, il est crucial de vérifier l'authenticité et l'intégrité de chaque aya, qu'elle soit complète ou incomplète. Les érudits coraniques utilisent des méthodes de critique textuelle pour évaluer la validité des versets, en examinant les variantes de texte, les styles de calligraphie et les pratiques de récitation à travers différentes écoles de récitation coranique. Pour les versets incomplets, les érudits se réfèrent également à des sources linguistiques et grammaticales pour déterminer l'intégrité structurelle du texte. Les normes d'authentification exigent que chaque aya soit conforme aux traditions bien établies de récitation et de transmission, garantissant ainsi la fidélité du texte coranique à son message originel.

4.3. Préservation des signes diacritiques

Les signes diacritiques (tels que les points et les accents) jouent un rôle crucial dans la prononciation correcte et la compréhension précise du texte coranique. Par conséquent, il est essentiel de préserver ces signes diacritiques pour maintenir l'intégrité du texte. Les défis techniques résident dans la reproduction précise de ces signes dans les manuscrits et les éditions imprimées du Coran, ainsi que dans les formats numériques. Les érudits coraniques travaillent en étroite collaboration avec des calligraphes et des spécialistes des langues pour garantir que les signes diacritiques sont reproduits avec précision, conformément aux règles de récitation et de lecture coraniques établies. Cela garantit que le texte est transmis avec la même précision phonétique que celle avec laquelle il a été révélé.

4.4. Conservation de l'ordre des mots ou des versets

L'ordre des mots et des versets dans le Coran est considéré comme sacré et fait partie intégrante de sa révélation divine. Tout changement dans cet ordre compromettrait l'intégrité du texte. Les érudits coraniques veillent à ce que l'ordre des mots et des versets soit préservé exactement tel qu'il est dans les manuscrits authentiques du Coran. Cela implique une attention particulière à la typographie, à la mise en page et à la pagination pour garantir que chaque verset est positionné correctement par rapport aux autres. Avec les avancées technologiques, il est également crucial de maintenir l'ordre des versets intact lors de la conversion du texte coranique dans différents formats et supports.

Sur les plateformes en ligne et les médias sociaux, l'authentification de l'intégrité du contenu coranique revêt une importance particulière en raison de plusieurs facteurs. Malgré l'importance des niveaux de documentation précédents et l'attention considérable accordée par la recherche, l'authentification de l'ordre des mots et des versets est d'autant plus cruciale. L'utilisation des signes diacritiques dans ces environnements est limitée et peut être absente. De plus, les citations du Coran sous forme de versets ou de séquences de versets sont privilégiées par rapport aux documents complets (Mushaf), en raison des contraintes de taille des publications sur ces plateformes. En outre, la vérification de la complétude des versets, qu'ils soient complets ou incomplets, n'est pas toujours applicable dans ce contexte, car les

citations sont généralement extraites d'une partie de verset et complétées par trois points, demeurant ainsi authentiques.

De plus, l'ordre des mots et des versets peut être altéré intentionnellement ou non, ce qui complique davantage l'authentification du contenu. Cette problématique n'a pas été pleinement prise en compte, rendant la distinction de l'ordre des mots et des versets encore plus difficile, même pour ceux qui mémorisent le Saint Coran. Il est donc nécessaire de prendre en considération ce problème et d'établir des mécanismes de vérification solides pour garantir que le texte coranique est présenté dans le bon ordre et n'est pas modifié, assurant ainsi sa préservation, sa qualité et son authenticité pour les internautes.

5. Approches pour l'authentification et la préservation de l'intégrité du contenu coranique

L'authentification du contenu constitue un processus essentiel pour déterminer l'originalité d'un contenu donné. Dans cette section, nous explorons les différentes approches utilisées pour garantir l'intégrité du contenu coranique. Chaque approche, qu'il s'agisse du la sécurité qui englobe le tatouage numérique (Watermarking), la cryptographie, la stéganographie, et la Blockchain ; de la recherche, la vérification, et la classification, est examinée en détail. Une synthèse des recherches récentes dans ce domaine est présentée à la fin de chaque section, offrant ainsi un aperçu complet des avancées réalisées en matière de préservation de l'intégrité du contenu coranique.

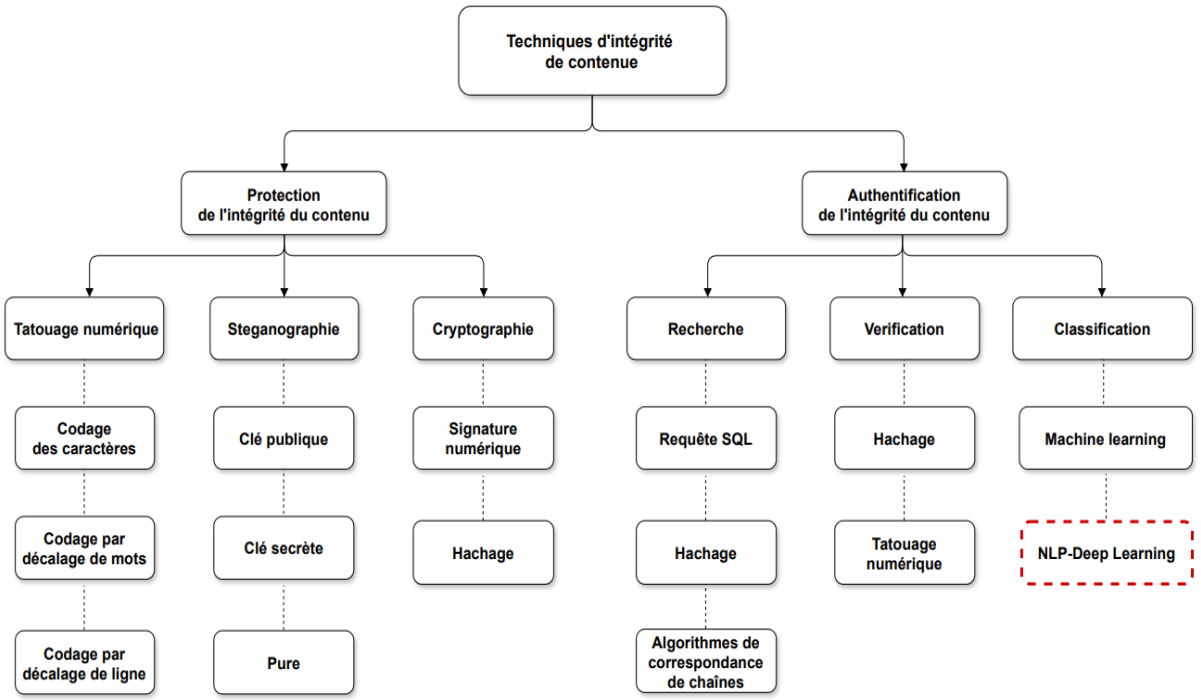


Figure 1. 4:Les techniques d'intégrité de contenue

Lorsqu'ils citent des versets coraniques en ligne, les utilisateurs ont souvent deux

préoccupations principales : s'assurer de l'exactitude et de la correction du verset ou de la phrase citée (e), et protéger leur propre contenu coranique téléversé contre toute altération ou modification. Face à ces préoccupations, Hakak [6] a classé les approches d'intégrité du contenu en deux catégories principales comme montre la figure 1.4 : la protection de l'intégrité du contenu (pour protéger le Saint Coran numérique) et l'authentification de l'intégrité du contenu (pour vérifier le Saint Coran numérique). La protection de l'intégrité du contenu inclut des techniques telles que le tatouage numérique, la cryptographie, la stéganographie et la blockchain. En revanche, l'authentification de l'intégrité du contenu englobe des approches basées sur la recherche, sur la vérification et sur la compréhension linguistique. Chacune de ces approches est brièvement décrite comme suit :

5.1. Protection de l'intégrité du contenu

Bien que notre étude se concentre principalement sur l'authentification de l'intégrité des textes coraniques en ligne, il est essentiel de comprendre les différentes méthodes utilisées pour protéger l'intégrité des textes numériques. Ces méthodes sont souvent mises en œuvre pour prévenir la manipulation ou la corruption des textes, ce qui est crucial pour garantir leur fiabilité et leur authenticité.

Dans cette section nous présentons un bref aperçu des principales approches de protection de l'intégrité.

Lorsqu'il s'agit de protéger l'intégrité du contenu , plusieurs aspects sont cruciaux à prendre en compte [7] [8] [9] :

- Imperceptibilité: Il est essentiel que tout élément de sécurité intégré au document reste invisible pour les utilisateurs ordinaires, préservant ainsi son intégrité [10] [11] .

- Robustesse: Le mécanisme de sécurité doit être robuste, capable de résister à diverses attaques potentielles sans nuire à son efficacité [10] [11] [12].

- Sécurité: La méthode utilisée pour protéger le contenu doit répondre à des normes strictes de sécurité pour garantir sa fiabilité à long terme [10].

- Faible charge de calcul : Il est impératif que le processus de protection soit efficace sur le plan informatique, tout en restant évolutif pour suivre les avancées technologiques futures, avec un minimum d'impact sur les performances [10].

En tenant compte de ces paramètres, diverses approches peuvent être envisagées pour préserver l'intégrité des données. Les techniques suivantes ont été utilisées avec succès pour sécuriser le contenu coranique des different formats contre toute altération ou modification envisagée.

5.1.1. Le tatouage numérique

Le tatouage numérique, une technique largement utilisée pour sécuriser les textes numériques, est souvent employé dans le but de protéger et d'authentifier leur contenu [13]. Cette méthode implique l'incorporation d'informations spécifiques, telles qu'un logo ou un texte, dans le média numérique, que ce soit une image, une vidéo ou un document texte [14]. Son objectif principal est d'assurer l'intégrité et l'identification du propriétaire du contenu, et il trouve des applications dans divers domaines tels que la surveillance des diffusions, l'authentification et la protection des droits d'auteur.

Il existe une variété de techniques de watermarking, qui peuvent être classées en fonction de leur utilisation sur des images ou du texte. Les méthodes basées sur l'image se concentrent généralement sur la manipulation des pixels ou des blocs d'une image, tandis que celles basées sur le texte modifient directement le contenu textuel [6]. Parmi ces méthodes, on retrouve le codage de caractères, le décalage de mots et le décalage de lignes, qui sont largement utilisés pour garantir l'intégrité des textes numériques, en particulier dans le cas des textes sensibles comme le Coran [15].

Dans la pratique, le watermarking a été appliqué avec succès pour protéger le contenu du Coran numérique contre la falsification et la manipulation [16]. Des approches telles que l'utilisation de symboles diacritiques spéciaux ou l'ajout d'espaces supplémentaires dans le texte ont été proposées pour dissimuler des informations de watermarking [17]. De plus, des techniques plus avancées comme l'encodage basé sur la caractéristique des caractères ont été développées pour renforcer la sécurité du watermarking [18].

Malgré ses avantages, le watermarking présente également des limites, notamment en termes de robustesse et de capacité à résister aux attaques malveillantes [19]. Les chercheurs continuent à explorer de nouvelles méthodes et approches pour améliorer la fiabilité et l'efficacité du watermarking dans le contexte de la préservation de l'intégrité des textes numériques, en tenant compte des exigences spécifiques des textes religieux tels que le Coran [20].

5.1.2. La cryptographie

La cryptographie, une technique largement utilisée pour sécuriser les données numériques, trouve également son application dans la protection et l'authentification des textes coraniques en ligne [21].Cette méthode implique la conversion du texte lisible par l'homme en un texte illisible, également connu sous le nom de texte chiffré ou ciphertext, afin de le rendre inaccessible à toute personne non autorisée, à travers des processus tels que l'encryption, la génération de clés et le déchiffrement [22]. Les avantages de la cryptographie dans ce contexte incluent la sécurisation des informations confidentielles entre l'expéditeur et le destinataire des textes coraniques, l'authentification pour prouver l'identité de ces textes, la vérification de leur intégrité pour garantir qu'ils n'ont pas été altérés et la non-répudiation, empêchant ainsi l'expéditeur de nier l'envoi des textes [21].

Dans le domaine de la protection et de l'authentification des textes coraniques, la cryptographie peut être classée en deux approches principales : basée sur les clés et sans clé

[21]. L'approche basée sur les clés comprend la cryptographie symétrique et asymétrique. Dans la cryptographie symétrique, une seule clé est utilisée pour l'encryption et le déchiffrement des textes coraniques, tandis que dans la cryptographie asymétrique, deux clés distinctes sont utilisées [21]. Les algorithmes couramment utilisés dans ce contexte incluent AES, DES et RSA, chacun présentant ses propres avantages et limitations [21].

Une approche populaire dans la cryptographie basée sur les clés pour les textes coraniques est l'utilisation de signatures numériques [23].Les signatures numériques sont des techniques mathématiques utilisées pour authentifier et valider l'intégrité des textes coraniques numériques, permettant de détecter toute tentative de falsification ou d'usurpation dans les communications numériques [23]. Les méthodes sans clé, telles que le hachage et la génération de nombres aléatoires, sont également largement utilisées pour garantir l'intégrité des textes coraniques numériques [24].

Cependant, malgré ses avantages, la cryptographie présente également des limites en termes de vitesse de traitement et de complexité des algorithmes, en particulier dans le contexte des textes coraniques numériques[21]. Les chercheurs continuent à explorer de nouvelles méthodes et approches pour améliorer la sécurité et l'efficacité de la cryptographie dans ce domaine spécifique.

5.1.3. La stéganographie

La stéganographie, une autre méthode utilisée dans le domaine de la protection et de l'authentification des textes coraniques en ligne, se distingue de la cryptographie en dissimulant l'existence même du message original plutôt que de simplement le rendre illisible [25]. Son objectif est d'incorporer le texte coranique dans un support de couverture de manière à ce qu'il soit inaccessible à toute personne non autorisée, offrant ainsi un niveau supplémentaire de confidentialité et de sécurité [26]. La stéganographie peut être classée en plusieurs catégories, chacune avec ses propres méthodes et techniques[24] .

La stéganographie pure ne nécessite pas de clé secrète car aucune autre partie n'est consciente de la communication en cours. Cependant, cette méthode peut être moins sécurisée en raison de son manque de protection contre l'interception [24] . La stéganographie à clé secrète utilise une clé de stéganographie pour la communication, offrant ainsi un niveau supplémentaire de sécurité. Enfin, la stéganographie à clé publique utilise des clés privées et publiques pour assurer une communication sécurisée, bien que cette méthode puisse être plus complexe à mettre en œuvre [24] .

Les approches de stéganographie utilisent différentes méthodes pour incorporer le texte coranique dans le support de couverture. Cela peut inclure des méthodes de substitution, où les parties redondantes de la couverture sont remplacées par le texte coranique, des techniques de transformation de domaine où le texte est incorporé dans l'espace de transformation de la couverture, des techniques de distorsion où le texte est stocké en modifiant les signaux de la couverture, et des méthodes statistiques où le texte est encodé en modifiant les propriétés statistiques de la couverture [24] .

Cependant, malgré ses avantages potentiels en termes de confidentialité et de sécurité, la stéganographie présente également des défis, notamment en termes de capacité de dissimulation du texte coranique, de robustesse face aux transformations et de détection par des tiers [25] . Les chercheurs continuent à explorer de nouvelles méthodes et techniques pour améliorer l'efficacité et la sécurité de la stéganographie dans le contexte spécifique de la protection et de l'authentification des textes coraniques en ligne.

5.1.4. La Blockchain

La blockchain est une technologie de registre distribué sécurisé qui permet d'enregistrer de manière transparente et immuable les transactions. La blockchain utilise Les contrats intelligents sont des programmes auto-exécutables qui définissent les règles et les conditions d'une transaction et sont enregistrés sur la blockchain de manière immuable [27] [28].

Dans le contexte de la préservation de l'intégrité des textes coraniques, les approches basées sur la blockchain utilisent cette technologie pour garantir l'authenticité et la non-altération des textes sacrés. Ces approches reposent sur des enregistrements décentralisés et sécurisés des transactions liées à l'intégrité du texte coranique, offrant ainsi un système de confiance numérique.

Chaque transaction liée à l'intégrité du texte coranique, telle que la vérification de son authenticité ou la protection contre la falsification, est enregistrée sur la blockchain de manière immuable. Les participants au réseau peuvent ainsi vérifier la validité de ces transactions sans avoir besoin d'une autorité centrale de confiance. Par exemple, en utilisant la blockchain Ethereum, un cadre de confiance numérique peut être développé pour sécuriser l'intégrité des textes coraniques en ligne, offrant ainsi une protection contre les altérations malveillantes ou accidentelles [28].

Bien que les approches basées sur la blockchain offrent un moyen prometteur de sécuriser l'intégrité des textes coraniques en ligne, elles présentent également certaines limites et considérations. La scalabilité et les coûts de transaction peuvent être des défis à prendre en compte. De plus, Les Blockchains peuvent être sujettes à des attaques de 51 %, où un groupe d'acteurs malveillants contrôle la majorité du pouvoir de calcul, compromettant ainsi l'intégrité des données enregistrées [27].

D'après l'analyse ci-dessus, il est évident que toutes les approches associées présentent plusieurs inconvénients en termes d'imperceptibilité, de robustesse, de sécurité et de coût informatique. Par conséquent, une attention accrue devrait être portée à ces domaines, et des tests rigoureux de robustesse face à diverses transformations, en utilisant différentes méthodologies et paramètres expérimentaux mentionnés précédemment, sont nécessaires.

Tableau 1. 1: Avantages et inconvénients des approches de protection de contenue coranique

Approches	**Avantages**	**Inconvénients**

Tatouage numérique	Offre une robustesse contre les attaques potentielles, garantissant ainsi l'intégrité du contenu [29]	Vulnérable à diverses attaques telles que les attaques géométriques, de bruit et autres, compromettant potentiellement l'intégrité du contenu.
Cryptographie	Assure une communication sécurisée sur des canaux non sécurisés, ce qui la rend adaptée aux attaques liées au réseau [30].	Plus adaptée aux attaques liées au réseau par rapport à la sécurisation des documents numériques, ce qui peut avoir des limites dans la préservation de l'intégrité des documents sensibles tels que le contenu coranique [31].
Steganographie	Permet la transmission de messages sans éveiller de soupçons, offrant ainsi une méthode de communication secrète [32].	Moins sécurisée que d'autres méthodes et peut ne pas être adaptée à la préservation de l'intégrité de documents sensibles comme les images ou le texte, ce qui peut entraîner des vulnérabilités [32].
Blockchain	Fournit un enregistrement décentralisé et immuable, garantissant l'intégrité et l'authenticité du contenu coranique [27] [28] .	Vulnérable à des attaques telles que les attaques de 51% et les vulnérabilités des contrats intelligents, ce qui peut compromettre la sécurité et l'intégrité des données stockées [27] [28].

Bien que ces méthodes soient essentielles pour garantir la sécurité et l'intégrité des textes numériques, notre recherche se concentre sur l'authentification de leur intégrité, en mettant l'accent sur les techniques qui permettent de vérifier et de valider l'authenticité des textes coraniques en ligne. En examinant ces méthodes de protection, nous comprenons mieux le contexte dans lequel notre travail s'inscrit et les défis auxquels il répond.

5.2. Authentification de l'intégrité du contenu

L'authentification de l'intégrité du contenu du Saint Coran est une démarche essentielle pour garantir la précision et la fidélité des textes sacrés numérique. Cette procédure revêt une importance cruciale dans le contexte de l'étude et de l'analyse du contenu coranique disponible sur Internet. Divisée en trois approches distinctes, à savoir l'approche basée sur les techniques de recherche, l'approche basée sur les techniques de vérification, et l'approche basée sur les techniques de classification, cette démarche nécessite une approche méthodique et rigoureuse pour assurer son efficacité. Dans cette perspective, la précision et la robustesse de représentation de contenue émergent comme des critères essentiels à prendre en compte pour garantir la fiabilité et la pertinence des résultats obtenus.

5.2.1. La recherche

Le Saint Coran, étant originellement enregistré en langue arabe, présente un défi pour de nombreux musulmans qui peuvent lire l'arabe mais ne sont pas capables de détecter toute modification indue. Afin de vérifier l'authenticité d'un verset spécifique, cette approche consiste à effectuer une recherche de ce verset et à le comparer avec une source vérifiable. Dans cette phase de correspondance et de recherche, l'élément crucial est la disponibilité d'un contenu vérifié, c'est-à-dire une base de données authentique. En l'absence d'un tel contenu vérifié, toute tentative de vérification de l'intégrité des données est vouée à l'échec. Cette approche englobe trois sous catégories distinctes pour mieux cerner les défis spécifiques rencontrés.

- **Requêtes SQL :** SQL constitue une base de données standard [33]. SQL est un langage de script basé sur les requêtes, et les requêtes disponibles dans SQL sont utilisées pour rechercher un contenu particulier. Certains des opérateurs de recherche puissants en SQL sont SELECT et LIKE. Cependant, l'utilisation d'expressions régulières qui inclut des préfixes et des suffixes est également utilisée dans la phase de recherche [34] [33].

- **Algorithmes de correspondance de chaînes :** Les algorithmes de correspondance de chaînes sont classés en algorithmes de correspondance exacte et algorithmes de correspondance approximative. Il existe de nombreux algorithmes qui peuvent être utilisés pour rechercher une requête particulière, tels que les algorithmes de Boyer-Moore qui sont considérés comme les algorithmes de correspondance de chaînes standard [35] [36].

- **Hashage :** Le hashage ou le condensé de message (MD) signifient tous deux la même chose, à savoir un processus unidirectionnel où des chaînes de caractères sont converties en une clé ou une valeur de longueur fixe. Il est considéré comme un processus rapide pour récupérer des éléments de la base de données [37]. Ces techniques sont plus préoccupées par la précision que par le surcoût de performance. Les approches de hachage ont le problème d'échanger des clés publiques entre de nombreuses parties communicantes. Certaines des approches de hachage les plus couramment utilisées sont les Message Digest (MD), Secure Hash (SHA) et le code d'authentification de message de hachage à clé (HMAC). Les messages digest sont également connus sous le nom de condensés de message adaptés au hachage de signatures numériques en valeurs plus courtes. De même, la famille SHA est adaptée pour créer de plus grands condensés de message [37].

Ces approches sont limitées pour vérifier l'authenticité de motifs spécifiques et courts uniquement, ce qui peut entraîner une précision réduite dans la détection du contenu coranique authentique au milieu d'un texte plus vaste. De plus, l'intégration de contenu coranique dans des textes en ligne, comprenant des métaphores et un langage familier ou figuré, complexifie davantage la tâche d'authentification. Cette complexité supplémentaire peut entraîner une augmentation du temps nécessaire pour mener à bien le processus d'authentification, car une analyse approfondie est requise pour distinguer le contenu

authentique des autres éléments. De ce fait, la nécessité d'une précision accrue et la gestion de la complexité temporelle constituent des défis majeurs à surmonter dans l'application de ces approches pour l'authentification du contenu coranique numérique.

5.2.2. La vérification

La vérification englobe le processus de confirmation de l'authenticité et de l'intégrité du contenu. Cela implique de valider que le contenu reflète fidèlement la source originale sans aucune altération ou modification non autorisée. Dans ce contexte, la vérification constitue un élément crucial pour garantir la fiabilité du contenu coranique dans les formats numériques. Diverses approches sont utilisées pour faciliter la vérification, chacune visant à évaluer différents aspects de l'intégrité du contenu, tels que la précision, la cohérence et le respect des normes établies. Il peut exister d'autres mécanismes de vérification, mais les plus utilisés sont les trois approches de vérification suivantes :

- **Le processus de hachage :** dans le processus de hachage, une chaîne de caractères est convertie en une clé spécifique pour représenter la chaîne originale. La vérification de l'intégrité de cette chaîne se fait en calculant sa valeur de hachage et en la comparant à une valeur de référence. Si les valeurs de hachage correspondent, les données sont considérées comme authentiques ; autrement, si elles diffèrent, cela indique une altération. Cependant, il est important de noter que le processus de hachage ne peut pas toujours détecter les modifications subtiles dans les données, ce qui pourrait compromettre l'authenticité des données si elles sont manipulées de manière malveillante. Les méthodes de tatouage numérique, de hachage ou de Message Digest (MD) sont des approches similaires utilisées dans ce contexte. Ces techniques sont principalement axées sur la précision plutôt que sur la surcharge de performance. Néanmoins, les approches de hachage rencontrent le défi initial de l'échange de "clés publiques" entre de multiples parties communicantes. Les approches de hachage couramment utilisées incluent Message Digest (MD), Secure Hash (SHA) et le Hash de code d'authentification de message (HMAC) [37].

- **Le tatouage numérique :** Le tatouage numérique constitue une autre méthode pour garantir l'intégrité des données. Il implique l'incorporation d'un tatouage ou d'un logo spécifique à un document ou à une image afin d'en assurer l'authenticité.

5.2.3. La classification

La classification des données coraniques implique le processus de catégorisation des textes pour déterminer s'ils sont d'origine coranique ou non. Cela nécessite des techniques sophistiquées de NLP et de ML pour identifier les caractéristiques distinctives des versets coraniques et les distinguer des autres textes. Bien que diverses approches aient été explorées dans le domaine de la récitation coranique, peu de travaux ont été réalisés dans le domaine de la classification des données coraniques. Parmi ces approches, l'utilisation de techniques de ML est l'une des plus prometteuses. Les algorithmes de ML ont démontré leur efficacité dans la classification des données textuelles et pourraient être adaptés pour la classification des

versets coraniques. Cette approche pourrait permettre une vérification rapide et précise de l'authenticité des versets coraniques dans les formats numériques.

- **La machine learning :** Dans le cadre de la vérification du contenu coranique, une technique de ML particulièrement pertinente est celle de la classification supervisée. Cette approche implique l'utilisation de modèles de ML entraînés sur des données annotées pour distinguer automatiquement entre le contenu coranique authentique et les altérations ou les contenus non autorisés. Les détails de cette approche incluent la collecte et la préparation d'un ensemble de données représentatif comprenant à la fois des exemples de contenu coranique authentique et des exemples d'altérations ou de contenus non autorisés. Ensuite, ces données sont utilisées pour entraîner un modèle de DL, tel qu'un classificateur binaire, qui est capable de distinguer avec précision entre les deux catégories de contenu. Le processus d'entraînement du modèle implique généralement plusieurs étapes, telles que la sélection des caractéristiques pertinentes à partir des données d'entrée, le choix d'un algorithme d'apprentissage adapté, et l'optimisation des paramètres du modèle pour améliorer ses performances. Une fois que le modèle est entraîné avec succès, il peut être utilisé pour classer automatiquement de nouveaux exemples de contenu coranique et identifier toute forme de manipulation ou d'altération. Cette approche de ML offre l'avantage de pouvoir détecter les altérations du contenu coranique avec une précision élevée et à grande échelle, ce qui en fait un outil précieux pour assurer l'intégrité et l'authenticité du texte sacré.

- **NLP-Deep Learning :** Les techniques de traitement automatique du langage naturel (NLP) jouent un rôle central dans la préservation et l'analyse du contenu numérique en ligne. Le NLP englobe un large éventail de méthodes et d'algorithmes qui permettent de comprendre et de traiter le langage humain de manière automatisée. Dans le contexte de la compréhension linguistique du Coran, ces techniques sont essentielles pour interpréter et analyser le texte sacré, en tenant compte des nuances linguistiques et des contextes culturels. Les méthodes de NLP comprennent l'analyse syntaxique pour décomposer les phrases en composants grammaticaux, la sémantique pour comprendre le sens des mots et des phrases, ainsi que la comparaison avec des sources authentiques pour garantir la fidélité du texte. En combinant le NLP avec des approches de ML et DL, il devient possible d'automatiser l'analyse et l'interprétation du contenu coranique avec une précision accrue, ouvrant ainsi la voie à de nouvelles possibilités de préservation et d'étude de ce texte sacré. Les techniques de DL , en particulier, permettent de traiter de vastes ensembles de données textuelles et d'extraire des motifs complexes, ce qui améliore la capacité des modèles à comprendre et à interpréter le Coran dans sa profondeur et sa richesse linguistique [38].

Les efforts visant à authentifier l'intégrité du contenu du Saint Coran sont résumés comme

suit :

Dans le but d'authentifier les citations du Coran lors de la recherche de texte, Alshareef et Saddik [34] ont proposé un cadre de détection des versets coraniques. Ce cadre prend un verset coranique en entrée et détermine son authenticité. Il comporte deux composantes principales : un filtre pour les citations coraniques et un mécanisme de vérification. Le processus de filtrage commence par éliminer les diacritiques arabes et les symboles spéciaux afin d'améliorer la compatibilité avec les moteurs de recherche traditionnels, bien que cette affirmation manque de justification. Ensuite, un mécanisme de vérification basé sur des requêtes SQL à expressions régulières est utilisé pour valider le texte. L'authenticité des échantillons de versets sélectionnés et de mots arabes spécifiques est évaluée et comparée aux résultats obtenus à partir des moteurs de recherche coraniques. L'algorithme proposé atteint un taux de précision de 89% par rapport aux autres moteurs de recherche. Cependant, cette précision peut fluctuer lorsque l'algorithme est appliqué à des ensembles de données plus importants, en particulier en raison de l'approche préfixe-suffixe utilisée par les moteurs de recherche réguliers face à de vastes corpus arabes.

Dans le but de vérifier les versets arabes accompagnés de diacritiques et de symboles, l'étude d'Alginahi et al. [39] a proposé un algorithme capable de détecter les versets complets et partiels. Pour l'évaluation, les paramètres "vérifié et authentifié" et "altéré" sont pris en compte. Cependant, aucun détail n'est donné sur le mécanisme de l'algorithme. De plus, le schéma de flux suggère l'utilisation d'une approche SQL simpliste pour la sélection de la requête, ce qui est considéré comme inefficace car cela nécessite de spécifier un emplacement particulier en premier lieu. De plus, l'algorithme représenté dans le schéma de flux prend en compte tous les symboles et diacritiques, bien qu'il les supprime avant la conversion au format Unicode.

Dans le but de détecter et d'authentifier les versets coraniques extraits de sources en ligne comme les forums, Sabbah et Selamat [40] ont introduit un nouveau cadre. L'objectif est d'améliorer la précision de détection du texte diacritique. Le cadre proposé comprend un extracteur qui identifie trois listes à partir du script coranique : les mots diacritiques coraniques distinctifs, les lettres distinctives, ainsi que les diacritiques et symboles. Chaque élément se voit attribuer un poids distinctif. Après cette pondération, les éléments sont regroupés en ensembles de caractères et de diacritiques. L'exactitude moyenne atteinte est de 62%, avec une précision de 75% et un rappel de 78%. Cependant, cet algorithme présente des limites. Il ne fonctionne pas sur du texte non diacritique, et le calcul des poids ainsi que la division des versets en deux groupes entraînent une surcharge importante. De plus, la complexité de l'algorithme augmente avec la présence de nombreux signes diacritiques dans le texte. Bien que visant à améliorer la détection des versets coraniques, cette approche rencontre donc des défis en termes de performance et de traitement des textes non diacritiques.

Dans le but de répondre aux exigences spécifiques de la vérification du contenu en ligne, Sabbah et Selamat [41] ont cherché à développer un modèle ML capable de classer les mots en ligne en mots coraniques et non coraniques en utilisant le model de support vector

machine. Un modèle de classification est appliqué aux mots extraits de la source en ligne. Cependant, avant l'extraction des mots, tous les symboles, diacritiques et lettres non arabes sont supprimés lors de la phase de filtrage. Le modèle de classification est évalué à l'aide de trois paramètres, à savoir l'exactitude, la précision et la mesure F [41]. Cette approche mentionne l'utilisation de différentes catégories de caractéristiques pour la classification, mais l'analyse détaillée de l'efficacité de chaque catégorie de caractéristiques n'est pas largement discutée.

Dans le but de vérifier l'intégrité et l'authenticité des versions électroniques du Coran en ligne, Alsmadi et Zarour [42] ont développé une méthode basée sur le hachage. Cette approche vise à authentifier et vérifier les versets coraniques, et les résultats ont indiqué que la vérification par hachage pourrait constituer une méthode fiable d'authentification automatique avec des niveaux de confiance élevés. Cependant, certains cas exceptionnels, tels que des lectures coraniques différentes, méritent une enquête et une évaluation plus approfondies. De plus, cette méthode est adaptée uniquement pour un seul verset à la fois, ce qui peut être inefficace pour les applications nécessitant la vérification de plusieurs versets simultanément. Il est également important de noter qu'une collision de hachage peut se produire, compromettant ainsi la fiabilité du processus d'authentification.

Dans le but de développer un système d'authentification novateur pour le Coran et les Hadiths, Kamsin et al. [43] vise à vérifier l'authenticité des copies numériques du Coran et à détecter les différences avec le Mushaf Uthmani standard. Le système est conçu pour analyser les variations entre les copies numériques du Coran à différents niveaux, y compris par sourate, verset, lettres, diacritiques, tajweed et autres symboles de récitation. Les tests impliquent de valider l'exactitude des textes du Coran numérique par rapport au mushaf Uthmani standard, en veillant à l'efficacité du système dans l'authentification. Pour atteindre ces objectifs, le système utilise une approche de correspondance de chaînes centrée sur l'Unicode pour comparer les données authentifiées et les données de test. Des techniques de ML sont également employées pour classer l'authenticité des textes coraniques et des Hadiths. L'algorithme développé vise à améliorer l'efficacité et la précision de la vérification de l'authenticité des applications numériques du Coran.

Dans le but de mettre en place un système d'authentification permettant de distinguer les versets falsifiés des originaux, Hakak et al. [5] ont proposé un modèle pour l'authentification automatique du Coran numérique. Le cadre de ce modèle d'authentification se compose de deux phases : la tokenisation et la recherche. La phase de tokenisation vise à diviser le verset d'entrée donné en caractères individuels et à convertir l'ensemble de la chaîne dans son format Unicode. Quant à la phase de recherche, son objectif est de vérifier les entrées de la base de données à l'aide d'un algorithme de correspondance exacte. L'approche est précise à 100 % en termes de détection du verset complet.

Le Tableau 1.2 résume les travaux réalisés pour déterminer l'authenticité de format textuel de Coran numérique.

Tableau 1. 2: Synthèse des travaux sur l'authentification du format textuel du Coran numérique

Ref	Objective	Approche	Résultats	Limitations	Evaluation
[34]	Détecter les faux versets coraniques	Les requêtes SQL	Un cadre de recherche pour authentifier le verset coranique	Il est obligé de saisir le nom de la sourate pour trouver un verset particulier.	La précision
[40]	Détecter et authentifier les versets coraniques	Techniques de correspondance de chaine	Un cadre pour détecter et authentifier le texte coranique.	N'étant pas adapté au traitement de textes non diacritiques et voyant sa complexité augmentée de manière significative avec la présence de nombreux signes diacritiques dans les versets	La précision, le recall
[39]	Vérification du texte coranique cité avec des diacritiques et des symboles de Tajweed	Technique de correspondance de chaîne centrée	Un algorithme pour la détection de versets arabes particuliers.	L'utilisateur doit connaître le verset particulier à vérifier. L'échec de l'algorithme est possible si le verset est le dernier d'une sourate et le premier de la suivante	Proposer deux paramètres pour évaluer la sortie de l'algorithme, à savoir "Vérifié et authentifié" et "altéré".
[41]	Vérification en ligne du contenu coranique	Machine Learning (SVM)	Un modèle de classification des mots en mots coraniques et non-coraniques	Adapté uniquement pour les mots.	La précision, Le recall et le F-score
[42]	Authentifier et vérifier les versets	Hachage	Une méthode basée sur le hachage pour authentifier et	Adapté uniquement pour un seul verset.	Aucune métrique. évaluation effectuée sur

	coraniques		vérifier les versets coraniques	Une collision de hachage peut se produire	la base de la comparaison des valeurs de hachage de différents textes.
[43]	Evaluer l'authenticité des applications numériques du Coran	Technique de correspondance de chaînes	Un prototype pour system d'authentification du Coran	Le système a été testé uniquement sur des copies de mushaf d'Uthman.	Aucune métrique
[5]	Authentifier la version diacritique du Coran	Correspondance de chaîne centrée sur Unicode	Un prototype pour l'authentification du Coran	N'étant adapté au traitement de textes non diacritiques	La précision

Le Tableau 1.3 répertorie les principaux inconvénients des approches utilisées pour déterminer l'authenticité du Coran numérique, tels que discutés dans les sections précédents.

Tableau 1. 3: Principaux inconvénients des approches pour l'authentification du Coran numérique

Approches	**Inconvénients**	**Recommandations**
Requêtes SQL	Si l'index n'est pas fourni. En cas de fourniture de l'index, l'utilisateur doit s'assurer que le verset d'entrée appartient à quel chapitre. Vulnérable aux attaques par injection SQL.	Non adapté pour rechercher des versets coraniques aléatoires
Algorithmes de correspondance de la chaine	Résultats en performances plus lentes en ce qui concerne le temps de traitement des textes non basés sur l'ASCII en raison de l'utilisation de différentes techniques de codage.	Approche potentielle pour rechercher et vérifier des versets coraniques.
Hachage	Risque de collision de hachage. Surcharge associée en cas de nécessité d'authentifier de nombreux versets coraniques individuels. pour la compression, La conservation des valeurs de hachage et des données compressées nécessite un espace de stockage supplémentaire	Non recommandé pour authentifier des citations coraniques individuelles. Recommandé pour des pages entières ou multiples.
Tatouage	Susceptible à de nombreuses attaques telles que géométriques, du bruit si un	Non recommandé en raison du nombre possible d'attaques et

numérique	logo d'image est intégré dans le texte brut. Dans le cas de bits intégrés dans le texte brut, la surcharge associée à sa sécurisation.	de la surcharge associée à l'incrustation de filigranes pour le texte brut.
Machine Learning	Limite de l'analyse au niveau du mot. Ne prenant pas en compte le contexte global du verset. Possibilité de classification erronée en raison du manque de compréhension du sens global.	Recommandé d'explorer des techniques plus sophistiquées pour une compréhension plus profonde.

Sur la base de l'analyse de toutes les approches qui ont été utilisées pour déterminer l'authenticité des versets coraniques basés sur du texte, l'examen des travaux connexes présentés dans le tableau 1.2 a indiqué que l'accent principal des approches a été mis sur la précision plutôt que sur la robustesse de représentation de contenue.

Tableau 1. 4: Les paramètres importants pour l'authentification de l'intégrité du contenu numérique coranique

Approche	[34]	[40]	[39]	[41]	[42]	[43]	[5]
Précision	✓	✓	✓	✓	✗	✗	✓
Robustesse	✗	✗	✗	✗	✗	✗	✗

Cela signifie que leur objectif principal était d'obtenir des résultats aussi précis que possible dans la classification des versets coraniques en fonction de leur authenticité. Cependant, un aspect qui a souvent été négligé est la robustesse de la représentation des données.

La robustesse de la représentation des données fait référence à la capacité d'un modèle ou d'une méthode à traiter efficacement la variabilité et la complexité du contenu coranique. Cela inclut la capacité à gérer les variations linguistiques, les différences de style et de syntaxe, ainsi que les nuances théologiques et culturelles présentes dans le texte coranique. Une représentation robuste des données est essentielle pour garantir que le modèle ou l'algorithme est capable de fournir des résultats fiables et cohérents, quel que soit le contexte dans lequel il est utilisé.

En négligeant la robustesse de la représentation des données, les approches existantes risquent de ne pas être pleinement adaptées à la diversité du contenu coranique. Elles peuvent être susceptibles de produire des résultats incohérents ou incorrects dans des cas où le texte présente des variations linguistiques importantes ou des nuances subtiles qui nécessitent une compréhension approfondie du contexte théologique et culturel.

Ainsi, il devient impératif pour les futures recherches dans ce domaine de prendre en compte non seulement la précision des résultats, mais également la robustesse de la représentation des données. Cela implique de développer des méthodes et des modèles qui sont capables de capturer la richesse et la diversité du texte coranique de manière précise et fiable, afin de

garantir une authentification efficace et précise des versets coraniques.

6. Problèmes ouverts, défis et solutions possibles

Les approches qui reposent sur des techniques telles que SQL, les algorithmes de correspondance de la chaine, le tatouage numérique et le ML pour l'authentification ou la vérification de l'intégrité du texte coranique peuvent être limitées dans leur capacité à fournir une solution complète et fiable. Ces méthodes exigent souvent une étape préalable d'identification manuelle du texte coranique ou nécessitent des informations supplémentaires telles que le numéro de verset ou de chapitre pour fonctionner efficacement. Cette dépendance à des données préalables peut constituer un obstacle dans un contexte en ligne où le texte peut être dynamique et sujet à des mises à jour fréquentes. De plus, ces approches traditionnelles peuvent ne pas être en mesure de capturer la complexité linguistique et contextuelle du texte coranique, ce qui compromet leur capacité à fournir une vérification précise de son intégrité. En outre, ces méthodes peuvent être plus sensibles aux tentatives de contournement ou d'attaques malveillantes, car elles reposent souvent sur des règles prédéfinies ou des modèles préexistants qui peuvent être contournés par des adversaires déterminés. Pour surmonter ces limitations, il est impératif d'explorer des approches plus avancées et adaptées, telles que le DL avec les techniques de représentation sémantique de NLP. Ces méthodes peuvent analyser de manière approfondie la structure et le contenu du texte coranique, en capturant les relations sémantiques et contextuelles entre les mots pour une vérification plus précise et fiable de son intégrité. De plus, les modèles de DL sont souvent robustes face aux tentatives de contournement et peuvent s'adapter de manière dynamique aux changements dans le texte, offrant ainsi une solution plus efficace et durable pour garantir l'intégrité du texte coranique en ligne. Cependant, il est important de noter que l'approche utilisant le ML est basée sur le calcul du pourcentage de diacritiques. Cependant, cette méthode peut présenter des limitations lorsqu'elle est appliquée à des textes arabes simples ne contenant pas de diacritiques, ce qui peut entraîner une précision réduite dans l'identification des mots.

Par conséquent, une approche alternative est nécessaire pour améliorer la phase d'identification et d'authentification afin de repérer et d'authentifier les textes coraniques.

7. Conclusion

Ce chapitre passe en revue les études récentes sur l'un des textes les plus sensibles, à savoir la protection et l'authenticité du Saint Coran numérique. Cette aire soulève de nombreuses questions qui demandent une réponse résolue et opportune sous la forme d'efforts de recherche intensifiés. L'authentification et la protection du Coran sont confrontées à de nombreux défis, notamment l'amélioration de la précision et de la justesse de detection de texte. Dans ce chapitre, les défis les plus courants sont identifiés et des solutions sont proposées. Un bref aperçu des recherches existantes dans ce domaine est présenté, les limitations possibles et leurs résultats sont évalués.

Par conséquent, sur la base de l'examen de la littérature, le problème de l'authentification du

Coran ainsi que la garantie de la précision et de la robustesse de la représentation du texte nécessitent une attention immédiate. L'appel à un modèle intelligent capable de représenter et de vérifier avec précision le contenu coranique en mettant l'accent sur la précision et la robustesse émerge comme une direction critique pour les futures recherches dans ce domaine. Cette recherche vise à relever ces défis pressants et à contribuer à l'avancement de l'authentification de l'intégrité du contenu textuel du Coran.

Chapitre 2

Représentation du contenu numérique du Saint Coran dans le domaine de NLP

1. Introduction

Dans ce chapitre, nous abordons le Traitement du Langage Naturel (NLP), apparu au début des années 1950 comme une branche des techniques d'intelligence artificielle travaillant du côté linguistique [44]. Cette technologie vise à obtenir des résultats automatiques dans le traitement des contenus des langues naturelles. Bien que l'arabe soit l'une des langues les plus anciennes et les plus répandues sur Internet, le traitement automatique de cette langue fait défaut, notamment dans le domaine de l'intelligence artificielle.

Ce chapitre se concentre sur l'étude du contenu des textes arabes les plus prestigieux, à savoir le texte sacré du Coran. La représentation du texte du Coran est essentielle pour utiliser les techniques de DL afin de découvrir les informations cachées derrière les mots et les phrases du Coran. Nous explorerons les techniques de représentation de texte, les statistiques relatives au contenu du Coran en termes de lettres, de mots, de versets et de chapitres, ainsi que les méthodes de conversion en vecteurs à l'aide de techniques de représentation de données. Ces éléments constituent une cible d'inférence utile dans les études reposant sur les réseaux neuronaux dans le traitement des langues naturelles. Le projet Tanzil [45] , le plus fiable a été utilisé pour ces statistiques. Nos résultats diffèrent de nombreuses autres statistiques, en particulier en ce qui concerne la fréquence des lettres et des mots, en raison des différentes sources de données et des styles d'écriture du Coran.

2. Le texte coranique

Le Saint Coran est un livre noble pour les musulmans. Il s'agit des paroles de Dieu révélées de la sourate Al-Alaq à la sourate Al-Nasr et enregistrées dans une copie physique appelée Mush'af, qui ont été arrangées de la sourate Al-Fatiha à la sourate Al-Nas.

2.1.Caractéristiques du texte coranique

Le Coran est divisé en 30 parties (Juz), chaque partie étant divisée en deux sections (Hizb), et chaque section étant divisée en quarts, pour un total de 114 chapitres (Sourates) divisés en "chapitres mecquois" révélés avant la migration à Médine, et "chapitres médinois" qui sont venus après la migration à Médine. Les chapitres contiennent plus de 6 mille versets (Ayah), leur longueur varie entre 1 et 129 mots, avec un total de plus de 78 000 mots. Le Tableau 1 montre les statistiques détaillées du Saint Coran.

Tableau 2. 1: Statistiques du Saint Coran [46]

Contenus	Total	Contenus	Total
Pages	604	Versets	6236
Parties	30	Versets Makki	4613
Pages par partie	20-23	Versets Madani	1623
Sections	60	Mots	78245
Chapitres	114	Mots uniques	14870
Chapitres Makki	86 (75.4%)	Mots vocalisés	19273
Chapitres Madani	28 (24.6%)	Caractères	332837

Le Coran est ponctué par un ensemble de signes, positionnés soit sur la même ligne du verset, soit au-dessus de celui-ci, chacun ayant sa propre signification. Les plus célèbres sont les signes d'arrêt (signes de Waqf), le signe de prosternation (signe de Sujud), le signe de chapitre et le signe de verset. Ces signes sont résumés et expliqués dans le tableau suivant.

Tableau 2. 2:Principaux signes du Coran

Type		Sign	Definition	Total
Waqf	Waqf lazim	ۘ	Doit s'arrêter: 100% arrêt	22
	Saktah	ۚ	Faire une petite pause	5
	Waqf Muraqabah	ۛ	Arrêt à un point	12
	Waqf Awla	ۗ	Mieux vaut s'arrêter: 70% arrêt, 30% continuer	603
	Waqf Jaiz	ۖ	S'arrêter ou continuer: 50% arrêt, 50% continuer	1972
	Waqf Hasan	ۘ	Mieux vaut continuer: 70% continuer, 30% arrêt	1682
	Waqf Qabih	ۙ	Ne jamais s'arrêter: 100% continuer	68
Sujud		۩	Obligation de faire le sujud	15
Chapitre		۞	Fin du ¼ de section	199
Verset		﴿﴾	Fin du verset et son numéro	6236

2.2.Statistiques du texte coranique

Le Coran se distingue par sa richesse linguistique et grammaticale ainsi que par la puissance de ses mots et de ses prononciations. Pour approfondir la composition du texte coranique, nous présentons dans la partie suivante diverses statistiques pour chaque lettre, mot, verset et thème qui constituent le Coran.

2.2.1. Lettres coraniques

Le Coran est écrit de droite à gauche par rapport à la langue arabe. La langue arabe se compose de 28 lettres de base (ا ،ب ، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، ل، م، ن، هـ، و، ي) auxquelles s'ajoutent les lettres modifiées (ة،ى،آ) et les hamzas (ء،أ،إ،ؤ،ئ) pour compléter les lettres qui composent le Coran. Le graphique ci-dessous (Figure 2.1) montre les fréquences des lettres dans le Coran. Il semble que la lettre "ا" soit la plus utilisée, avec un total d'utilisation de 43878 fois, suivie des lettres "ن" ,"ل" et "م" avec un taux de 38639, 27382 et 27071 fois respectivement. Les autres caractères apparaissent de 1 000 à 15 000 fois, à l'exception des lettres "ؤ" et "ظ", qui apparaissent moins d'un millier de fois.

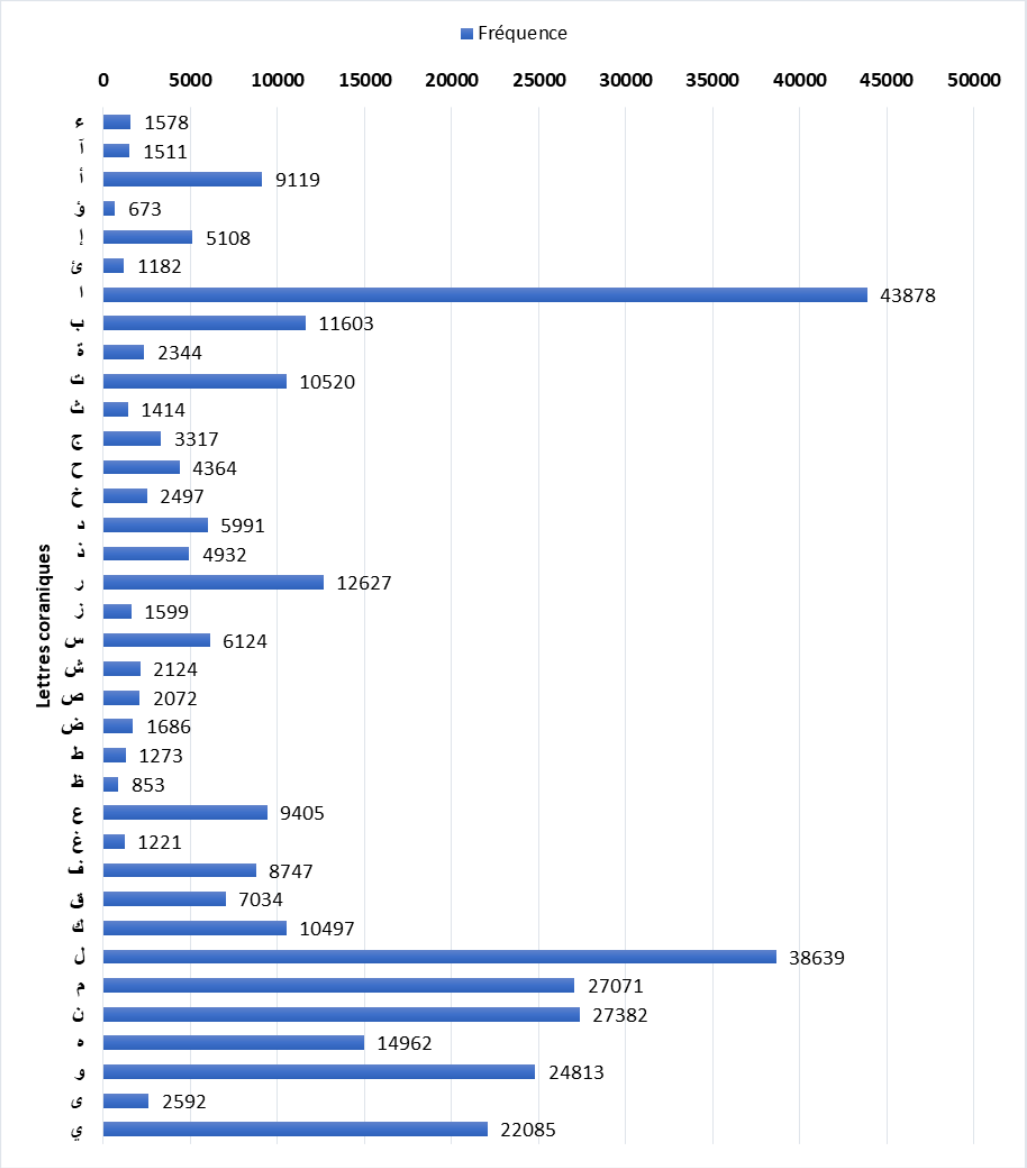


Figure 2. 1:Répartition des lettres dans le Coran

2.2.2. Lettres disjointes du Coran

Les lettres disjointes, ou "HUROOF AL MUQATTA'AT" ou "AL fawātiḥ", tirent leur nom de leur apparition dans les premiers 25,43 % des sourates après le Bismillāh. Leur longueur varie de 1 à 5 lettres, et elles apparaissent 30 fois dans 29 versets. Le graphique suivant (Figure 2.2) résume le nombre et les emplacements des lettres disjointes.

- La lettre "الم" apparaît dans les deux chapitres Madaniens n° 2 et 3 (Al-Baqara et El-Imran) ainsi que dans les quatre chapitres Makkiens successifs du chapitre n° 29 au chapitre n° 32, à savoir (al-ʿAnkabūt, Ar-Rūm, Luqmān et As-Sajdah).

- La lettre "الر" apparaît dans les chapitres Makkiens successifs du chapitre n° 10 au chapitre n° 15 (Yūnus, Hūd, Yūsuf, Ibrāhīm, al-Ḥijr) à l'exception du chapitre n° 13 (Ar-Raʿd) qui commence par "المر".

- Le chapitre Makki n° 19 (Maryam) commence par "كهيعص".

- "طس" apparaît dans le chapitre Makki n° 27 (anNaml) entre les deux chapitres Makkiens (Ash-Shuʿārāʾ et al-Qaṣaṣ), qui commencent par "Tasam".

- Chacune des lettres "ص" ,"يس", "طه" et "ق" ouvrent respectivement les chapitres Makkiens numérotés 20, 36, 38 et 50, qui ont un nom identique aux lettres (Qāf, Ṣād, Yā Sīn, Ṭā Hā).

- La lettre "حم" apparaît dans chacun des chapitres Makkiens successifs du chapitre n° 40 au chapitre 46 (Ghāfir, Fuṣṣilat, Ash-Shūrā, Az-Zukhruf, Al Dukhān, Al-Jāthiya et Al-Aḥqāf), où le chapitre n° 42 (Ash-Shūrā) se distingue par la présence d'une deuxième lettre "عسق".

- Enfin, la lettre "ن" apparaît dans le chapitre Makki n° 68 (Al-Qalam).

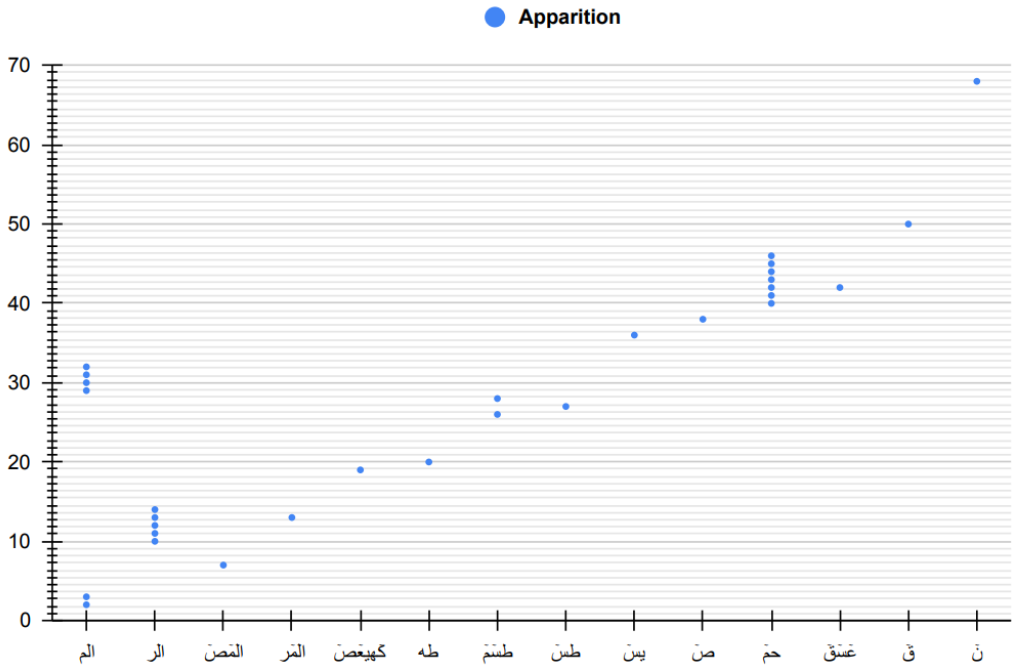


Figure 2. 2:Emplacements de lettres disjointes

2.2.3. Mots du Coran

Parmi les secrets des miracles du Noble Coran se trouve la répétition de ses mots et de ses phrases. Cette répétition diffère de la répétition ennuyeuse contenue dans le langage ordinaire, car elle est une méthode d'éloquence, et son but est l'affirmation et la vénération. La répétition est divisée en deux types, le premier étant une répétition synonymique (ou sémantique), dans laquelle les mots diffèrent afin de maintenir le même sens. Par exemple, rappeler à Dieu Tout-Puissant le paradis dans plus d'un verset.

سَلَامٌ عَلَيْكُم بِمَا صَبَرْتُمْ ۚ فَنِعْمَ عُقْبَى الدَّارِ ﴿٢٤﴾

إِنَّ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ كَانَتْ لَهُمْ جَنَّاتُ الْفِرْدَوْسِ نُزُلًا ﴿١٠٧﴾

Le deuxième type est une répétition identique (ou verbale), dans laquelle les mots sont répétés sans compromettre le sens. Cette répétition peut être liée soit au début, au milieu ou à la fin du verset, ou elle peut être une répétition de l'ensemble du verset.

هَيْهَاتَ هَيْهَاتَ لِمَا تُوعَدُونَ ﴿٣٦﴾

وَيُطَافُ عَلَيْهِم بِآنِيَةٍ مِّن فِضَّةٍ وَأَكْوَابٍ كَانَتْ قَوَارِيرَا ﴿١٥﴾ قَوَارِيرَ مِن فِضَّةٍ قَدَّرُوهَا تَقْدِيرًا ﴿١٦﴾

وَجَاءَ رَبُّكَ وَالْمَلَكُ صَفًّا صَفًّا ﴿٢٢﴾

Ou bien c'est une répétition distincte au niveau du verset ou de l'ensemble du Coran. L'histogramme suivant (Figure 2.3) montre les fréquences des 50 premiers mots les plus utilisés dans le Coran. Cette déclaration montre que les mots de liaison arabes sont les plus fréquents et doivent être pris en compte dans les études nécessitant un traitement de données. Le mot de "الله" occupe la première place et prend plusieurs formes, notamment "الله" apparaît 2 265 fois, "والله" apparaît 240 fois, "بالله" apparaît 139 fois, "تالله" apparaît 8 fois, "فالله" apparaît 6 fois, "اللهم" apparaît 5 fois, "أالله" apparaît 2 fois, "أبالله" et "وتالله" apparaissent une fois.

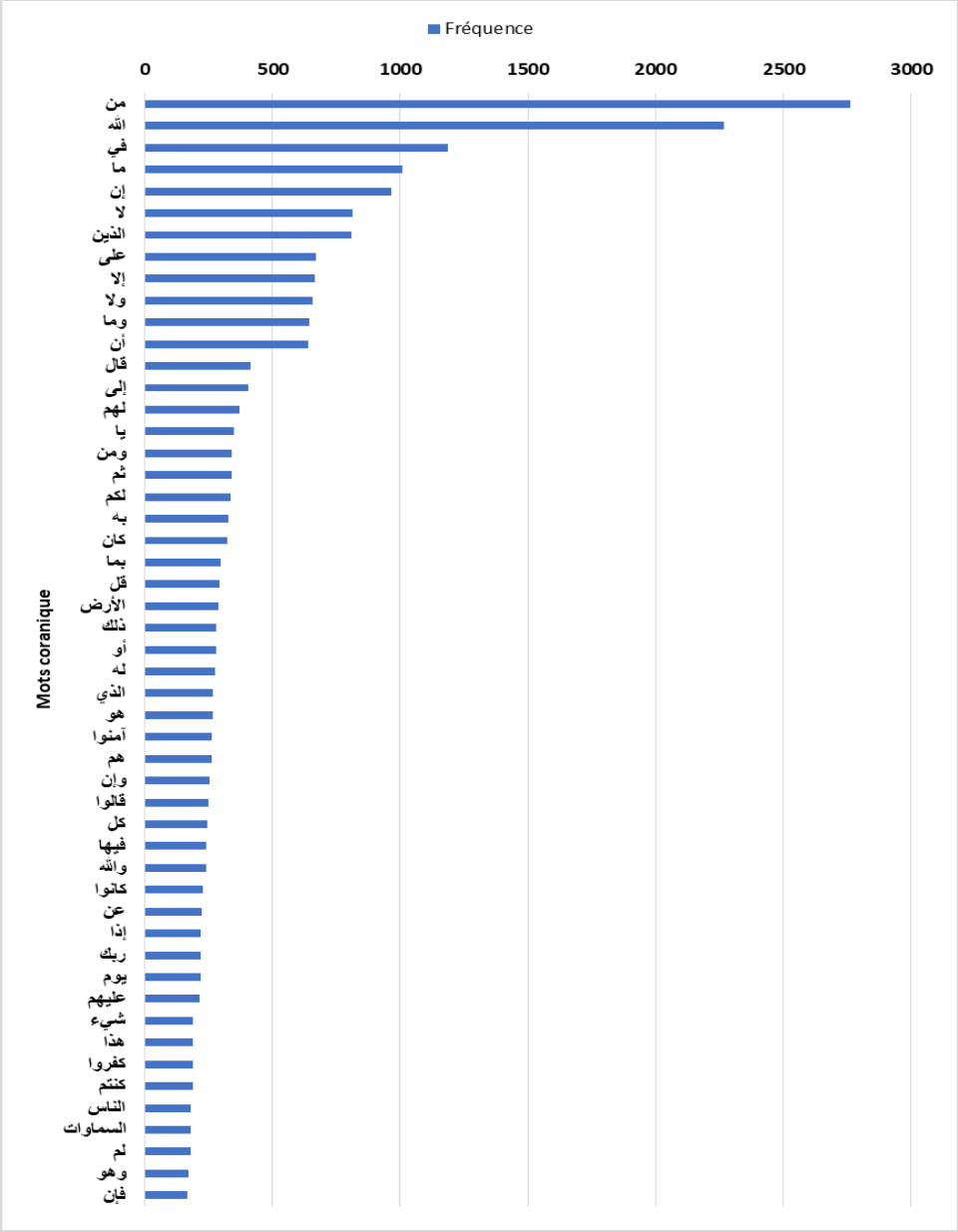


Figure 2. 3: Les 50 mots les plus fréquents dans le Coran

Les mots de liaison sont souvent associés aux mots simples les plus courants, formant des bi-grammes, qui donnent également des taux de répétition élevés dans le texte coranique. L'histogramme ci-dessous (Figure 2.4) répertorie les 50 bi-grammes les plus courants.

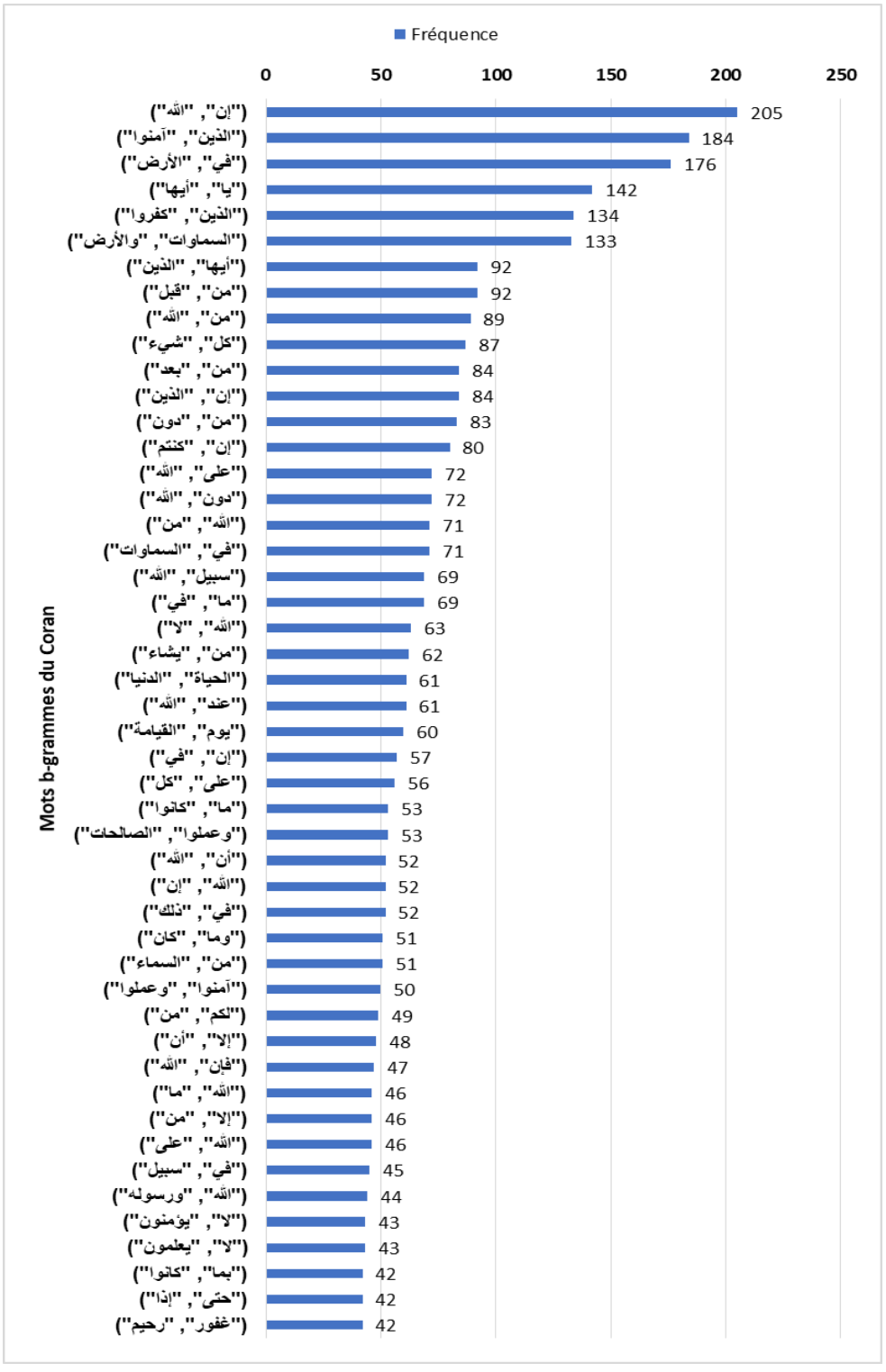


Figure 2. 4:Les 50 bi-grammes les plus fréquents dans le Coran

Concernant la répétition des versets, le graphique suivant (Figure 2.5) montre la répétition du même verset au niveau du même chapitre. De toute évidence, le chapitre n° 26 (Ash-Shu'ara') contient de 5 à 8 répétitions de cinq versets différents. Il est suivi du chapitre n° 37 (As-Saffat), qui contient de 3 à 4 répétitions de 3 versets différents. Le chapitre n° 54 (Al-Qamar) contient également de 3 à 4 répétitions de deux versets différents. La fréquence la plus élevée

se trouve au chapitre n° 55 (Ar-Rahman) avec un taux de 31 répétitions du même verset, tandis que la fréquence la plus faible apparaît au niveau du chapitre n° 56 (Al-Waqi'ah) avec un taux de 2 répétitions. Le reste des versets n'apparaît qu'une seule fois, et il convient de noter que certaines parties des versets sont également répétées dans différentes proportions dans de nombreux chapitres.

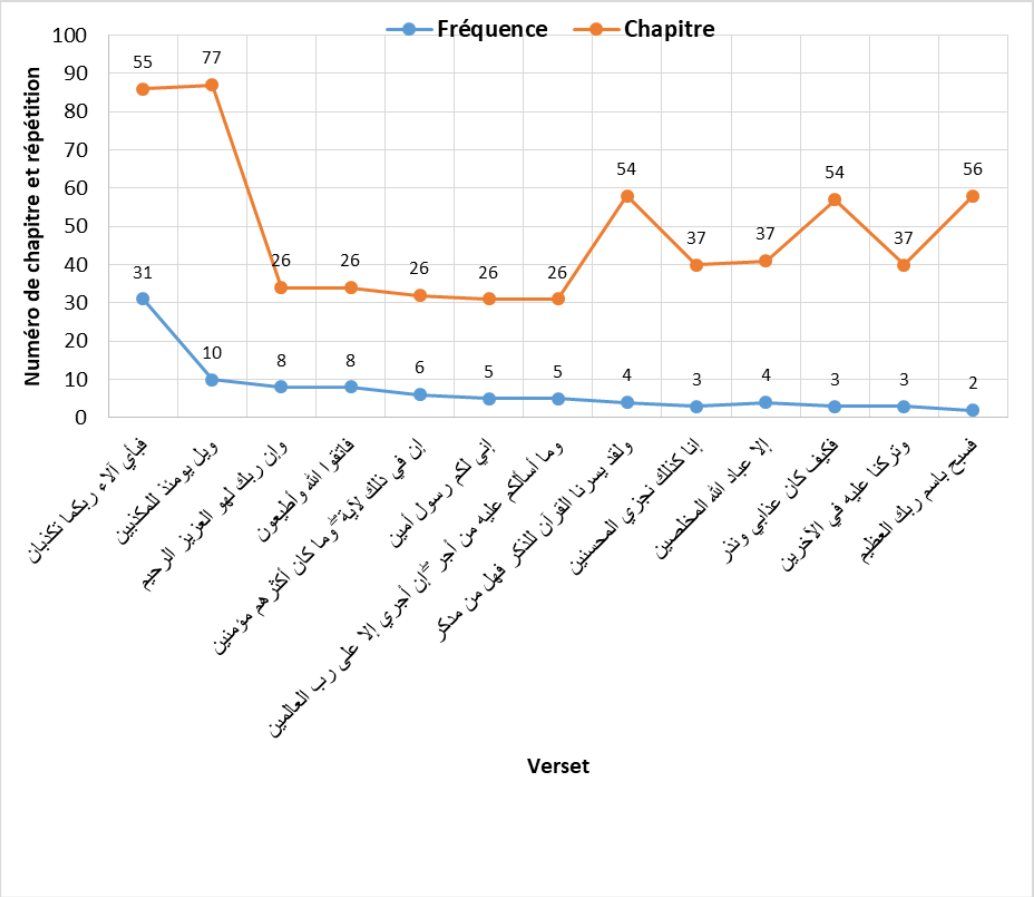


Figure 2. 5:Répétition des versets du Coran

2.2.4. Thèmes du Coran

Le Coran aborde toutes les questions de la vie et traite des règles et des fondements sur lesquels les musulmans comptent dans leur vie. Le graphique suivant (Figure 2.6) représente les thèmes les plus importants du Coran et leurs pourcentages. Les thèmes de l'invisible et de l'isthme arrivent en tête avec 24,4 % des versets du Coran, qui parlent généralement de questions de jugement, de feu, de paradis, de Satan, d'anges, de la création d'Adam et de l'humanité. Les thèmes des récits et de l'histoire apparaissent également à un taux considérable de 22,3 %. Quant au reste des thèmes, ils apparaissent à des taux inférieurs à 10 %. Il convient de noter que chacun de ces thèmes contient un groupe de sous-thèmes avec des pourcentages différents. De plus, un verset peut être trouvé sur plusieurs thèmes.

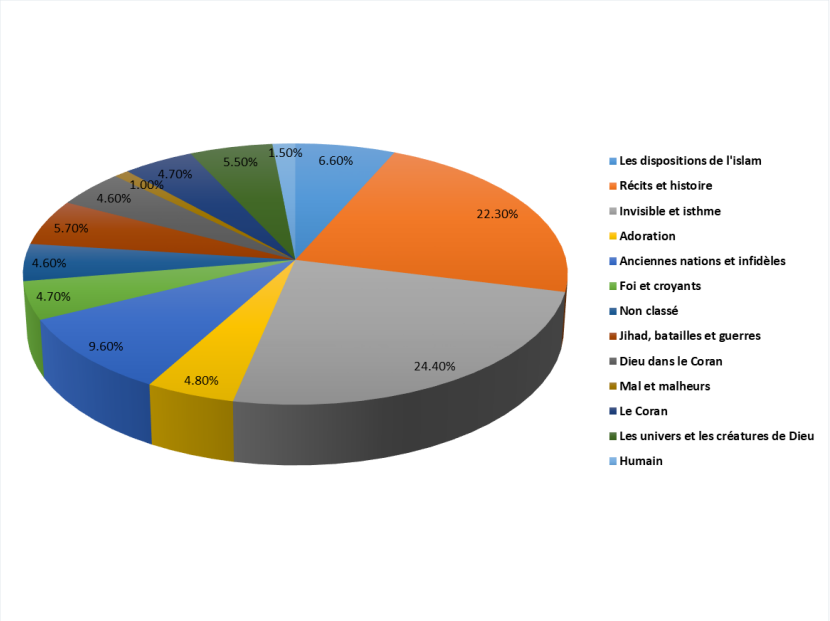


Figure 2. 6:Thèmes du Coran

Ces caractéristiques et statistiques ouvrent un champ de compétition et un défi pour l'utilisation des techniques d'intelligence dans l'extraction des informations cachées du Coran, en tenant compte du fait que seuls quelques versets sont répétés et que les versets ne correspondent pas toujours aux significations.

3. Traitement du langage naturel (NLP)

Le Traitement du Langage Naturel est une partie intégrante de l'Intelligence Artificielle (IA). En tant que simulation de l'esprit humain, cette technologie aide la machine à comprendre et à analyser les langues humaines. Afin d'améliorer les performances des algorithmes de ML, cette technique est utilisée pour représenter l'ensemble de données utilisé pour atteindre l'objectif souhaité. Comme le montre la Figure 7, les étapes de l'analyse des données consistent d'abord à collecter les données puis à les traiter pour qu'elles soient prêtes pour la représentation, qui est la principale entrée pour l'application des algorithmes de ML.

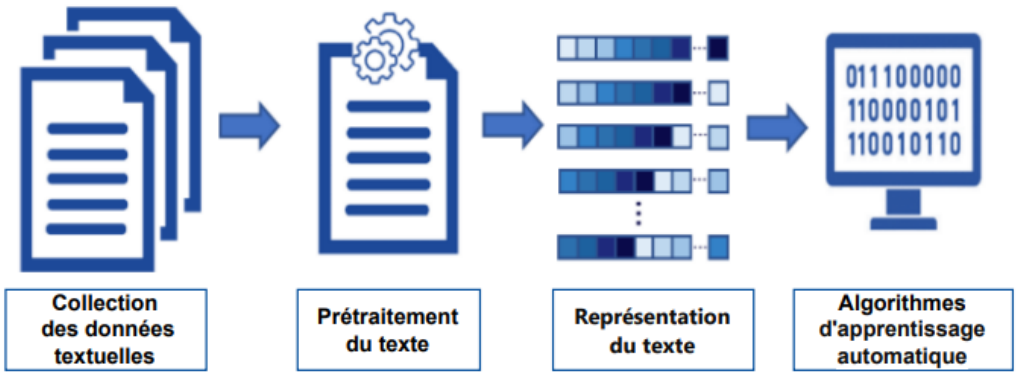


Figure 2. 7:Étapes de ML basées sur le traitement du langage naturel

L'utilisation de ML et du NLP peut être une clé pour affronter les défis posés par la langue arabe dans le futur. L'un de ces défis est le texte coranique que nous abordons dans cette étude. La section suivante présente une étude des différentes techniques de représentation des données qui peuvent être appliquées au texte coranique en arabe.

4. Techniques de représentation des mots du coran

Un mot est une séquence de lettres significative, généralement considérée comme la plus petite unité du langage humain [47] , et le composant principal des phrases et des textes. Pour le Coran, les mots sont au cœur des significations qui enrichissent la compréhension dans l'esprit humain. La question diffère dans les classificateurs et les modèles, où les mots doivent être représentés dans les tâches de NLP pour être compréhensibles par la machine. La représentation des mots est connue comme un processus de transformation mathématique, qui extrait les caractéristiques des mots et les représente sous forme d'espace vectoriel numérique avec une dimension spécifique. Chaque caractéristique de dimension peut même avoir une interprétation sémantique ou grammaticale, c'est pourquoi nous l'appelons une caractéristique de mot [48].

Les types de représentation des mots se divisent en deux familles : une représentation locale et une représentation distribuée [49] . Comme le montre la figure 8, dans une représentation locale, chaque concept (rectangle vertical, rectangle horizontal, ellipse verticale, ellipse horizontale) est représenté de manière unique et séparée du reste. Alors que dans la représentation distribuée, une relation est établie entre tous les concepts, de sorte que chaque concept est représenté sous forme d'un modèle d'activation à travers plusieurs caractéristiques (vertical, horizontal, rectangle, ellipse).

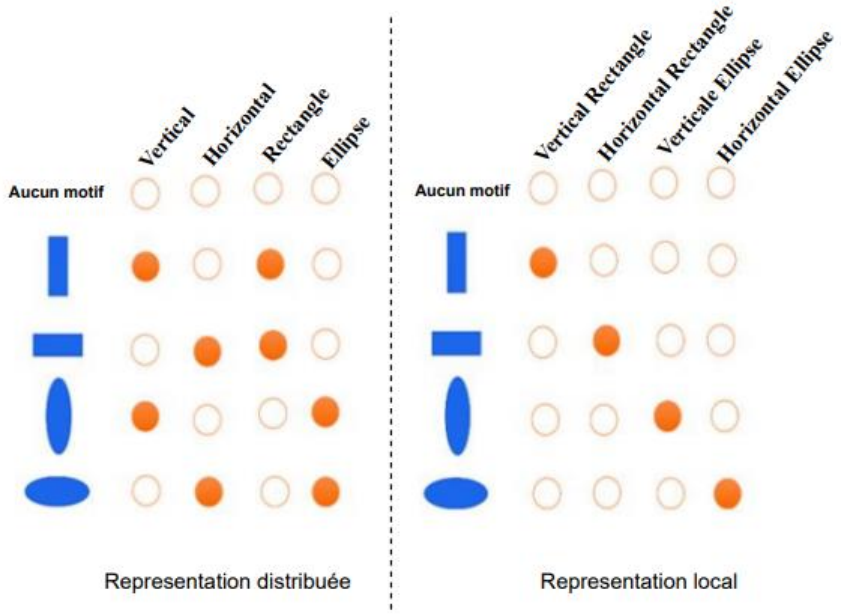


Figure 2. 8:Représentations locales & distribuées [49].

Dans la section suivante, nous étudions les techniques les plus célèbres de représentation qui peuvent être utilisées pour représenter une variété de textes tels que des mots et des versets issus de différents chapitres du Coran.

4.1.Représentation localisée

4.1.1. Modèle du sac de mots (Bag of Word : BOW)

Le modèle BOW est une forme de représentation locale qui compte le nombre de fois où les mots apparaissent dans le sac (le texte) et les représente sous forme de vecteurs avec une dimension égale au contenu du sac [50].

En supposant que le nombre de mots dans le sac soit $\boldsymbol{n}$, nous extrayons d'abord tous les mots $\boldsymbol{w}$ du sac $\boldsymbol{B}$ sous la forme $\boldsymbol{B}$ = { $\boldsymbol{w_1, w_2}$ ... $\boldsymbol{w_3}$ }. Maintenant, pour représenter n'importe quel mot $\boldsymbol{i}$ du sac $\boldsymbol{B}$, nous utilisons l'équation suivante:

$$W_i = \begin{cases} K: Nombre\ de\ fois\ où\ i\ apparaît\ dans\ B \\ 0: Si\ i\ n'apparaît\ pas \end{cases}$$

4.1.2. N-grammes

Le modèle précédent représente chaque mot comme un caractéristique unigramme. En revanche, les N-grammes capturent les mots adjacents et conservent l'ordre des mots. N représente la longueur de la séquence de mots et prend une valeur supérieure ou égale à 2.

4.1.3. TF-IDF

TF-IDF, acronyme pour fréquence du terme et fréquence inverse du document. En plus de décrire l'apparition et l'absence de termes dans les documents, ce modèle calcule la pertinence des mots dans un document particulier [51].

TF [52] mesure le poids du mot dans le document. Le poids est calculé à l'aide de l'équation suivante [53].

$$tf_{t,d} = \frac{f_{t,d}}{n_d}$$

Où :

$\boldsymbol{f_{t,d}}$ est la fréquence du terme $\boldsymbol{t}$ dans les documents $\boldsymbol{d}$.

$\boldsymbol{n_d}$ est le nombre total de mots dans le document $\boldsymbol{d}$.

Souvent, la répétition d'un mot dans tous les documents est insignifiante, comme c'est le cas pour les mots vides dans le Coran (... من ، لا ، أي). Le biais de ces mots doit être évité en leur donnant moins de poids. IDF [54] mesure le poids d'un terme dans un ensemble de documents, la valeur du poids étant inversement proportionnelle à la visibilité accrue du terme dans les documents. L'équation suivante calcule ce poids [53]:

$$idf_t = 1 + \log\left(\frac{D}{C_t}\right)$$

Où :

$\boldsymbol{D}$ est le total des documents.

$\boldsymbol{C_t}$ est le nombre de documents contenant le mot $\boldsymbol{t}$.

La relation suivante capture les poids des mots importants et peu impressionnants à travers tous les documents [53]:

$$W_{t,d} = tf_{t,d} * idf_t$$

Où :

$\boldsymbol{W_{t,d}}$ est le score TF-IDF du terme $\boldsymbol{t}$ dans le document $\boldsymbol{d}$.

4.2.Représentation distribuée

4.2.1. Word Embedding

Cette technique se caractérise par la capacité à aborder l'aspect sémantique du contenu coranique et à formuler le contexte de ses mots. Le terme "word embedding" est apparu au début des années 2003 par Bengio [55], et est le modèle le plus populaire de représentation distribuée des mots. En s'appuyant sur les réseaux neuronaux, ce modèle extrait les similarités syntaxiques et sémantiques entre les mots et les phrases, puis les représente dans des vecteurs spatiaux avec des nombres réels. Cette idée a été adoptée et exploitée dans la création d'outils spéciaux pour les mots et les phrases.

4.2.1.1. Word2vec

Word2vec est un modèle basé sur les réseaux neuronaux spécialisé dans l'apprentissage des liens entre les mots et dans la production de vecteurs avec des relations mathématiques dans un espace multidimensionnel, ce qui indique les similarités sémantiques présentes dans les mots [56]. L'outil Word2Vec fournit des algorithmes pour l'entraînement de word embeddings.

Comme le montre la figure 2.9, le premier algorithme est le Continuous Bag of Words (CBOW), cet algorithme est basé sur l'architecture linguistique FeedForward Neural Net Language qui se compose de couches d'entrée, de projection, cachée et de sortie. Le degré de calcul augmente en complexité entre les couches de projection et cachée. Dans le CBOW, la couche cachée non linéaire est supprimée et la couche de projection est partagée avec tous les mots, de sorte que le mot actuel est prédit en fonction du contexte des mots restants [56]. Le deuxième algorithme est le SkipGram (SG), qui a une architecture similaire à celle du CBOW, mais dans lequel le contraire se produit, où les mots environnants sont prédits en regardant le mot actuel [56].

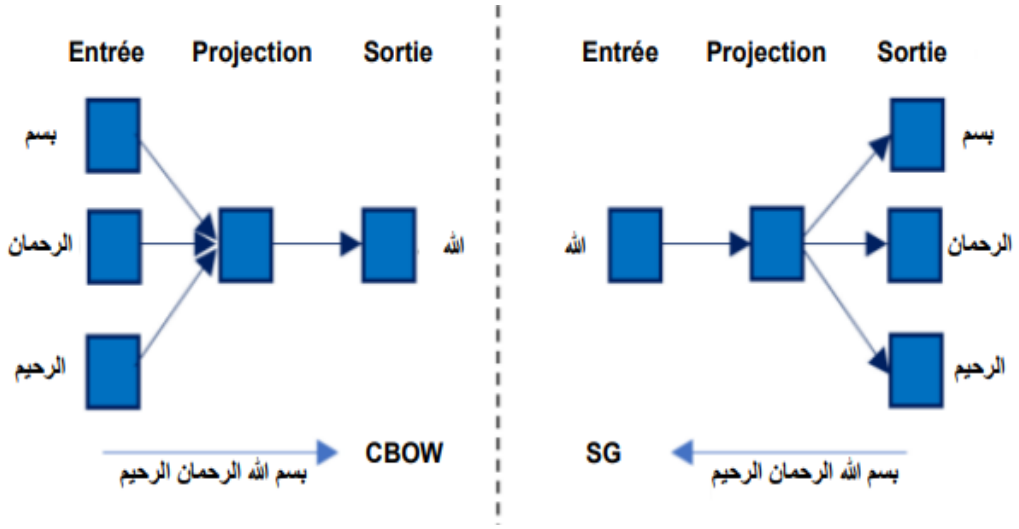


Figure 2. 9:Algorithmes Word2vec [57].

4.2.1.2. Doc2vec

Doc2vec s'appuie sur word2vec pour représenter des documents (versets, chapitres, parties ou sections). Contrairement aux mots, il n'y a pas de structure logique dans les documents, donc un identifiant unique est ajouté à chaque paragraphe en utilisant word2vec. Comme pour word2vec, il existe deux algorithmes de doc2vec (Figure 2.10) : le premier appelé Mémoire distribuée de Vecteur de Paragraphe (Distributed Memory version of Paragraph Vector:PV-DM) est similaire à la méthode de sac continu de mots (CBOW) dans word2vec. Et le second appelé Sac de mots distribué de Vecteur de Paragraphe (Distributed Bag of Words version of Paragraph Vector :PV-DBOW) est similaire à (SG) dans word2vec [57].

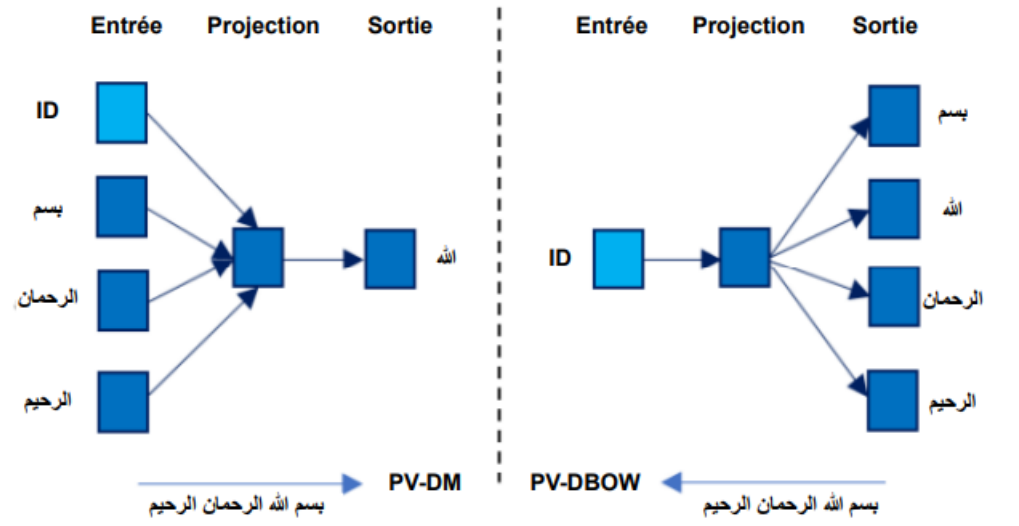


Figure 2. 10:Algorithmes Doc2vec [57].

5. Evaluation

Il est nécessaire d'évaluer les modèles de représentation des mots pour comprendre l'efficacité de chaque méthode dans les applications du monde réel intéressées par le Coran. Les critères d'évaluation sont divisés en normes externes qui se soucient de la structure des mots et en normes internes qui se soucient du sens et des connotations des mots. Dans la représentation locale, chaque mot est représenté comme une seule caractéristique en très peu de temps avec une densité de mémoire minimale et sans besoin de données externes, mais ni les règles de phrase ni l'ordre des mots (à l'exception des N-grammes) ne sont préservés. De plus, la longueur du vecteur augmente à mesure que le nombre de mots des phrases augmente, et le nombre de zéros dans les vecteurs augmente également, ce qui donne une représentation éparse qui n'est pas souhaitable. La représentation distribuée reflète l'intérieur des mots et leurs significations, et maintient l'interconnexion des mots entre eux. Cette représentation nécessite de grandes données externes et une mémoire plus intensive. Il est difficile de décrire une solution idéale pour toutes les applications du Saint Coran nécessitant des tâches de programmation linguistique.

Les évaluations sont limitées à la tâche en cours, donc un petit changement dans la représentation des données change significativement les résultats.

6. Conclusion

Le texte sacré du Coran n'a pas encore reçu une attention considérable de la part des chercheurs habitués à l'intelligence artificielle. Dans cet article, nous avons présenté une clé pour ouvrir le champ d'application des sous-ensembles de l'intelligence artificielle afin de découvrir des informations sur le Coran. Les statistiques fournies seront utiles pour diviser et organiser les ensembles de données, et le choix de la méthode de représentation des données sera basé sur la nature de la recherche. Dans nos prochains chapitres, nous utiliserons des techniques de représentation des mots avec des réseaux neuronaux pour identifier et valider des citations coraniques afin de préserver l'intégrité du Coran.

Chapitre 3

Identification du contenu coranique arabe à l'aide de Deep Learning et Word Embeddings

1. Introduction

Le Saint Coran numérique est l'un des livres les plus sacrés parmi 1,3 milliard de musulmans à travers le monde. Le Saint Coran se compose de 114 chapitres (sourates) de longueurs variables et est disponible dans différents styles d'écriture tels que l'Uthmani, non-Uthmani, etc. [58]. Le Saint Coran a été révélé en langue arabe et a été préservé de toute distorsion et corruption, contrairement au reste des livres célestes. Par conséquent, le rôle de chaque musulman est de travailler à préserver l'authenticité et l'intégrité de ce noble Livre [6] [3].

L'identification des versets/sourates coraniques peut être définie comme la distinction des mots du texte correspondant au contenu du Saint Coran, tandis que l'authentification consiste à s'assurer qu'ils sont écrits exactement comme ils le sont dans toutes les copies du Saint Coran [41].

Avec le développement des applications Internet, les versets coraniques sont de plus en plus cités sur des blogs, des forums et des sites de médias sociaux. Les versets sont copiés à partir de sites Web non fiables, il est donc difficile pour le lecteur de distinguer entre les versets coraniques authentiques et le texte ordinaire [59] [41]. Le problème est plus difficile pour les locuteurs non natifs de la langue arabe qui ne sont pas familiers avec la langue et peuvent être facilement confus. Pour les locuteurs natifs, bien qu'il soit plus facile de distinguer les versets écrits en script uthmani, les versets diacritiques et les versets écrits entre deux crochets, le problème réside dans la distinction des versets non diacritiques où il devient difficile de déterminer le début et la fin de chaque verset ou sourate incorporé dans un texte en arabe.

Les techniques de traitement de texte classées en dehors du domaine du Traitement du Langage Naturel (NLP) donnent des résultats moins qualifiés, surtout avec les textes coraniques arabes, car elles nécessitent du temps et une analyse approfondie de la version originale du texte coranique [60]. Il est donc très important de développer des systèmes intelligents qui peuvent aider à analyser ce type de ressource en ligne.

Le deep learning (DL) est une branche de l'intelligence artificielle, l'une de ses études les plus importantes étant l'application de réseaux neuronaux profonds pour résoudre des problèmes dans le domaine de la linguistique informatique et du NLP [61]. Le DL repose sur l'entraînement progressif des couches sur des représentations de données de plus en plus complexes en même temps. La puissance du DL réside dans l'élimination de la phase d'ingénierie des caractéristiques, qui est l'un des principaux goulets d'étranglement dans les pipelines de ML [61]. De plus, les classificateurs de DL surpassent les classificateurs traditionnels en réalisant davantage de gains lorsqu'ils traitent des textes sensibles à la séquence. Par conséquent, il est considéré comme une excellente solution pour traiter les textes coraniques.

À l'heure actuelle, il n'existe aucune application automatisée capable de reconnaître les versets coraniques en utilisant des techniques de DL.

Par conséquent, la principale contribution de ce chapitre est de proposer une nouvelle approche basée sur le DL et les techniques de plongements de mots (WE) pour construire des modèles capables de classer automatiquement le contenu des textes coraniques et arabes.

2. Travaux Connexes

Il n'existe pas beaucoup de travaux disponibles dans le domaine de l'authentification du Coran numérique. Les travaux dans [3] [62] présentent une revue exhaustive des approches d'authentification de pointe du Coran numérique et des Hadiths. De même, [6] [63] présente les récents progrès dans l'authentification du contenu numérique avec un accent principal sur le contenu arabe disponible respectivement en formats texte et image.

Quelques-uns des travaux connexes incluent les travaux de [64] [5] [59] [65] [66] [67] [68] où les auteurs ont proposé des approches basées sur la correspondance exacte des chaînes [69] pour détecter les versets coraniques uthmani et non-uthmani. Kamaruddin et al. [70] ont proposé une approche basée sur le tatouage numérique pour protéger et détecter les versets coraniques.

Dans l'un des travaux les plus récents, Almazrooie et al. [71] ont proposé une fonction de hachage cryptographique pour préserver l'intégrité du Coran numérique. Toutes les approches mentionnées ci-dessus se sont concentrées sur l'authentification et la préservation de l'intégrité du contenu du Coran numérique. Aucun de ces travaux n'a abordé le problème de la classification du contenu coranique par rapport au texte ordinaire en utilisant des techniques de DL.

Bien que quelques travaux se soient concentrés sur des problèmes de classification concernant le contenu arabe, comme le travail de [72] [73] [74] .

Cependant, la portée de ces travaux est limitée et s'est concentrée sur différents aspects de la langue arabe tels que la détection des abus, la détection de l'humeur et les aspects de fouille d'opinions.

3. Défis et Motivations

Les sites de réseaux sociaux sont les plateformes les plus populaires pour partager des versets coraniques. Les versets coraniques peuvent être identifiés en se basant sur certains mots de début tels que "قال الله تعالى" ou des mots de fin tels que "صدق الله العظيم", ou en déterminant le pourcentage de diacritiques, et ainsi de suite. Ces techniques sont considérées comme traditionnelles et nécessitent beaucoup de temps et d'efforts, car elles nécessitent toujours une référence pour approbation. Il n'est pas difficile de documenter les versets coraniques, mais cela nécessite une identification manuelle préalable, ce qui rend le processus fastidieux et non instantané. Malheureusement, ces formes sont considérées davantage comme des moteurs de recherche que comme des systèmes d'authentification [40]. La puissance du DL dans le traitement automatique du texte apparaît dans la compréhension et la classification des textes de manière automatisée et en temps réel. La formation exhaustive est une caractéristique clé

du DL qui en fait un outil puissant pour le NLP. Il existe plusieurs tâches dans lesquelles les modèles DL sont nettement supérieurs au reste des modèles précédents. Tout d'abord, le DL peut extraire automatiquement des caractéristiques car il peut gérer même un nombre illimité de caractéristiques. Il réside également la capacité à travailler avec une connaissance insuffisante, dans notre cas, nous entraînons les modèles sur seulement un certain pourcentage du Coran et le reste est utilisé pour la vérification. Enfin, certains réseaux neuronaux sont dédiés au traitement de textes en général, et d'autres réseaux neuronaux profonds sont dédiés au traitement de textes séquentiels en particulier [75]. De nombreux algorithmes de classification classiques peuvent faire un assez bon travail. Si nous examinons d'autres techniques de classification non basées sur des réseaux neuronaux, elles sont entraînées sur plusieurs mots en tant qu'entrées séparées qui ne sont rien d'autre qu'un mot sans réel sens en tant que phrase, et lors de la prédiction de la classe, elle donnera le résultat en fonction des statistiques et non en fonction du sens. Une des bonnes raisons d'utiliser des modèles DL est qu'ils sont efficaces pour mémoriser des informations importantes. Nous pouvons utiliser une chaîne de mots multiples pour savoir à quelle catégorie elle appartient. Cela est très utile lors du travail avec le NLP. Si nous utilisons des couches d'encastrement et de codage appropriées dans les modèles DL, le modèle connaîtra le sens réel de la chaîne d'entrée et donnera la classe de sortie la plus précise.

4. Méthodologie Proposée

La Figure 3.1 illustre le schéma globale proposée pour identifier les versets coraniques arabes, composée de plusieurs étapes. Les données ont d'abord été collectées à partir des deux principales sources de tanzil.net[5] [45] et du Corpus d'Apprenant de l'Arabe (ALC[6]) [76]. Ensuite, les données collectées ont été soumises à l'étape de prétraitement. L'étape de représentation du texte reçoit les ensembles de données afin de représenter les mots sous forme de vecteurs dimensionnels. Afin de préserver les relations entre les mots dans le texte, nous avons choisi la technique WE pour extraire des caractéristiques numériques du texte. Nous avons entraîné les modèles de Réseau Neuronal Convolutif (CNN) et de Réseau Neuronal Convolutif avec Mémoire à Court et Long Terme (CNN-LSTM) en utilisant les données d'entraînement. Enfin, nous testons les modèles en utilisant les vecteurs de l'ensemble de données de test.

[5] Tanzil.net est une initiative coranique fondée en 2007 dans le but de générer un texte coranique Unicode soigneusement examiné à utiliser dans les sites web et applications coraniques.

[6] Arabiclearnercorpus.com (ALC : Arabic Learner Corpus) est un projet présentant une collection de documents écrits et parlés provenant d'apprenants de l'arabe en Arabie saoudite.

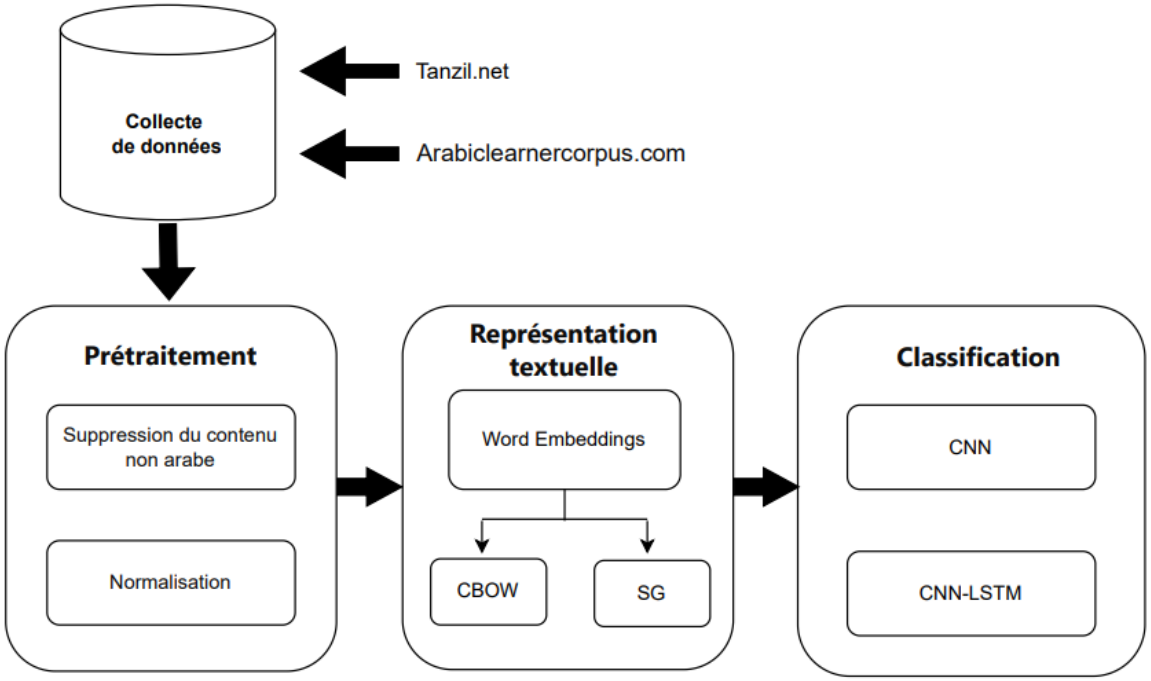


Figure 3. 1: Structure de l'approche proposée.

4.1. Collecte de l'ensemble de données

Pour entraîner l'algorithme, nous avons exploré deux ensembles de données populaires. En ce qui concerne le Saint Coran, le texte coranique Unicode vérifié, qui sert de source fiable fournie par tanzil.net [45] , a été utilisé et nommé la classe "Texte de coran". Sur tanzil.net, les textes coraniques sont généralement représentés dans deux formats généraux : le script Uthmani et le script simple. Les autres formats disponibles comprennent le format Texte (TEXT), le format Langage de Requête Structurée (SQL) et le format Langage de Balisage Extensible (XML). La figure 3.2 montre le verset "Al Fatiha" au format XML, structuré sous forme de nœuds arborescents, où le nœud "aya" est l'enfant du nœud "surah", qui est l'enfant du nœud racine du Coran.

```
<quran>
        <sura index="1" name="الفاتحة">
        <aya index="1" text="بسم الله الرحمن الرحيم" />
        <aya index="2" text="الحمد لله رب العالمين" />
        <aya index="3" text="الرحمن الرحيم" />
        <aya index="4" text="مالك يوم الدين" />
        <aya index="5" text="إياك نعبد وإياك نستعين" />
        <aya index="6" text="اهدنا الصراط المستقيم" />
        <aya index="7" text="صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين" />
    </sura>
...
</quran>
```

Figure 3. 2: Texte coranique simple au format XML de tanzil.net.

Pour les textes arabes ordinaires, nous avons exploré le site web du Corpus des Apprenants en Arabe (ALC), qui propose une collection de documents écrits et parlés produits par des apprenants de l'arabe en Arabie Saoudite. Ces données ont été utilisées dans notre approche sous la classe "Texte Arabe". La Figure 3.3 montre un exemple de ces données au format XML, où les nœuds "titre" et "corps du texte" sont tous deux enfants du nœud racine du texte utilisé dans nos données.

<ALC>
<doc ID="S001_T1_M_Pre_NNAS_W_C">
<header>
…
</header>
<text>
<title> رحلة إلى أبها </title>
<text_body>

قمنا برحلة إلى أبها، كنا تسعة أشخاص، كان موعد تجمعنا في ملعلب إسكان الطلاب الساعة 9 15، وبعد ذلك انطلقنا إلى المطار الملك خالد بالرياض، عند وصولنا في المطار، ركبنا بطائر إلى أبها، لما وصلنا في أبها جاء رجلان للاستقبالنا، ثم بعد السلام نقلوانا بحافلة في طريق إلى الفندق أبها فيها مناظر جميلة، شهادة القرون في طريق والجبال، وذهبنا لذيارت جامعة الملك حالد الجامعة مشاء الله جميل جداً، أشكر عمادة بقيام هذه رحلة جزاهم الله خيراً

</text_body>
</text>

Figure 3. 3:Texte arabe simple au format XML de ALC

4.2. Prétraitement

Le jeu de données coranique comprend 6236 documents étiquetés comme "Texte des versets", et le jeu de données arabes se compose de 1585 documents arabes étiquetés comme "Texte arabe".

Pour la normalisation des ensembles de données, les étapes de prétraitement suivantes ont été effectuées :

- Organisation du deuxième jeu de données : les paragraphes du deuxième jeu de données étaient très longs, dans cette étape, nous les avons divisés en phrases appropriées en fonction des signes de ponctuation. À la fin, nous avons obtenu 6375 phrases utiles.

- Suppression de la ponctuation et des dialectes : les signes de ponctuation et les Harakat, tels que le Fatha /َ/, le Dammah /ُ/, le Kasrah /ِ/, le Sukoon /ْ/, le Shaddah /ّ/ et le Tanween /ً ٌ ٍ / ont été filtrés.

- Suppression des caractères non arabes : filtrer le contenu non arabes est une étape importante lors du traitement de données numériques provenant du web car il existe une autre langue dont les lettres se chevauchent avec l'alphabet arabe, tels que l'ourdou et le persan.

- Suppression des symboles spéciaux et des liens : dans cette étape, tous les emojis, émoticônes, dates, heures et URL ont été supprimés.

- Suppression de la Kasheeda : la kasheeda est un type de caractère étendu, par exemple, "Bismi Allahi" avec et sans kasheeda peut ressembler à ce qui suit:

Mot (en Phonétique)	Normal	Kasheeda
Bismi	بسم	بســـــــــم
Allahi	الله	اللـــــــــه

- Normalisation des Alefs, Alef Maksura et Tah Marbutah : Remplacer le Hamza / أ et إ / et Maddah / آ / par un simple Alif / ا /, tout en remplaçant la lettre Alif Couteau / ى / par un Yā / ي /, et en remplaçant le Tāʾ marbūṭah / ة / par un hāʾ / ه /. Cela est dû au fait que les utilisateurs d'Internet abusent souvent de ces caractères en raison de leur similitude [77] et de la difficulté à en trouver certains sur le clavier.

4.3. Représentation textuelle

Word Embeddings (WE) [55] est une représentation distribuée du vocabulaire d'un document dans un espace vectoriel. Une des techniques courantes pour construire une telle intégration est word2vec [78] [56] [79]. Il prend en entrée des mots individuels du corpus et produit un vecteur de valeurs réelles dans un espace de dimensions élevées. Cette technique capture le contexte, le sens ainsi que la structure des mots, ce qui permet de mapper les mots similaires dans des espaces vectoriels géométriquement proches. L'outil Word2Vec est considéré comme une collection de deux architectures de réseaux neuronaux différentes : Continuous Bag of Words (CBOW) et Skip-Gram (SG). Comme illustré dans la Figure 3.4, ces deux architectures sont des réseaux neuronaux avec une couche cachée destinée à représenter les vecteurs de mots. La différence réside dans le fait que CBOW prend le contexte en entrée $(w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2})$ et tente de produire le mot cible (w_t) avec une grande précision grammaticale, tandis que SG fait l'inverse : il prend le mot cible en entrée (w_t) et essaie de produire un contexte approprié pour ce mot $(w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2})$ avec une plus grande précision sémantique [56].

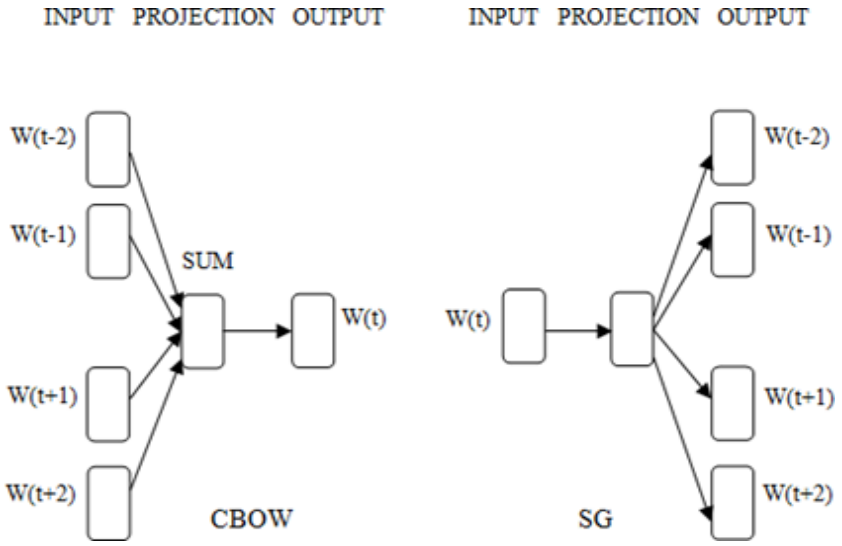


Figure 3. 4: Architectures CBOW et Skip-Gram

La qualité de WE peut être affectée par la dimension de l'espace vectoriel, qui varie de cent à mille. Plus le vecteur est grand, plus la précision et la complexité du modèle en termes de temps de calcul sont importantes. Dans notre étude, nous avons utilisé plusieurs modèles de dimensions 100 et 300 fournis dans la deuxième version d'AraVec par Soliman et al. [80]. Les modèles AraVec couvrent trois domaines de contenu arabes différents : Wikipédia, Twitter et le World Wide Web, avec les deux architectures WE : CBOW et SG. Ces modèles sont résumés dans le Tableau 3.1.

Tableau 3. 1:Description des différents modèles AraVec [80].

Modèle	**Documents**	**Vocabulaires**	**Dimension**
Twitter-CBOW	66,900,000	331,679	300
Twitter-SG	66,900,000	331,679	300
Twitter-CBOW	66,900,000	331,679	100
Twitter-SG	66,900,000	331,679	100
Wikipedia-CBOW	1,800,000	162,516	300
Wikipedia-SG	1,800,000	162,516	300
Wikipedia-CBOW	1,800,000	162,516	100
Wikipedia-SG	1,800,000	162,516	100
Www-CBOW	132,750,000	234,961	300
Www-SG	132,750,000	234,961	300
Www-CBOW	132,750,000	234,961	100
Www-SG	132,750,000	234,961	100

4.4. Représentation de Word2vec

En NLP, le traitement des caractéristiques textuelles nécessite de coder chaque caractéristique comme une dimension unique. Dans les attributs basés sur les mots tels que word2vec, chaque mot est inclus dans l'espace de dimension et représenté sous forme de vecteur dans cet espace. Supposons qu'un texte en arabe A se compose de N mots, $A = (w_1, w_2, \dots, w_N)$ Tout d'abord, chaque mot W_i du texte est représenté par le vecteur de caractéristiques V_i de dimension d, et donc tous les mots du texte sont représentés par la matrice de caractéristiques $d * N$, où chaque élément de la matrice correspond au mot correspondant dans le texte. Afin de représenter le texte avec un vecteur d'attributs unidimensionnel pour servir de seule entrée aux algorithmes de DL, nous appliquons l'équation suivante :

$$V(A) = \frac{\sum_{i=1}^{N} v_i}{N}$$

4.5. Classification

Les réseaux neuronaux (NN) ont été utilisés dans le traitement du langage dans diverses recherches. Le NN simple se compose d'un ensemble connecté de neurones. Chacun produit une série d'activations de valeurs réelles [81] . Le DL est un NN de grande taille composé de multiples couches de traitement pour apprendre des représentations de données avec plusieurs niveaux d'abstraction [81]. Dans cette étude, deux classificateurs DL ont été considérés pour les expériences : le réseau neuronal convolutif (CNN) et le réseau neuronal récurrent (RNN). Les CNN [82] sont des réseaux neuronaux à propagation avant (FNN), cette architecture utilise efficacement des couches avec des filtres de convolution appliqués à des caractéristiques locales [82]. Alors que les RNN diffèrent des FFNs en ce qu'ils ont une mémoire [83]. Un type spécial de réseaux neuronaux récurrents (RNN) est LSTM [83], composé d'une cellule de mémoire, d'une porte d'entrée, d'une porte de sortie et d'une porte d'oubli. La cellule stocke une valeur pour des périodes longues ou courtes. Cela est réalisé en utilisant la fonction d'activation pour la cellule de mémoire.Dans des recherches récentes, les deux réseaux ont été consacrés à une variété de problèmes de ML. La recherche actuelle utilise un CNN unidimensionnel (CNN 1D), où le noyau se déplace dans une seule direction et LSTM pour distinguer le verset coranique dans le texte arabe. Ce qui suit est une description complète des architectures réseau pour les deux sous-modèles de notre système.

4.5.1. Modèle CNN

Pour le modèle CNN, chaque entrée doit être représentée sous forme de matrice (dimension d'incorporation × taille du vocabulaire). Nous représentons la dimension d'incorporation avec la dimension du modèle AraVec utilisé (300 ou 100), et nous représentons la taille du vocabulaire avec la longueur du verset le plus long dans le Saint Coran "Ayat al-Din : 129 mots". Par conséquent, les dimensions de la matrice sont soit 300×129 soit 100×129. Un rembourrage avec des zéros est utilisé pour les phrases contenant moins de 129 mots. Ensuite,

cela passe dans la couche Conv1D, qui se compose de 128 filtres pour obtenir la profondeur maximale compte tenu des possibilités de calcul disponibles, suivie d'une couche de max-pooling avec une taille de pool de 5 et des dropouts de 0,5 pour éviter le surajustement. Ensuite, nous prenons la sortie du max-pooling, la mettons à plat et la passons dans le réseau neuronal entièrement connecté avec une couche cachée dense de 128 unités pour réduire les erreurs de classification intra-classe. La fonction d'activation est ReLU [84]. La couche de sortie se compose de 2 unités softmax pour prédire la classe du texte. Pour l'optimisation, nous utilisons l'optimiseur de descente de gradient stochastique (SGD) [85]. Nous avons sauvegardé les poids de prédiction de sortie pour prédire les ensembles de données de test. La fonction d'ajustement utilise un nombre d'époques=50, Batch size =10, validation split =20.

4.5.2. Modèle CNN-LSTM

L'architecture CNN-LSTM combine à la fois des couches convolutionnelles pour apprendre des caractéristiques locales et des couches LSTM pour apprendre des caractéristiques globales. Comme le modèle CNN, notre modèle CNN-LSTM utilise une couche d'incorporation, une couche de convolution 1D de 128 filtres et une taille de noyau de 5 suivie d'un max-pooling avec une taille de pool de 5, plus une couche LSTM avec 300 ou 100 unités (selon la dimension du modèle WE utilisé), suivie d'une couche dense de 128 unités, et enfin, une fonction d'activation sigmoïde appliquée à la sortie de la LSTM.

5. Expérimentations

5.1. Mesures d'évaluation

Pour évaluer les modèles de réseaux neuronaux proposés, nous nous appuyons sur le calcul de l'Accuracy, de la Precision (Précision), du Recall (Rappel) et du F1 score de chaque modèle sur les ensembles de données de test. Ces mesures vont de 0 % à 100 % et sont calculées comme suit pour chacune des catégories positive et négative :

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

$$Accuracy = \frac{TP + TN}{P + N} * 100$$

Où signifie Positif (P); signifie Négatif (N); signifie Vrai Positif (VP); signifie Faux Positif (FP); signifie Vrai Négatif (VN) et signifie Faux Négatif (FN).

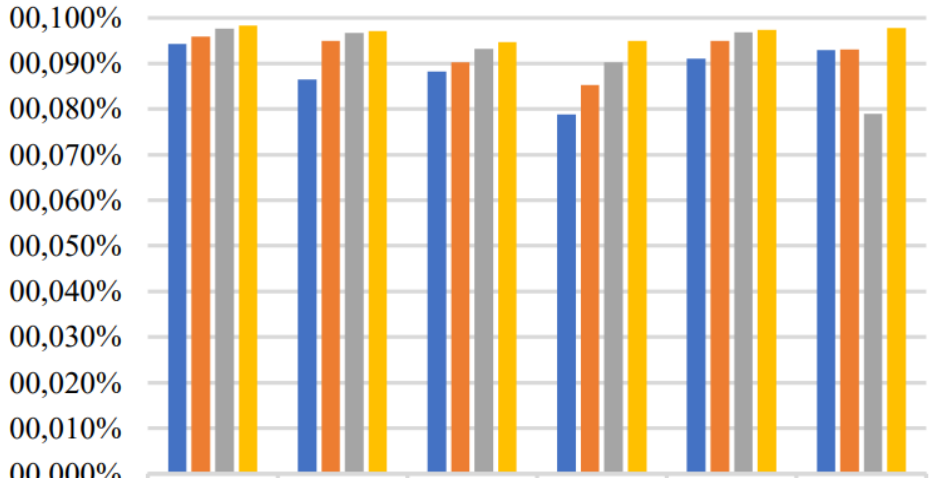
5.2. Résultats expérimentaux

Le jeu de données est divisé en 3 ensembles : 60 % pour l'entraînement, 20 % pour les tests et 20 % pour la validation. Le Tableau 3.2 résume le nombre d'éléments dans chaque ensemble.

Tableau 3. 2: Division du jeu de données

Class	Train	Validation	Test	Total
Texte de Coran	3991	998	1247	6236
Text arabe	4080	1020	1275	6375
Total	8071	2018	2522	12611

L'ensemble d'entraînement des deux modèles a atteint une précision de 100 % dans la plupart des cas. Le Tableau 3.3 présente les résultats de l'ensemble de test en termes de précision, de rappel et de score F1, et la Figure 3.5 montre les résultats de l'ensemble de test en termes de précision considérée dans cette étude, où 0 indique la classe "Texte arabe" et 1 indique la classe "Texte de Coran".



	Twitter-CBOW	Twitter-SG	Wiki-CBOW	Wiki-SG	Www-CBOW	Www-SG
CNN 100	00,094%	00,086%	00,088%	00,079%	00,091%	00,093%
CNN 300	00,096%	00,095%	00,090%	00,085%	00,095%	00,093%
CNN-LSTM 100	00,098%	00,097%	00,093%	00,090%	00,097%	00,079%
CNN-LSTM 300	00,098%	00,097%	00,095%	00,095%	00,097%	00,098%

Figure 3. 5: Résultats de précision des tests

Pour le premier modèle, l'architecture CNN utilisant la technique CBOW, apprise à partir d'un corpus collecté sur Twitter avec une dimension de 300, atteint la meilleure précision sans biais (93,87 % pour la classe '0' et 94,94 % pour la classe '1') ainsi que le meilleur score F1 (94,47 % pour la classe '0' et 94,32 % pour la classe '1'). En termes de rappel, la même architecture obtient les meilleurs résultats (95,07 % pour la classe '0' et 93,71 % pour

la classe '1').

Pour le deuxième modèle, l'architecture hybride CNN-LSTM utilisant la technique CBOW, apprise à partir d'un corpus collecté sur Twitter avec une dimension de 300, atteint la meilleure précision sans biais (98,07 % pour la classe '0' et 97,66 % pour la classe '1') ainsi que le meilleur score F1 (97,83 % pour la classe '0' et 97,90 % pour la classe '1'). En termes de rappel, la même architecture obtient les meilleurs résultats (97,58 % pour la classe '0' et 98,14 % pour la classe '1').

Le Tableau 3.4 présente la précision, le rappel et le score F1 moyens pour le meilleur résultat dans les architectures CNN et CNN-LSTM. La combinaison du CNN avec le LSTM améliore l'exactitude du CNN seul de 2,46 %, la précision de 3,46 %, le rappel et le score F1 de 3,47 %. Cela signifie que cette combinaison peut distinguer les versets et les textes arabes de 3,46 % de plus que le CNN seul. Par conséquent, le modèle hybride CNN-LSTM surpasse le modèle CNN. Les résultats de la comparaison précédente tournent autour de deux points. Premièrement, l'utilisation de modèles WE pré-entraînés dans une dimension fixe donne de meilleurs résultats que les modèles avec une représentation vectorielle classique. En outre, la combinaison des couches CNN et LSTM permet de capturer plus de caractéristiques, améliorant ainsi toutes les mesures et indiquant une meilleure reconnaissance des versets et des textes arabes.

Tableau 3. 3: Résultats des modèles proposés

Modèle	Technique	Dimension	Classe	Precision	Recall	F1 Score
CNN	Twitter-CBOW	300	0	93.87	95.07	94.47
			1	94.94	93.71	94.32
		100	0	90.13	96.25	93.09
			1	95.92	89.32	92.50
	Twitter-SG	300	0	93.28	97.04	95.12
			1	96.87	92.91	94.85
		100	0	76.88	99.50	86.74
			1	99.28	69.66	81.87
	Wiki-CBOW	300	0	91.43	86.12	88.69
			1	86.71	91.81	89.19
		100	0	89.23	84.05	86.56
			1	84.73	89.72	87.15
	Wiki-SG	300	0	99.56	68.11	80.88
			1	75.51	99.70	85.93
		100	0	99.82	56.00	71.75
			1	69.12	99.90	81.71
	Www-CBOW	300	0	93.91	94.19	94.05
			1	94.09	93.81	93.95
		100	0	93.67	88.87	91.21
			1	89.27	93.91	91.53
	Www-SG	300	0	86.88	98.42	92.29
			1	98.15	84.93	91.06
		100	0	86.56	98.32	92.07
			1	98.03	84.53	90.78

CNN-LSTM	**Twitter-CBOW**	300	0	98.07	97.58	97.83
			1	97.66	98.14	97.90
		100	0	98.22	95.76	96.98
			1	95.61	98.15	96.87
	Twitter-SG	300	0	97.29	98.67	97.97
			1	98.52	96.98	97.74
		100	0	95.79	99.24	97.48
			1	99.13	95.21	97.13
	Wiki-CBOW	300	0	99.08	92.04	95.43
			1	91.89	99.06	95.53
		100	0	99.26	89.01	93.85
			1	89.16	99.27	93.94
	Wiki-SG	300	0	98.02	94.03	95.98
			1	93.73	97.92	95.78
		100	0	99.65	81.53	89.68
			1	83.10	99.68	90.64
	Www-CBOW	300	0	99.70	96.21	97.92
			1	95.99	99.68	97.80
		100	0	99.60	95.17	97.33
			1	94.94	99.58	97.20
	Www-SG	300	0	99.21	96.21	97.69
			1	95.97	99.16	97.54
		100	0	99.19	58.42	73.53
			1	68.55	99.48	81.17

Tableau 3. 4: Comparaison des résultats

Modèle	Accuracy	Precision	Recall	F1 Score
CNN	95.87%	94.40%	94.39%	94.39%
CNN-LSTM	98.33%	97.86%	97.86%	97.86%

6. Conclusions

Ce chapitre met en lumière les implications de l'utilisation de modèles de DL pour identifier les versets coraniques dans le contenu textuel en arabe. Nos modèles sont basés sur des classificateurs de DL, CNN et LSTM, et reposent uniquement sur des WE pré-entraînés basés sur trois sources de données complètement différentes. Malgré la difficulté de la langue arabe et le manque de répétition des versets coraniques, nos modèles ont obtenu un résultat satisfaisant.

De cette étude, nous concluons que :

- Les techniques de DL et de WE sont capables de détecter les versets arabes avec un accuracy de 98,33 % et 97,86 % pour la précision, le rappel et le score F1.

- Le classificateur CNN-LSTM surpasse largement le classificateur CNN. Cela explique la capacité du modèle CNN à extraire des caractéristiques profondes et à les transmettre au modèle LSTM, qui effectue la classification basée sur les caractéristiques extraites.

- Les modèles construits avec le contenu textuel de Twitter surpassent les autres sources de données utilisées pour apprendre les modèles WE, en raison de son utilisation fréquente par les pionniers de l'internet arabes et de la diversité de leurs contenus de données.

- Les modèles construits avec CBOW ont pu représenter les mots répétés dans le Coran de manière plus efficace, et ils ont obtenu une meilleure vitesse d'entraînement que SG.

- Les vecteurs WE pré-entraînés de plus grande taille stockent plus d'informations car il existe de nombreux cas possibles, et la qualité de la représentation des mots se détériore chaque fois que la taille du vecteur est inférieure à 300.

Ce travail aidera à identifier le contenu coranique, ainsi qu'à renforcer la confiance des utilisateurs d'Internet à l'égard des citations coraniques.

Dans certains cas, ces modèles échouent à déterminer l'ordre des mots des versets. Dans le chapitre suivant, nous nous concentrerons sur ce point et essayerons différents modèles de DL pour améliorer les résultats.

Chapitre 4

Authentification des séquences de contenu coranique à l'aide de Deep Learning

1. Introduction

À l'heure actuelle, la détection de textes falsifiés est devenue un axe majeur de recherche dans les domaines du Traitement du Langage Naturel (NLP) et de l'Intelligence Artificielle (IA). Dans le contexte du texte coranique, garantir à la fois l'authentification et l'intégrité est primordial pour la lecture et l'écriture des versets coraniques [42]. Originaire de la langue arabe, le Saint Coran demeure le texte le plus authentique depuis l'antiquité, mais il est également hautement vulnérable à la manipulation dans le monde numérique. Révélé initialement au Prophète Muhammad "Que la paix soit sur lui" sur une période de 23 ans, le Coran a été transcrit sur divers matériaux tels que la peau, les os, les feuilles et les pierres pour protéger son contenu et maintenir la séquence des versets. Cependant, à l'ère numérique, des initiatives frauduleuses cherchent à modifier les règles standard d'arrangement du contenu coranique, ce qui pourrait entraîner la création de nouveaux versets/sourates contenant des versets incohérents avec la version authentique du Coran. Détecter de telles altérations, en particulier les chevauchements entre les versets, constitue un défi pour les lecteurs coraniques numériques, soulignant la nécessité d'outils de vérification automatisés pour garantir l'authenticité du Coran. Par conséquent, il est urgent de développer des algorithmes de DL spécialisés en NLP, tels que les modèles LSTM réputés pour leur efficacité dans le traitement des données séquentielles [83]. Dans ce chapitre, nous nous concentrons sur la conception d'un modèle de DL pour authentifier l'arrangement du contenu du Saint Coran.

2. Travaux connexes

Certains travaux de recherche se sont intéressés à l'authentification, l'intégrité et la sécurité du Saint Coran en appliquant plusieurs techniques à plusieurs formes du Coran telles que des images [18] [20] [86] [87] [88] [89] [90] des textes, etc.

Pour la forme textuelle, les travaux connexes peuvent être divisés en fonction de la technologie utilisée. Les travaux [70] et [91] ont utilisé des techniques de tatouage numérique pour protéger et vérifier le contenu du Coran.

Divers types d'algorithmes de recherche de chaînes Boyer-Moore ont été appliqués pour faire correspondre le contenu du Coran [5] [64] [66]. Tandis que [68] a combiné les deux technologies précédentes. Dans un travail similaire, [92] a utilisé un algorithme de similarité de chaînes pour détecter les altérations des versets. Pour vérifier l'authenticité des versets, [40] et [34] ont utilisé des techniques de recherche et de filtrage traditionnelles nécessitant l'identification manuelle des versets. D'autres travaux se sont penchés sur les fonctions de hachage cryptographiques pour vérifier l'intégrité des versets [42] [71] [93] [94] .

Un travail a tenté l'application d'algorithmes de ML pour l'identification automatique des mots du Coran intégrés dans des textes numériques [41].

Une des faiblesses de ces travaux est de négliger l'ordonnancement et la séquence des versets. Malgré la puissance des modèles de DL [81] dans le traitement des chaînes et des mots

adjacents des textes, la littérature précédente n'a montré aucune approche basée sur cette technique. Considérant ce problème comme un processus de classification, dans ce travail, nous proposons de construire un modèle de réseau neuronal basé sur le modèle LSTM pour classer l'ordonnancement des versets.

3. Méthodologie proposée

Ce travail vise à résoudre le problème de la perversion du Coran basée sur la manipulation au niveau de l'arrangement.

Il existe de nombreuses formes de manipulation de l'ordre des mots / versets du Saint Coran. Dans la plupart des cas, il est difficile pour le facteur humain de découvrir ce type d'erreur en un temps record lors de la lecture ou de l'écriture du contenu du Coran. L'arrangement peut être délibérément inversé, ce qui est très courant dans les travaux de sorcellerie et de magie, ou cela peut être le résultat d'un hasard, résultant de l'arrangement de mots arabes qui donne un mélange du Coran.

Par conséquent, nous avons limité ces possibilités en trois catégories, comprenant la catégorie correcte telle que rapportée dans les Corans d'Othman, la catégorie désordonnée résultant de la combinaison de mots arabes, et la catégorie délibérément inversée utilisée pour les actions diaboliques indésirables. La figure 4.1 représente notre méthodologie proposée pour construire un modèle intelligent capable de distinguer entre ces catégories.

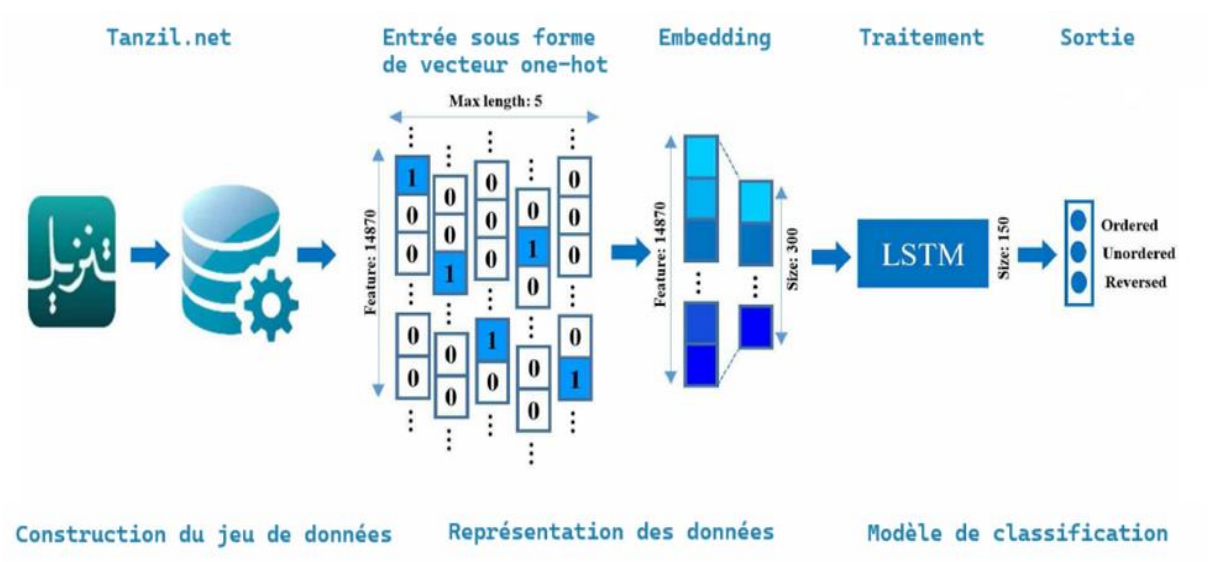


Figure 4. 1: Méthodologie proposée

3.1. Construction des données

En nous appuyant sur Tanzil [45], le site le plus fiable du Coran, nous avons extrait tous les mots du Coran tels qu'ils apparaissaient dans la même séquence. Dans le travail présenté par [25], nous avons découvert que la séquence de deux, trois ou quatre mots peut exister sous plusieurs formes dans le Coran, nous avons donc choisi le nombre cinq comme longueur fixe pour diviser la séquence de mots, et éviter de créer des phrases similaires.

Ensuite, nous avons généré 3 catégories distinctes de phrases sans nous soucier du sens des phrases composées :

- Phrases coraniques ordonnées : En considérant que les 5 mots dans une séquence constituent une phrase valide, nous avons obtenu 78 248 phrases appartenant à cette catégorie.

- Phrases coraniques désordonnées : En sélectionnant aléatoirement 5 mots de la série de mots à chaque fois, nous avons également formé 78 248 phrases désordonnées.

- Phrases coraniques inversées : Contrairement à chacune des chaînes de phrases ordonnées, nous avons à nouveau obtenu 78 248 phrases inversées.

La version finale de cette collection est présentée sous forme de fichier CSV, chaque ligne de ce fichier contient un morceau de texte d'une longueur de cinq mots et une balise indiquant l'une des catégories équilibrées : O pour Ordre Coranique (Ordered), U pour Coran Désordonné (Unordred) et R pour Coran Inversé (Reversed). Un échantillon du jeu de données est présenté dans le Tableau 4.1.

Tableau 4. 1: Un Échantillon du Jeu de Données Développé

Ordered	Unordered	Reversed
['بسم', 'الله', 'الرحمن', 'الرحيم', 'الحمد']	['بسم', 'الله', 'رب', 'العرش', 'العظيم']	[' الحمد', 'الرحيم', 'الرحمن', 'الله', 'بسم']
['الله', 'الرحمن', 'الرحيم', 'الحمد', 'لله']	['قل', 'اعوذ', 'من', 'شر', 'الفلق']	[' لله', 'الحمد', 'الرحيم', 'الرحمن', 'الله']
['الرحمن', 'الرحيم', 'الحمد', 'لله', 'رب']	['لا', 'نفرق', 'بين', 'المنافقين', 'والكفار']	[' رب', 'لله', 'الحمد', 'الرحيم', 'الرحمن']
...	...	...

3.2. Représentation des données

Tout d'abord, chaque mot est représenté sous forme d'un vecteur one-hot d'une longueur égale au nombre de mots dans le Coran sans répétition (14870 mots), où le nombre 1 est attribué au mot existant et le nombre 0 au mot non existant.

Ensuite, chaque phrase dans le jeu de données (chaque groupe de 5 mots consécutifs) est introduite en entrée dans la couche d'incorporation (Embedding layer) qui apprend les poids pour réduire l'ordre de 14870 à 300 vecteurs numériques.

3.3. Modèle de classification

Pour construire le modèle, nous avons sélectionné le réseau LSTM [95] pour le distinguer du reste des réseaux dans le traitement des données séquentielles. LSTM est le roi des réseaux neuronaux récurrents (RNN) qui étend l'idée d'inclure un retour d'information agissant comme une sorte de mémoire en créant un composant de mémoire à court et à long terme. Il se compose d'une porte d'entrée, qui est responsable de l'entrée des données, d'une porte

oubliée qui contrôle la durée de conservation des données, et d'une porte de sortie, qui est responsable de la valeur de l'état caché suivant.

La sortie de la couche d'incorporation est alimentée à la couche LSTM de 150 unités de mémoire en batches de 64 pour effectuer l'entraînement et activer la couche de sortie pour produire un vecteur numérique de longueur 3 via la fonction d'activation Softmax. Ensuite, l'indicateur de la plus grande valeur est la couche attendue.

4. Expérimentations

4.1. Mesures d'évaluation

Pour mieux comprendre les performances du modèle, nous avons préféré utiliser une matrice de confusion pour trouver les prédictions des vrais positifs (TP), des faux positifs (FP), des vrais négatifs (TN) et des faux négatifs (FN). Sur la base de ces prédictions, nous avons calculé quatre mesures supplémentaires : l'exactitude, le rappel, la précision et le score F1, qui sont basés sur les valeurs prédites par rapport aux vraies valeurs.

- Accuracy : une portion d'exemples correctement classés.

- Reacall (Rappel) : La fraction de séquences ordonnées / désordonnées / inversées qui sont classées comme ordonnées / désordonnées / inversées.

- Precison (Précision) : La fraction de séquences classées comme ordonnées / désordonnées / inversées qui sont réellement ordonnées / désordonnées / inversées.

- Score F1 : Le score F1 pour la classe ordonnée / désordonnée / inversée est la moyenne harmonique de sa précision et de son rappel.

4.2. Résultats expérimentaux

Après avoir entraîné le modèle proposé pendant 50 époques et en utilisant une valeur de dropout de 0,2, sans se fier à une quelconque intégration de mots pré-entraînée, nous avons testé le modèle sur 20 % du jeu de données total, et une précision de test de 99,98 % a été obtenue.

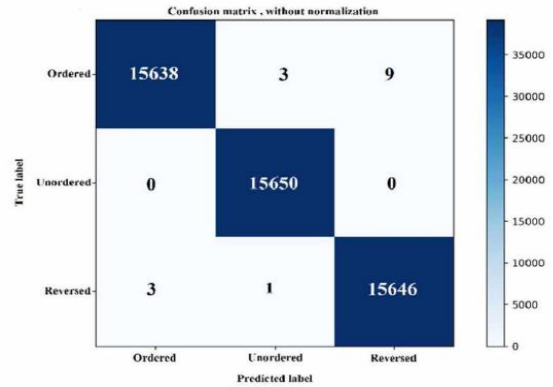


Figure 4. 2: Matrice de confusion pour les prédictions du modèle LSTM

La figure 4.2 représente les prédictions du modèle LSTM. Le pourcentage de test est de 20 % du jeu de données total. La majorité des prédictions du modèle se trouvent dans la diagonale de la matrice, qui représentent respectivement : ce que le modèle a prédit comme étant ordonné et la séquence étant étiquetée comme ordonnée, ce que le modèle a prédit comme étant désordonné et la séquence étant étiquetée comme désordonnée, et ce que le modèle a prédit comme étant inversé et la séquence étant étiquetée comme inversée.

Le tableau 4.2 présente les mesures de performances avec des valeurs entre 0 et 1. Comme mentionné ci-dessus, le modèle LSTM a été capable d'identifier 99 % des séquences ordonnées, 100 % des séquences désordonnées et 99 % des séquences inversées. Le modèle a échoué à classer 1 % des séquences ordonnées ainsi que les séquences inversées.

Tableau 4. 2: Résultat de modèle proposé

Classe	Recall	Precison	F1 score
Ordered	0.9992	0.9998	0.9995
Unordered	1.000	0.9997	0.9998
Reversed	0.9997	0.9994	0.9995

5. Discussion

Comparé aux systèmes précédents, nous avons choisi de travailler sur l'authentification du verset coranique le plus célèbre utilisé sur les réseaux sociaux, qui est :

‘’وبشر الصابرين الذين اذا اصابتهم مصيبة قالوا إنا لله و انا اليه راجعون‘‘

Nous avons d'abord recherché ce verset sur le site le plus honnête dans la fourniture du Coran "Tanzil.net", et le site le plus utilisé pour télécharger des données dans tous les travaux connexes mentionnés précédemment.

Le résultat de la recherche apparaît dans la figure 3, où il a été constaté que le site a échoué à authentifier ce verset, ce qui prouve que la plupart des sites et des méthodes de recherche précédentes appellent à prédéterminer le début et la fin du verset, et cela est classé dans les méthodes de recherche traditionnelles qui ne sont pas compatibles avec l'intégration de plus d'un verset pendant le processus de recherche et d'authentification, et ignore également l'ordre des positions des versets.

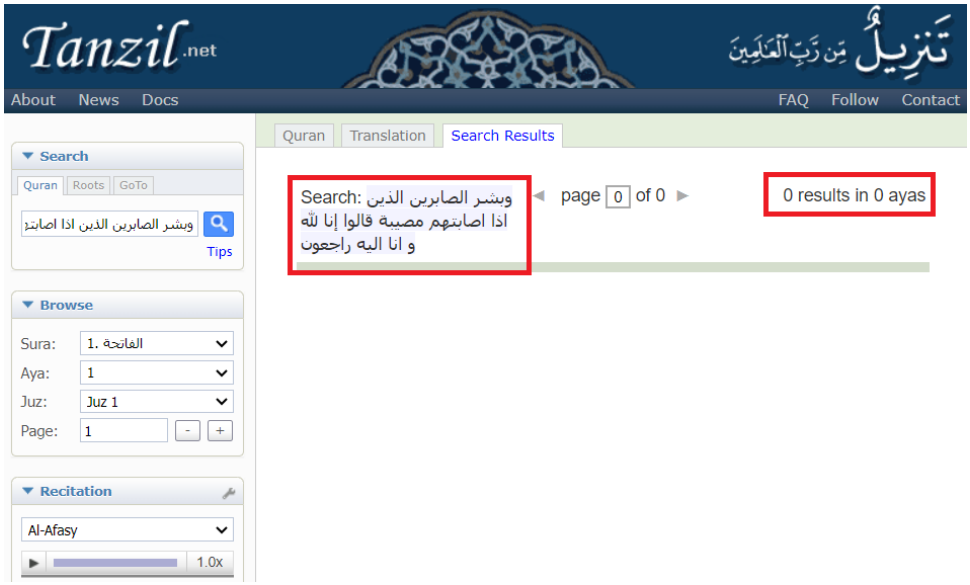


Figure 4. 3: Le résultat de Tanzil.net

D'autre part, nous avons appliqué notre modèle proposé dans une tentative d'authentifier ce verset coranique, où le modèle divise d'abord le passage en séquences de 5 mots, puis documente chaque séquence. Les résultats préliminaires sont présentés dans la figure 4, où il devient clair que le modèle proposé a pu surpasser le site précédent dans l'authentification du verset coranique le plus célèbre utilisé sur les réseaux sociaux.

```
وبشر الصابرين الذين اذا اصابتهم
result : authenticated as correct

الصابرين الذين اذا اصابتهم مصيبة
result : authenticated as correct

الذين اذا اصابتهم مصيبة قالوا
result : authenticated as correct

اذا اصابتهم مصيبة قالوا انا
result : authenticated as correct

اصابتهم مصيبة قالوا انا لله
result : authenticated as correct

مصيبة قالوا انا لله وانا
result : authenticated as correct

قالوا انا لله وانا اليه
result : authenticated as correct

انا لله وانا اليه راجعون
result : authenticated as correct
```

Figure 4. 4: Le résultat de notre modèle

6. Conclusion

Ce chapitre a démontré avec succès la puissance de DL dans la préservation des séquences de texte les plus sensibles à la distorsion pendant l'entraînement et les tests. En construisant un ensemble de données qui inclut toutes les instances de séquences de mots coraniques continues et discontinues, et en s'appuyant sur l'algorithme LSTM pour développer un modèle de classification permettant de distinguer entre les versets corrects, désordonnés et inversés, les résultats ont indiqué que l'algorithme LSTM a pu résoudre ce problème efficacement et surpasser le reste des méthodes, en atteignant une précision de 99,98%.

Conclusion générale & Perspectives

CONCLUSION GENERALE & PERSPECTIVES

Cette conclusion traite de l'ensemble des travaux qui ont été menés pour compléter la recherche proposée. Dans un premier temps, les objectifs de la recherche sont revisités pour expliquer le but et les résultats de cette étude. Deuxièmement, les contributions des recherches menées sont mises en évidence. Enfin, les limites de la recherche menée ainsi que les recommandations futures sont présentées.

1. Objectifs de recherche revisités

Cette section revisite les objectifs de recherche des recherches menées

1.1. Objectif de recherche 1

Le premier objectif passe en revue l'état de l'art dans l'authentification du texte coranique. Différentes approches telles que la correspondance de motifs, le tatouage numérique, le hachage, le SQL et les approches basées sur ML sont identifiées comme des approches potentielles pouvant être utilisées pour authentifier différents contenus coraniques. Les études ayant exploré la stéganographie, la cryptographie, le tatouage numérique et la blockchain pour protéger le texte coranique ont également été étudiées. Après avoir mené une investigation préalable de toutes les études pertinentes, une taxonomie basée sur l'authentification de l'intégrité du contenu du texte coranique est élaborée. L'objectif est d'aider les chercheurs à identifier les approches d'authentification appropriées. Il convient de noter que notre étude s'appuie sur l'état de l'art établi par Hakak [3], tout en intégrant des mises à jour spécifiques correspondant à nos besoins de recherche.

1.2. Objectif de recherche 2

Le deuxième objectif est d'identifier la meilleure technique de représentation des données du contenu du Coran. Pour atteindre cet objectif, la structure du contenu du Coran est étudiée et certaines techniques de NLP sont testées. La technique de Word Embeddings a été choisie en fonction de sa robustesse. Une discussion détaillée est présentée dans le Chapitre 2.

1.3. Objectif de recherche 3

Le troisième objectif est d'identifier le contenu coranique intégré dans du texte arabe, en particulier dans des textes simples (sans diacritiques). Suite à la réalisation de l'Objectif 1, il a été observé que la plupart des versets n'ont pas été récupérés correctement, ou ont été récupérés manuellement. Par conséquent, une nouvelle approche basée sur le DL et le word Embeddings est proposée. Cette approche améliore considérablement le processus d'identification. Les étapes nécessaires à la réalisation de cet objectif sont présentées dans le Chapitre 3.

1.4. Objectif de recherche 4

L'objectif final est de surmonter la limitation de l'Objectif 3 en authentifiant les versets au niveau de l'ordre. À cette fin, une approche basée sur le DL est explorée. Les détails de cette approche sont fournis dans le Chapitre 4.

2. Contribution de la Recherche

La contribution de cette recherche au domaine du traitement des données en termes d'authentification (identification et authentification) est mise en évidence par les méthodes novatrices proposées pour l'authentification des textes coraniques. Différents problèmes liés à l'authentification de l'intégrité du contenu des textes en arabe, en particulier le contenu coranique, ont été identifiés. Sur la base de ces problèmes, différentes méthodes ont été proposées pour rechercher et authentifier des textes complexes afin de détecter les altérations du contenu. Le champ d'étude des textes sensibles a été restreint à l'étude de cas du Coran numérique arabe, compte tenu de sa nature grammaticale complexe et de sa sensibilité à la manipulation minutieuse, ce qui en fait un matériau particulièrement adapté pour une étude de cas.

Un cadre complet relatif à l'identification et à l'authentification du texte sensible est proposé avec l'étude de cas du Saint Coran numérique. Étant donné que le champ de cette étude est limité à la phase d'authentification de l'intégrité, quatre objectifs différents ont été identifiés comme discuté précédemment.

Le premier objectif de notre recherche consiste à déterminer les méthodes les plus efficaces de représentation des textes coraniques dans le domaine NLP. Après avoir atteint le premier objectif, il est observé que cette étape conduit à une analyse approfondie des recherches existantes sur l'identification de textes coraniques. Le deuxième objectif est spécifiquement axé sur la détection des textes coraniques intégrés ou cités dans des contextes de texte arabe à l'aide de styles d'écriture simples. Pour ce faire, nous avons élaboré un ensemble de données en combinant des versets coraniques avec des textes arabes, et notre modèle a obtenu une précision impressionnante de 98,33% dans cette tâche. Cette performance a été atteinte grâce à l'utilisation de techniques de DL, notamment les plongements de mots (Word Embeddings), combinées à des classificateurs CNN et LSTM pour la classification binaire. Les résultats expérimentaux ont démontré la supériorité de cette approche par rapport aux méthodes traditionnelles pour distinguer avec précision entre les versets coraniques et le texte en arabe.

Dans certains cas, ce modèle échoue à déterminer l'ordre des mots des versets. Pour aborder cette lacune, le troisième objectif est défini. Une nouvelle méthode d'authentification des versets basée sur l'ordre est proposée pour répondre à cet objectif. Cette méthode utilise des algorithmes de DL pour authentifier automatiquement l'intégrité de l'arrangement des mots dans le contenu coranique. Nous avons choisi l'algorithme de Mémoire à Long Terme et Court Terme (LSTM) pour cette tâche, et les résultats ont été exceptionnels, atteignant une précision de test de 99,98% sur l'ensemble de données que nous avons constitué en utilisant les données du site Tanzil.

Ces différentes méthodes peuvent être utilisées pour développer un système capable d'identifier et d'authentifier les versions numériques des textes coraniques.

Toutes les conclusions de cette recherche sont publiées dans des revues académiques et des actes de conférence. De plus, une autre approche de protection du Saint Coran numérique a été publiée, les détails sont donnés dans la section des publications.

3. Limitations et travaux futurs

Le domaine d'étude des textes du Coran arabe et de leur authentification offre de nombreuses opportunités de recherche. Toutes les étapes, comme la présentation des données et les phases d'identification/d'authentification, présentent certaines limites qui peuvent être étudiées, évaluées et corrigées. Les limitations inhérentes à cette étude sont mises en évidence ci-dessous :

3.1.Limitations de l'objectif 2

Bien que l'objectif principal soit d'identifier la meilleure technique de représentation des données du contenu du Coran, une limitation notable réside dans le fait que les techniques de NLP testées peuvent ne pas être exhaustives. Malgré la sélection de la technique de Word Embeddings en raison de sa robustesse, il est possible que d'autres approches de représentation des données puissent offrir des résultats différents ou complémentaires. Par conséquent, il est important de reconnaître que l'objectif 2 ne peut garantir l'exhaustivité dans l'exploration des techniques de représentation des données pour le contenu coranique.

3.2.Limitations de l’objectif 3

Bien que l'objectif soit de repérer le contenu coranique intégré dans du texte arabe en utilisant des styles d'écriture simples, une limitation importante réside dans la complexité potentielle des styles d'écriture. Les textes arabes peuvent présenter une grande variabilité linguistique et stylistique, ce qui peut rendre difficile la détection automatique du contenu coranique. Par conséquent, malgré l'utilisation de techniques de DL et de plongements de mots pour améliorer la précision, il peut y avoir des cas où le modèle ne parvient pas à détecter avec précision les versets coraniques, en particulier dans des contextes d'écriture complexe.

3.3.Limitations de l’objectif 4

Bien que l'objectif soit de surmonter la limitation de l'objectif 3 en authentifiant les versets au niveau de l'ordre, une limitation potentielle réside dans la capacité du modèle à généraliser à des ensembles de données plus vastes ou à des contextes plus variés. Bien que l'algorithme LSTM ait démontré une précision exceptionnelle dans l'authentification de l'ordre des mots dans les versets, il est possible qu'il rencontre des difficultés lorsqu'il est confronté à des verset long / surates ou contextuelles significatives.

Par conséquent, il est important de valider la robustesse et la généralisabilité de l'approche proposée sur une gamme diversifiée de données et de contextes pour garantir sa fiabilité dans

des applications réelles.

Pour aborder ces limitations et améliorer davantage les résultats, plusieurs pistes de recherche peuvent être envisagées. Premièrement, une exploration plus large des techniques de représentation des données en NLP pourrait être entreprise, en intégrant des approches telles que les modèles de langage pré-entraînés et les techniques de traitement des séquences pour capturer de manière plus efficace la structure et le sens des versets coraniques. Deuxièmement, des recherches supplémentaires pourraient être menées pour affiner les modèles d'authentification des versets au niveau de l'ordre, en explorant des architectures de réseaux neuronaux plus complexes ou en intégrant des mécanismes d'attention pour mieux modéliser les relations entre les mots dans les versets. Enfin, une évaluation plus approfondie de la généralisabilité des modèles proposés sur des ensembles de données variés et dans des contextes d'utilisation réelle pourrait être réalisée pour garantir leur applicabilité dans divers scénarios d'authentification de texte coranique.

Références

Liste des références

1. Moukdad, H., Cui, H.: How do search engines handle Chinese queries. Webology. 2, 18 (2005).

2. Pinkerton, B.: Webcrawler: Finding what people want. University of Washington (2000).

3. Hakak, S., Kamsin, A., Tayan, O., Idris, M.Y.I., Gani, A., Zerdoumi, S.: Preserving content integrity of digital holy Quran: Survey and open challenges. Ieee Access. 5, 7305–7325 (2017).

4. Pan, P.J., Huang, H.-C., Jain, L.C.: Intelligent watermarking techniques (with Cd-rom). World scientific (2004).

5. Hakak, S.I., Kamsin, A., Idris, M.Y.I., Gani, A., Amin, G., Zerdoumi, S.: Diacritical digital Quran authentication model. Pertanika J Sci Tech. 25, 133–142 (2017).

6. Hakak, S., Kamsin, A., Tayan, O., Idris, M.Y.I., Gilkar, G.A.: Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges. Inf. Process. Manag. 56, 367–380 (2019).

7. Daraee, F., Mozaffari, S.: Watermarking in binary document images using fractal codes. Pattern Recognit. Lett. 35, 120–129 (2014).

8. Haouzia, A., Noumeir, R.: Methods for image authentication: a survey. Multimed. Tools Appl. 39, 1–46 (2008).

9. Tao, H., Chongmin, L., Zain, J.M., Abdalla, A.N.: Robust image watermarking theories and techniques: A review. J. Appl. Res. Technol. 12, 122–138 (2014).

10. Arnold, M., Schmucker, M., Wolthusen, S.D.: Techniques and applications of digital watermarking and content protection. Artech House (2002).

11. Makbol, N.M., Khoo, B.E., Rassem, T.H.: Block-based discrete wavelet transform-singular value decomposition image watermarking scheme using human visual system characteristics. IET Image Process. 10, 34–52 (2016).

12. Nin, J., Ricciardi, S.: Digital watermarking techniques and security issues in the information and communication society. In: 2013 27th International Conference on Advanced Information Networking and Applications Workshops. pp. 1553–1558. IEEE (2013).

13. Singh, P., Chadha, R.S.: A survey of digital watermarking techniques, applications and attacks. Int. J. Eng. Innov. Technol. IJEIT. 2, 165–175 (2013).

14. Brassil, J.T., Low, S., Maxemchuk, N.F.: Copyright protection for the electronic distribution of text documents. Proc. IEEE. 87, 1181–1196 (1999).

15. Khare, V., Shivakumara, P., Raveendran, P.: Multi-oriented moving text detection. In: 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). pp. 347–352. IEEE (2014).

16. Gutub, A.A.-A., Al-Alwani, W.: Improved method of Arabic text steganography using the extension 'Kashida'character. Bahria Univ. J. Inf. Commun. Technol. BUJICT. 3, (2010).

17. Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. IBM Syst. J. 35, 313–336 (1996).

18. Laouamer, L., Tayan, O.: An enhanced SVD technique for authentication and protection of text-images using a case study on digital Quran content with sensitivity constraints. Life Sci. J. 10, 2591–2597 (2013).

19. Abuhaija, B., Shilbayeh, N., Alwakeel, M.: Security protocol architecture for website authentications and content integrity. In: 2013 World Congress on Computer and Information Technology (WCCIT). pp. 1–6. IEEE (2013).

20. Kurniawan, F., Khalil, M.S., Khan, M.K., Alginahi, Y.M.: Authentication and tamper detection of digital Holy Quran images. In: 2013 International Symposium on Biometrics and Security Technologies. pp. 291–296. IEEE (2013).

21. Mitali, V.K., Sharma, A.: A survey on various cryptography techniques. Int. J. Emerg. Trends Technol. Comput. Sci. IJETTCS. 3, 307–312 (2014).

22. Schellekens, M.: Electronic signatures: authentication technology from a legal perspective. Springer (2004).

23. Tayan, O., Kabir, M.N., Alginahi, Y.M.: A hybrid digital-signature and zero-watermarking approach for authentication and protection of sensitive electronic documents. Sci. World J. 2014, (2014).

24. Sumathi, C.P., Santanam, T., Umamaheswari, G.: A study of various steganographic techniques used for information hiding. ArXiv Prepr. ArXiv14015561. (2014).

25. Katzenbeisser, S., Petitcolas, F.: Information hiding. Artech house (2016).

26. Bennett, K.: Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text. CERIAS Tech Rep. 2004-13. (2004).

27. Touati-Hamad, Z., Laouar, M.R., Bendib, I.: A Secure Framework Design for Digital

Sensitive Content Certification. In: International Conference on Computing and Information Technology. pp. 352–358. Springer (2022).

28. Touati-Hamad, Z., Laouar, M.R., Bendib, I.: Towards blockchain-based document authentication: application for digital mushaf al-quran authentication. Int. J. Organ. Collect. Intell. IJOCI. 12, 1–15 (2022).

29. Xuehua, J.: Digital watermarking and its application in image copyright protection. In: 2010 International Conference on Intelligent Computation Technology and Automation. pp. 114–117. IEEE (2010).

30. Coron, J.-S.: What is cryptography? IEEE Secur. Priv. 4, 70–73 (2006).

31. Delfs, H., Knebl, H., Delfs, H., Knebl, H.: Symmetric-key cryptography. Introd. Cryptogr. Princ. Appl. 11–48 (2015).

32. Cole, E.: Hiding in plain sight. Wiley Hoboken (2002).

33. Date, C.J., Darwen, H.: A guide to the SQL standard: a user's guide to the standard database language. Addison-Wesley Longman, Inc., 4th edition edition (1997).

34. Alshareef, A., El Saddik, A.: A Quranic quote verification algorithm for verses authentication. In: 2012 International Conference on Innovations in Information Technology (IIT). pp. 339–343. IEEE (2012).

35. Navarro, G.: A guided tour to approximate string matching. ACM Comput. Surv. CSUR. 33, 31–88 (2001).

36. Aho, A.V., Corasick, M.J.: Efficient string matching: an aid to bibliographic search. Commun. ACM. 18, 333–340 (1975).

37. Chi, L., Zhu, X.: Hashing techniques: A survey and taxonomy. ACM Comput. Surv. Csur. 50, 1–36 (2017).

38. Touati-Hamad, Z., Laouar, M.R., Bendib, I., Hakak, S.: Arabic quran verses authentication using deep learning and word embeddings. Int. Arab J. Inf. Technol. 19, 681–688 (2022).

39. Alginahi, Y.M., Tayan, O., Kabir, M.N.: Verification of qur'anic quotations embedded in online arabic and islamic websites. Int. J. Islam. Appl. Comput. Sci. Technol. 1, 41–47 (2013).

40. Sabbah, T., Selamat, A.: A framework for Quranic verses authenticity detection in online forum. In: 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences. pp. 6–11. IEEE (2013).

41. Sabbah, T., Selamat, A.: Support vector machine based approach for quranic words detection in online textual content. In: 2014 8th. Malaysian Software Engineering Conference

(MySEC). pp. 325–330. IEEE (2014).

42. Alsmadi, I., Zarour, M.: Online integrity and authentication checking for Quran electronic versions. Appl. Comput. Inform. 13, 38–46 (2017).

43. Kamsin, A., Gani, A., Suliaman, I., Jaafar, S., Mahmud, R., Sabri, A.Q.M., Razak, Z., Idris, M.Y.I., Ismail, M.A., Noor, N.M.: Developing the novel Quran and Hadith authentication system. In: The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M). pp. 1–5. IEEE (2014).

44. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. J. Am. Med. Inform. Assoc. 18, 544–551 (2011).

45. Tanzil, https://tanzil.net/docs/home, last accessed 2024/04/12.

46. Hammo, B., Sleit, A., El-Haj, M.: Enhancing retrieval effectiveness of diacritisized Arabic passages using stemmer and thesaurus. In: Proc. of The 19th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS2008) (2008).

47. Liu, Z., Lin, Y., Sun, M., Liu, Z., Lin, Y., Sun, M.: Word representation. Represent. Learn. Nat. Lang. Process. 13–41 (2020).

48. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 384–394 (2010).

49. Bhardwaj, A., Di, W., Wei, J.: Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling. Packt Publishing Ltd (2018).

50. Zhang, Y., Jin, R., Zhou, Z.-H.: Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cybern. 1, 43–52 (2010).

51. Ramos, J.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. pp. 29–48. Citeseer (2003).

52. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165 (1958).

53. Kowalski, G.J., Maybury, M.T.: Information storage and retrieval systems: theory and implementation. Springer Science & Business Media (2000).

54. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28, 11–21 (1972).

55. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. Adv. Neural Inf. Process. Syst. 13, (2000).

56. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word

representations in vector space. ArXiv Prepr. ArXiv13013781. (2013).

57. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196. PMLR (2014).

58. Touati-Hamad, Z., Laouar, M.R., Bendib, I.: Quran content representation in NLP. In: Proceedings of the 10th International Conference on Information Systems and Technologies. pp. 1–6 (2020).

59. Hakak, S., Kamsin, A., Palaiahnakote, S., Tayan, O., Idna Idris, M.Y., Abukhir, K.Z.: Residual-based approach for authenticating pattern of multi-style diacritical Arabic texts. PloS One. 13, e0198284 (2018).

60. Arkok, B., Zeki, A.M.: Classification of quranic topics using ensemble learning. In: 2021 8th International Conference on Computer and Communication Engineering (ICCCE). pp. 244–248. IEEE (2021).

61. Pudaruth, S., Soyjaudah, S., Gunputh, R.: Classification of legislations using deep learning. Int Arab J Inf Technol. 18, 651–662 (2021).

62. Hakak, S., Kamsin, A., Zada Khan, W., Zakari, A., Imran, M., bin Ahmad, K., Amin Gilkar, G.: Digital Hadith authentication: Recent advances, open challenges, and future directions. Trans. Emerg. Telecommun. Technol. 33, e3977 (2022).

63. Zerdoumi, S., Sabri, A.Q.M., Kamsin, A., Hashem, I.A.T., Gani, A., Hakak, S., Al-Garadi, M.A., Chang, V.: Image pattern recognition in big data: taxonomy and open challenges: survey. Multimed. Tools Appl. 77, 10091–10121 (2018).

64. Gilkar, G.A., Hakak, S., Kamsin, A., Rahman, M.M., Rahman, M.: An Exact matching approach to enhance retrieval process for Quranic texts. In: Proceedings of the 4th ACM International Conference of Computing for Engineering and Sciences. pp. 1–4 (2018).

65. Hakak, S., Kamsin, A., Shivakumara, P., Idna Idris, M.Y., Gilkar, G.A.: A new split based searching for exact pattern matching for natural texts. PloS One. 13, e0200912 (2018).

66. Hakak, S., Kamsin, A., Shivakumara, P., Tayan, O., Idris, M.Y.I., amin Gilkar, G.: An efficient text representation for searching and retrieving classical diacritical arabic text. Procedia Comput. Sci. 142, 150–157 (2018).

67. Hakak, S., Kamsin, A., Shivakumara, P., Idris, M.Y.I.: Partition-based pattern matching approach for efficient retrieval of Arabic text. Malays. J. Comput. Sci. 31, 200–209 (2018).

68. Hakak, S., Kamsin, A., Veri, J., Ritonga, R., Herawan, T.: A framework for authentication of digital Quran. In: Information Systems Design and Intelligent Applications: Proceedings of Fourth International Conference INDIA 2017. pp. 752–764. Springer (2018).

69. Hakak, S.I., Kamsin, A., Shivakumara, P., Gilkar, G.A., Khan, W.Z., Imran, M.: Exact string matching algorithms: survey, issues, and future research directions. IEEE Access. 7, 69614–69637 (2019).

70. Kamaruddin, N.S., Kamsin, A., Hakak, S.: Associated diacritical watermarking approach to protect sensitive arabic digital texts. In: AIP Conference Proceedings. AIP Publishing (2017).

71. Almazrooie, M., Samsudin, A., Gutub, A.A.-A., Salleh, M.S., Omar, M.A., Hassan, S.A.: Integrity verification for digital Holy Quran verses using cryptographic hash function and compression. J. King Saud Univ.-Comput. Inf. Sci. 32, 24–34 (2020).

72. Abdellatif, M., Elgammal, A.: Offensive language detection in Arabic using ULMFiT. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. pp. 82–85 (2020).

73. Elnagar, A., Al-Debsi, R., Einea, O.: Arabic text classification using deep learning models. Inf. Process. Manag. 57, 102121 (2020).

74. Hussein, A., Al Kafri, M., Abonamah, A.A., Tariq, M.U.: Mood detection based on Arabic text documents using machine learning methods. Int. J. 9, (2020).

75. Touati-Hamad, Z., Laouar, M.R., Bendib, I.: Authentication of quran verses sequences using deep learning. In: 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI). pp. 1–4. IEEE (2021).

76. Alfaifi, A.Y.G., Atwell, E., Hedaya, I.: Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners. In: Proceedings of Learner Corpus Studies in Asia and the World 2014. pp. 77–89. Kobe International Communication Center (2014).

77. Abozinadah, E.A., Mbaziira, A.V., Jones, J.: Detection of abusive accounts with Arabic tweets. Int J Knowl Eng-IACSIT. 1, 113–119 (2015).

78. Goldberg, Y., Levy, O.: word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. ArXiv Prepr. ArXiv14023722. (2014).

79. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 26, (2013).

80. Soliman, A.B., Eissa, K., El-Beltagy, S.R.: Aravec: A set of arabic word embedding models for use in arabic nlp. Procedia Comput. Sci. 117, 256–265 (2017).

81. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Netw. 61, 85–117 (2015).

82. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to

document recognition. Proc. IEEE. 86, 2278–2324 (1998).

83. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9, 1735–1780 (1997).

84. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv 2018. ArXiv Prepr. ArXiv180308375. (1803).

85. Ruder, S.: An overview of gradient descent optimization algorithms. ArXiv Prepr. ArXiv160904747. (2016).

86. AlAhmad, M.A., Alshaikhli, I., Alduwaikh, A.E.: A new fragile digital watermarking technique for a PDF digital Holy Quran. In: 2013 International Conference on Advanced Computer Science Applications and Technologies. pp. 250–253. IEEE (2013).

87. Hisham, S.I.: Localization Watermarking for Authentication of Text Images in Quran. (2013).

88. Kurniawan, F., Khalil, M.S., Khan, M.K., Alginahi, Y.M.: DWT+ LSB-based fragile watermarking method for digital Quran images. In: 2014 international symposium on biometrics and security technologies (ISBAST). pp. 290–297. IEEE (2014).

89. Kurniawan, F., Khalil, M.S., Khan, M.K., Alginahi, Y.M.: Exploiting digital watermarking to preserve integrity of the digital Holy Quran images. In: 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences. pp. 30–36. IEEE (2013).

90. Tuncer, T., Ertam, F., Avci, E.: A watermarking application for authentication of Holy Quran. In: 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences. pp. 37–41. IEEE (2013).

91. Alginahi, Y.M., Tayan, O., Kabir, M.N.: A zero-watermarking verification approach for Quranic verses in online text documents. In: 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences. pp. 42–46. IEEE (2013).

92. Milon Islam, M., Kabir, M.N., Sadi, M.S., Morsalin, M.I., Haque, A., Wang, J.: A novel approach towards tamper detection of digital holy quran generation. In: InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019. pp. 297–308. Springer (2020).

93. Dewi, A.S., Setiawan, H.: Implementation of SHA-256 and AES-256 for securing digital Al Quran verification system. In: 2019 Fourth International Conference on Informatics and Computing (ICIC). pp. 1–8. IEEE (2019).

94. AlAhmad, M.A., Alshaikhli, I., Jumaah, B.: Protection of the Digital Holy Quran hash

digest by using cryptography algorithms. In: 2013 International Conference on Advanced Computer Science Applications and Technologies. pp. 244–249. IEEE (2013).

95. Sundermeyer, M., Schlüter, R., Ney, H.: Lstm neural networks for language modeling. In: Interspeech. pp. 194–197 (2012).