Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen Faculté des Sciences Département d'Informatique

MÉMOIRE

Pour l'obtention du diplôme de Master

Spécialité : *Système d'information et de connaissance (SIC)*

Thème

Détection automatique de plagiat dans les textes arabes

Soutenue le **29/09/2024**

Par

Lakehal Ibrahim El-khalil

Berrabah Nasreddine

Devant le jury composé de

M M. MERZOUG	***Président***
Mme S.KHITRI	***Examinatrice***
Mme Y.SELADJI	***Experte I2E***
M A. ABDERRAHIM	***Encadrant***

Année universitaire 2023/2024

Dédicace

Tout d'abord, je tiens à remercier DIEU De m'avoir donné la force et le courage de mener à bien ce modeste travail.

Je tiens à dédier cet humble travail :

À ma tendre mère

pour leur soutien inconditionnel, leur amour et leurs encouragements constants tout au long de ce parcours académique.

À mes sœurs , chère frère et sa femme

pour leur dévouement, leur compréhension et leur grande tendresse, qui en plus de m'avoir encouragé tout le long de mes études.

À mon binôme Nasro. À mes meilleurs amis

Merci d'être toujours là pour moi.

LAKEHAL IBRAHIM EL-KHALIL

Dédicace

Tout d'abord, je tiens à remercier DIEU De m'avoir donné la force et le courage de mener à bien ce modeste travail.

Je tiens à dédier cet humble travail :

À ma mère pour leur soutien inconditionnel, leur amour et leurs encouragements constants tout au long de ce parcours académique.

À mon père pour ses efforts ,leurs conseils et leurs encouragements.

À ma sœur Wafaa et son mari Samir et ses enfants et ma petite sœur

À mes frère Ismail , Ilyes et sa femme et ses enfants.

À mes potes : Ibrahim ,Bilal ,Billel,Moha,Ayoub,Anes et autres amis et ma famille pour leur encouragement et leur confiance.

Merci d'être toujours là pour moi.

BERRABAH NASREDDINE

Remerciement

Nous remercions tout d'abord الله pour l'achèvement de ce modeste travail.

Nous remercions M. A. ABDERRAHIM, notre encadrant, pour ses conseils et suggestions avisés qui nous ont aidé à mener à bien ce travail, ainsi que pour les remarques et conseils apportés à ce mémoire.

Nous remercions infiniment les membres du jury : le président M. Merzoug, l'examinatrice Mme S. Khitri et l'experte I2E Mme Y.Seladji, pour leur présence, pour avoir accepté de lire et d'évaluer ce travail, ainsi que pour les remarques qu'ils nous adresseront lors de cette soutenance afin d'améliorer notre travail.

Nous voudrions exprimer notre reconnaissance envers les amis et collègues qui nous ont apporté leur soutien moral et intellectuel tout au long de notre démarche.

Table de matière

Table de matière

Liste des figures

Liste des tableaux

Introduction générale

1) Contexte :

Au 21e siècle, les avancées technologiques ont révolutionné de nombreux domaines, y compris la recherche scientifique. L'accès facilité à l'information via les moteurs de recherche offre de nombreux avantages, mais il pose également des défis en matière de créativité et de production intellectuelle.

Le manque de réflexion et de créativité dans les travaux de recherche est devenu apparent, certains chercheurs et étudiants optant pour la copie directe de contenu provenant de livres, d'articles ou de travaux antérieurs sans attribution de la source. Cette pratique viole les droits d'auteur et compromet la crédibilité des travaux. Ce phénomène de réutilisation non légitime du texte est connu sous le nom de plagiat.

Le plagiat est particulièrement répandu ces dernières années, notamment chez les étudiants, et se manifeste de diverses manières : copier-coller, paraphrase, traduction non autorisée ou empreint d'idées sans attribution. La détection automatique de plagiat dans les textes arabes est devenue un sujet crucial pour maintenir l'intégrité académique. [1]

Il existe deux approches principales pour détecter le plagiat dans les textes arabes : l'approche extrinsèque et l'approche intrinsèque (Potthast, Barrón-Cedeño, Stein, & Rosso, 2009).L'approche extrinsèque compare un document suspect avec d'autres documents pour identifier des similitudes. L'approche intrinsèque, quant à elle, analyse le document suspect pour repérer des changements de style qui pourraient indiquer un plagiat.

Compte tenu de la diversité et de la complexité des textes arabes, le développement de méthodes automatiques efficaces pour détecter le plagiat est essentiel pour préserver la qualité et la crédibilité de la recherche académique dans le monde arabe.

[1]Ce lien représente des statistiques sur le plagiat « https://plagiarism.org/article/plagiarism-facts-and-stats »

2) Motivation

Le plagiat reste un problème majeur dans le monde académique et est considéré comme une infraction sérieuse[2]. Qu'il soit intentionnel ou non, les étudiants qui commettent du plagiat risquent des sanctions telles que la suspension ou l'expulsion. De même, dans le milieu professionnel, le plagiat peut nuire à la réputation des chercheurs et des auteurs, et même entraîner la perte de leur emploi.[3]

Bien que le plagiat ait existé bien avant l'ère numérique, notamment sous la forme de vols de poésie, son incidence n'était pas aussi répandue qu'aujourd'hui. L'émergence d'Internet a facilité le plagiat académique grâce à l'accessibilité de nombreuses sources en ligne et à la culture du copier-coller. Cependant, une récente étude américaine publiée dans le Journal for 'Academic Ethics' révèle que le plagiat était aussi courant avant l'ère numérique qu'il ne l'est aujourd'hui. [4]

La détection manuelle du plagiat en utilisant des mots-clés sur les moteurs de recherche est souvent laborieuse et inefficace, surtout lorsque les passages plagiés proviennent de plusieurs sources différentes. Cette méthode demande du temps et des efforts considérables, soulignant ainsi la nécessité de développer des logiciels de détection de plagiat plus efficaces.

3) Importance du sujet :

La vérification du plagiat est cruciale non seulement dans les domaines académiques et éducatifs, mais aussi pour tous les créateurs de contenu, y compris les blogueurs, journalistes et artistes. Dans le domaine académique, les établissements d'enseignement évaluent les étudiants sur la base de leurs travaux et mémoires. Par conséquent, les étudiants doivent produire des travaux authentiques sans contenu plagié.

[2] Voir par exemple les sanctions selon la loi algérienne dans l'arrêté du journal officiel qui détermine les sanctions appliquées sur le plagiaire (c'est l'arrêté n° :=1082, le chapitre 4, la section 3 de l'année 2020).

[3]Dans ce lien, il y a des faits réels de plagiat qui ont causé des scandales et des suspensions d'emploi : https://www.acfas.ca/publications/magazine/2013/04/scandales-plagiat-universitaire

[4] https://www.lemonde.fr/campus/article/2015/08/28/internet-n-a-pas-augmente-le-plagiat-chez-les- etudiants_4739525_4401467.html

La détection du plagiat est également essentielle pour la recherche, car un article scientifique doit apporter de nouvelles perspectives sur un sujet donné. Si un article contient du contenu plagié, il perd sa crédibilité et sa valeur scientifique.

Pour les blogueurs et créateurs de contenu indépendants, il est important de produire un contenu original. Les lecteurs peuvent facilement repérer les parties copiées, ce qui peut entraîner une perte d'audience. Par conséquent, les créateurs de contenu doivent vérifier leurs œuvres avant publication en utilisant des outils de détection de plagiat.

Un avantage majeur de l'originalité est que les œuvres authentiques sont mieux classées par les moteurs de recherche, ce qui améliore leur visibilité et leur capacité à atteindre davantage de lecteurs. [5]

4) Objectif

L'objectif de notre travail est de développer et d'évaluer une méthode efficace pour la détection automatique du plagiat dans les textes arabes. Bien que de nombreux logiciels aient été conçus pour identifier le plagiat dans les documents en anglais, ces outils ne garantissent pas nécessairement le même niveau d'efficacité pour la langue arabe (Farghaly & Shaalan, 2009).

La plupart des recherches actuelles portent sur la détection de plagiat par des approches extrinsèques, tandis que les travaux axés sur les approches intrinsèques restent limités (Potthast, Stein, Barrón-Cedeño, & Rosso, 2011).Notre travail cherche à combler cette lacune en explorant l'utilisation d'algorithmes d'apprentissage automatique, en particulier les méthodes non supervisées.

L'objectif de ce projet est de développer un système de détection de plagiat pour les textes en langue arabe en utilisant l'architecture des réseaux de neurones convolutionnels **(**CNN**)** et les représentations sémantiques de l'Arabic WordNet (AWN)**.** Ce système vise à identifier efficacement les similarités textuelles, même en présence de reformulations, pour répondre au besoin croissant d'outils adaptés à la langue arabe dans le contexte de l'augmentation des contenus numériques.

Nous avons testé notre approche sur le corpus AraPlagDet, conçu spécifiquement pour évaluer les méthodes de détection intrinsèque du plagiat dans les textes arabes (Bensalem, et al., 2015).Notre étude vise à démontrer l'efficacité de cette méthode et à contribuer à l'avancement de la détection automatique du plagiat dans les textes arabes.

[5]Voir plus de détails sur ce sujet sur : Copyleaks. « why-plagairism-is-important»,https://copyleaks.com/fr/plagiarism-checker/why-plagairism-is-important

I. Chapitre 1 : État de l'art

Ce chapitre sert à introduire les concepts de base abordés dans ce mémoire. Nous mettons l'accent sur les notions de détection de plagiat, d'incorporation de mots, de technologies de transformateurs, et examinons quelques aspects de la langue arabe.

I.1 Phénomène de plagiat

Dans cette section, nous étudions le phénomène du plagiat en général, avec une attention particulière portée au plagiat textuel.

I.1.1 Définition de plagiat

Le plagiat est l'utilisation d'idées, de concepts, de mots ou de structures sans reconnaître de manière appropriée la source, dans un contexte où l'originalité est attendue (Fishman, 2009). Le plagiat englobe l'utilisation non autorisée ou l'imitation proche du langage, des idées ou des illustrations d'un auteur, considéré comme une œuvre originale. Cela inclut le vol littéraire, le fait de copier des paragraphes, des mots ou des idées d'autrui et de les présenter comme les siens sans en citer la source. Nombreux sont ceux qui voient le plagiat comme l'action de copier le travail ou d'emprunter les idées originales d'une autre personne (Abdelrahman, 2017).

I.1.2 Définition du plagiat textuel

Le plagiat textuel, également connu sous le nom de plagiat d'un texte, implique de s'approprier une œuvre. Cela peut consister à reproduire intégralement ou partiellement un texte sans mentionner la source, ou à résumer les idées d'un auteur dans ses propres termes. C'est aussi le cas de la traduction complète ou partielle d'un texte sans citation adéquate, ou de la présentation du travail d'une autre personne comme étant le sien, même si l'auteur a donné son accord. Le plagiat implique également l'utilisation d'idées déjà écrites (Maurer, Kappe, & Zaka, 2006)même en les exprimant avec ses propres mots.

I.1.3 L'histoire de la détection du plagiat

Au début, la détection du plagiat reposait sur la mémoire et l'attention des lecteurs, qui devaient identifier manuellement les ressemblances entre les textes. Cette méthode avait cependant des limites, car elle dépendait de la capacité des individus à se souvenir d'un grand nombre de textes, ce qui rendait la tâche particulièrement ardue et souvent inefficace. La découverte du plagiat

se faisait généralement quand un lecteur avait une impression de "déjà-vu" en lisant un passage familier.

Avec l'augmentation des contenus à analyser, les limites de cette approche manuelle sont devenues plus évidentes. Même les lecteurs les plus attentifs ne pouvaient plus suivre le rythme, et leur subjectivité ou fatigue rendaient le processus encore plus difficile. Il était donc nécessaire de créer des méthodes plus fiables et précises, ce qui a conduit à l'introduction d'outils automatisés et basés sur des statistiques pour comparer rapidement les textes.

En 1927, Bird a introduit l'une des premières méthodes automatisées, en utilisant des statistiques pour repérer les plagiats dans les tests à choix multiples. Cela a marqué le début d'une nouvelle ère. Aujourd'hui, les outils modernes de détection du plagiat sont capables de comparer des milliards de documents, de repérer même les paraphrases subtiles, et de protéger ainsi l'intégrité académique et littéraire avec une efficacité sans précédent.

Dans les années 1960, une nouvelle approche pour détecter le plagiat a été introduite, en se concentrant sur les tests à choix multiples. Cela a marqué le début des tentatives de rendre ce processus plus automatisé. Au début des années 1990, les premiers outils basés sur des méthodes statistiques ont vu le jour, permettant de comparer les textes et d'identifier les similitudes. La plupart de ces obensqutils étaient orientés vers la détection de plagiat dans les écrits, mais certains ont également été développés pour analyser le code source.

Au cours des dernières années, l'utilisation de ces systèmes a fortement augmenté, en grande partie à cause de l'explosion du nombre d'étudiants et de la demande croissante pour des outils de détection efficaces. En 2000, seulement cinq systèmes étaient bien établis pour détecter le plagiat, mais ce chiffre a rapidement grimpé à 47 en 2010, soulignant un besoin accru de protéger l'intégrité académique. Cependant, malgré cette prolifération, il restait encore des lacunes dans l'efficacité des outils disponibles.

Aujourd'hui, bien que les logiciels de détection de plagiat soient performants pour repérer les similitudes évidentes, ils peinent encore à détecter les changements subtils dans les textes, comme les paraphrases ou les réarrangements syntaxiques. Pour surmonter ces défis, il est essentiel de développer des outils capables d'analyser non seulement les mots, mais aussi le sens et la structure profonde des textes, un domaine encore en pleine exploration.

I.1.4 Les catégories de plagiat :

Les catégories de plagiat sont :

— **Accidentel** : en raison d'un manque de connaissances sur le plagiat et les styles de

citation ou de référence.

— **Non intentionnel** : l'abondance d'informations disponibles influence les pensées et conduit à des idées similaires exprimées comme les siennes.

— **Intentionnel** : copie délibérée tout ou partie du travail d'autrui sans citer la source.

— **Auto-plagiat** : utilisation d'un travail publié précédemment par soi-même sans référence à l'original.

I.1.5 Les méthodes de plagiat courantes :

Il y a une multitude de techniques de plagiat fréquemment employées. Quelques-unes de ces approches comprennent (Maurer, Kappe, & Zaka, 2006):

- **Copier-coller** : copie mot à mot de contenu textuel.
- **Plagiat d'idées** : appropriation de concepts ou d'opinions qui ne sont pas de notoriété publique.
- **Paraphrase** : modification de la grammaire ou des mots, réorganisation des phrases, ou reformulation du contenu original.
- **Plagiat Artistique** : présentation du travail d'autrui dans différents formats, tels que le texte, l'image, la voix ou la vidéo.
- **Plagiat de code** : utilisation de code de programme, d'algorithmes ou de fonctions sans autorisation ni citation appropriée.
- **Liens oubliés ou périmés** : utilisation de guillemets ou de références sans fournir des liens à jour vers les sources.
- **Usage incorrect des guillemets** : incapacité à identifier précisément les parties du contenu emprunté.
- **Références erronées** : ajout de références incorrectes ou inexistantes à des sources.
- **Plagiat traduit** : traduction d'un contenu multilingue sans citer l'œuvre originale.

I.2 Détection du plagiat :

Cette section traite des types de détection de plagiat intrinsèque et extrinsèque, ainsi que des mesures de similarité couramment utilisées dans ce domaine. Elle se termine par la présentation de quelques logiciels de détection de plagiat.

I.2.1 Détection de plagiat intrinsèque

Le plagiat intrinsèque consiste à détecter la possibilité de plagiat en examinant le document lui-même à la recherche de variations subtiles dans le style d'écriture (Alzahrani, Salim, & Abraham, 2012).

L'analyse intrinsèque du plagiat est étroitement liée à la vérification de l'authenticité de l'auteur : cette méthode vise à identifier la possibilité de plagiat en évaluant un document pour des changements subtils dans l'orthographe et d'autres caractéristiques textuelles (Eissen & Stein, 2006).

Une fonction fondamentale de cette approche consiste à transformer une séquence de mots et de ponctuations en un ensemble de caractéristiques qui reflètent le style de l'écrivain. Dans la littérature, diverses caractéristiques ont été suggérées en raison de la complexité de l'analyse du style d'écriture. Selon (Muhr, et al., 2010), il existe des fonctions de transformation qui reposent sur des parties du mot, comme les n-grammes de caractères. Selon (Stamatatos, 2009), la méthode utilisant les n-grammes de caractères a été suggérée. Dans le but d'analyser l'incohérence de style, elle compare les passages du document suspect représentés par les fréquences de caractères avec le modèle du document suspect lui-même. Cela permet de détecter les passages plagiés.

I.2.2 Détection de plagiat extrinsèque

La détection du plagiat extrinsèque consiste à comparer un document suspect à un corpus ou à un ensemble de sources de référence préexistantes. Cette collection de références peut être disponible en ligne ou hors ligne, par exemple, à partir de sources sur le web ou de bases de données hors ligne contenant des documents sources (Gupta & Awasthi, 2016).

Chaque document suspect est vérifié en le comparant aux sources disponibles pour déterminer s'il a été copié ou manipulé à partir de l'une de ces références (Figure 1).

La tâche de détection de plagiat extrinsèque est souvent divisée en deux sous-tâches : la collecte de documents sources candidats et la comparaison entre le document suspect en cours d'analyse et chacune des sources identifiées par la tâche de collecte.

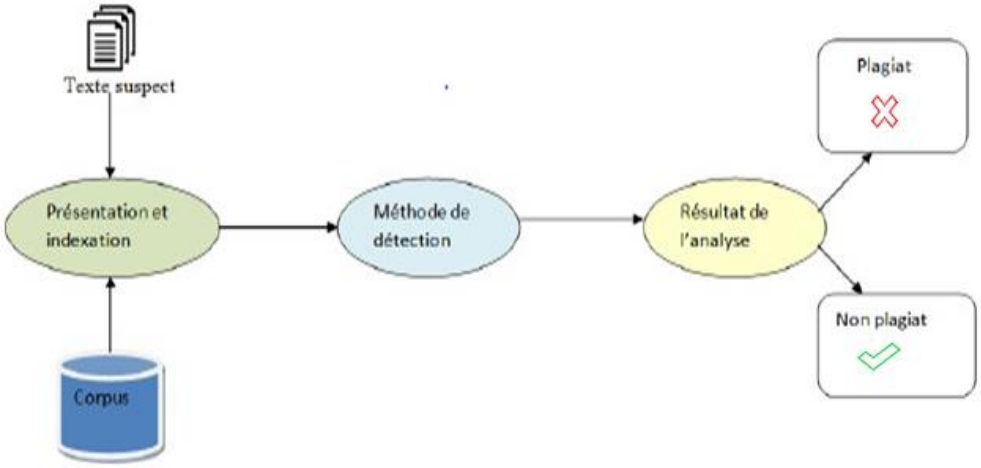


Figure 1: Architecture générale de la détection de plagiat extrinsèque

I.2.3 Différences entre la détection de plagiat extrinsèque et intrinsèque

Différences entre la détection de plagiat extrinsèque et intrinsèque :

Plagiat intrinsèque	**Plagiat extrinsèque**
Cette méthode n'a pas besoin de corpus externe car elle se base uniquement sur l'analyse interne du document suspect.	Cette méthode nécessite un corpus ou une base de données de documents de référence pour effectuer les comparaisons.

L'analyse se concentre sur des caractéristiques stylistiques internes du document, telles que le choix des mots, la longueur des phrases, la syntaxe, ou la structure des paragraphes.	La comparaison se fait entre le document suspect et les sources externes à l'aide de mesures de similarité textuelle ou d'autres techniques de comparaison.

Tableau 1:Différences entre la détection de plagiat extrinsèque et intrinsèque

I.3 Mesures de similarité :

La mesure de similarité est une technique utilisée pour évaluer à quel point deux objets, tels que des vecteurs, des ensembles de données, des images ou des séquences de texte, sont semblables. Elle est essentielle dans les domaines de l'apprentissage automatique, du traitement du langage naturel et de la vision par ordinateur.[6]

I.3.1 Distance euclidienne

La distance euclidienne est une mesure de la distance entre deux points dans l'espace euclidien. En deux dimensions, la distance euclidienne entre deux points (x1, y1) et (x2, y2) peut être calculée à l'aide du théorème de Pythagore :

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad 1\text{-}1$$

[6] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

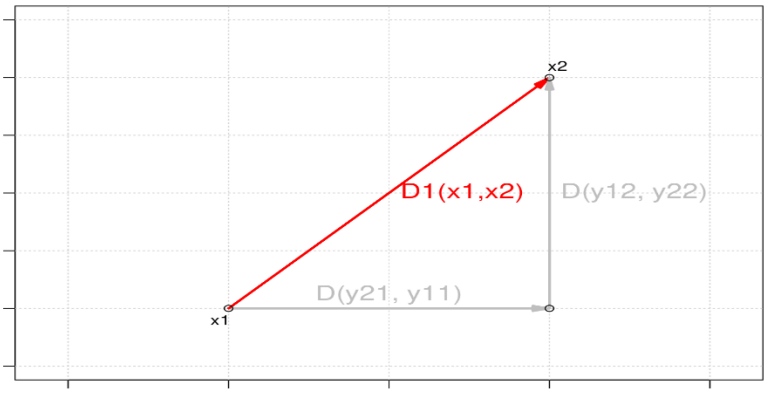


Figure 2: distance euclidienne entre deux points

La distance de Jaccard DJ mesure la « proximité » de deux mots, sans tenir compte de l'ordre des lettres.
La formule pour le calcul de la distance de Jaccard entre deux mots est : DJ = 1 – (nombre de lettres communes / nombre total de lettres distinctes).

Une distance de Jaccard est donc toujours comprise entre 0 et 1. Plus la distance est « proche de 0 », plus les mots sont « proches de l'anagramme ».

$$DJ(A,B) = 1 - \frac{A \cap B}{A \cup B} \quad \text{I-2}$$

Où $A \cap B$ nombre de lettres communes et $A \cap B$ nombre total de lettres distinctes.

I.3.2 Distance Manhattan

La distance de Manhattan est une mesure de distance entre deux points dans un espace n-dimensionnel. Il est calculé comme la somme des valeurs absolues des différences entre les coordonnées correspondantes des points.

Pour deux points $P=(p1,p2,\ldots,pn)$ et $Q=(q1,q2,\ldots,qn)$ dans un espace n-dimensionnel,

la distance de Manhattan est donnée par :

$$DM(P,Q) = \sum_{i=1}^{n} |p_i - q_i| \quad \text{1-3}$$

I.3.3 Similarité Cosinus

La similarité en cosinus est une mesure de similarité entre deux vecteurs non nuls d'un espace de produit intérieur qui mesure le cosinus de l'angle entre eux. La similarité cosinus entre les deux vecteurs est une représentation numérique de leur similarité ; elle va de -1, ce qui est complètement différent, à 1, ce qui est une correspondance exacte et la valeur de 0 des vecteurs indépendants (orthogonaux).

Le calcul de la similarité en cosinus est le suivant :

$$cos(A,B) = \frac{A.B}{\| A \| \| B \|} \qquad \text{I-4}$$

où :

- $A.B$ est le produit scalaire de A et B.
- $\| A \|$ est la norme (ou longueur) de A.
- $\| B \|$ est la norme (ou longueur) de B.

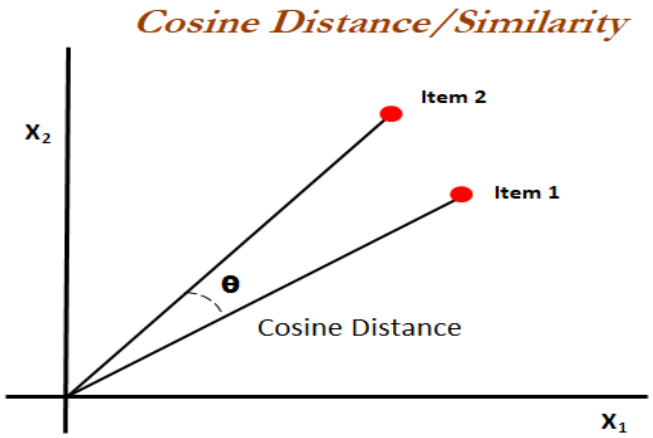


Figure 3: : Similarité cosinus entre deux vecteurs.

I.4 Incorporation des mots (Word Embedding)

L'incorporation de mots, ou Word Embedding, est une technique d'apprentissage automatique qui mappe les mots ou les phrases présents dans des données textuelles à un espace vectoriel, permettant ainsi de capturer les informations sémantiques et syntaxiques (Li & Yang, 2018). Les mots ayant des contextes similaires finissent par avoir des significations similaires. D'autres noms pour cette technique incluent Neural Embeddings ou Prediction-Based

Embeddings, et en français, elle est parfois appelée représentation vectorielle ou continue du mot (Bengio, Ducharme, & Vincent, 2000).

Le modèle d'incorporation de mots consiste à entraîner un réseau de neurones pour estimer la probabilité d'apparition d'un mot suivant, en s'appuyant sur la représentation continue des mots précédents. Cette représentation est apprise au fur et à mesure que le réseau de neurones s'entraîne. L'incorporation de mots permet de transformer le langage humain en une forme numérique. L'idée est que chaque mot peut être converti en un vecteur de N dimensions, et que des mots similaires se retrouvent avec des vecteurs plus proches les uns des autres (Balikas, 2017).

Plusieurs approches de l'incorporation de mots existent. Les premières approches, basées sur des techniques de réduction de la dimensionnalité, datent des années 1960 (Harris, 1954).Des méthodes plus récentes, telles que Word2Vec, s'appuient sur des modèles probabilistes et des réseaux de neurones pour améliorer les performances (Abbas & Hamdad, 2020).

I.4.1 Méthode Word2vec

Word2vec est une technique d'apprentissage automatique non supervisée qui produit des représentations distribuées de mots et de phrases dans un espace vectoriel à haute dimensionnalité

[Jansen, 2017]. Cette méthode s'appuie sur un réseau de neurones entraîné pour identifier les relations entre les éléments du langage et leur contexte. Les représentations vectorielles obtenues capturent ainsi les relations syntaxiques et sémantiques entre les mots et les phrases.

Word2vec offre deux architectures distinctes : le modèle CBOW (Continuos Bag of Word) et le modèle Skip-gram (voir Figure 4). Chacune de ces architectures a ses propres caractéristiques et avantages, mais toutes deux sont utilisées pour générer des vecteurs de mots qui préservent les relations linguistiques.

I.4.2 Continuos Bag-of-Word

Le modèle Continuos Bag-of-Word (CBOW) est une technique de word embedding utilisée pour prédire un mot à partir de son contexte. En d'autres termes, CBOW cherche à deviner un mot donné en se basant sur les termes qui l'entourent dans une phrase. Cette approche consiste à prendre les mots environnants comme entrée pour estimer le mot cible, en s'appuyant sur les relations de voisinage au sein du texte.

I.4.3 Skip-gram

Le modèle Skip-gram est une architecture conçue pour prédire les mots du contexte donné un mot en entrée. Contrairement au modèle CBOW, qui vise à prédire un mot central à partir de son contexte environnant, Skip-gram fait l'inverse : il utilise la représentation distribuée d'un mot d'entrée pour anticiper les mots de son contexte.

En pratique, les modèles CBOW tendent à apprendre plus rapidement, mais les modèles Skip-gram ont généralement de meilleures performances en termes de qualité de représentation des mots et des phrases (Mikolov, Chen, Corrado, & Dean, 2013).

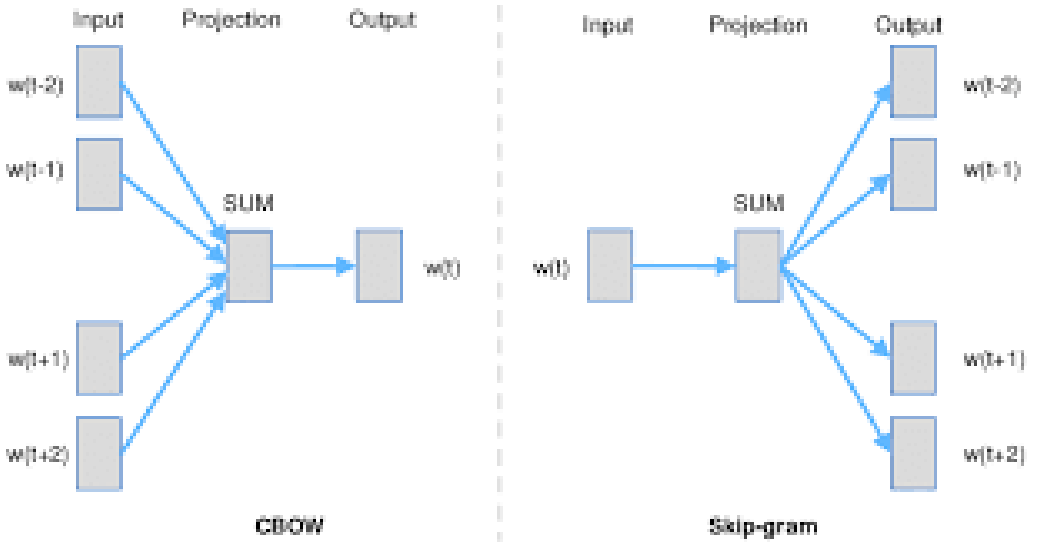


Figure 4: Modèles CBOW et Skip-gram [Mikolov et al., 2013]

I.5 Technologie des Transformateurs

Le modèle d'encodeur-décodeur basé sur un transformateur a été introduit par Vaswani et al. dans leur célèbre article intitulé "Attention is All You Need" (Vaswani, et al., 2017).Aujourd'hui, cette architecture est devenue la norme dans le traitement du langage naturel (NLP).

Le transformateur est le premier modèle de transformation à s'appuyer entièrement sur l'auto-attention pour calculer des représentations de son entrée et de sa sortie, sans recourir à des réseaux

de neurones récurrents (RNN) alignés sur la séquence ou à des convolutions (voir Figure 5). Cette approche innovante permet de traiter le langage de manière plus efficace et précise.

I.5.1 Encodeur

L'encodeur dans l'architecture de transformateur est constitué d'un empilement de N = 6 couches identiques. Chaque couche est composée de deux sous-couches principales :

1. Mécanisme d'auto-attention multi-tête : La première sous-couche met en œuvre un mécanisme d'auto-attention multi-têtes qui permet au modèle de se concentrer sur différentes parties de la séquence d'entrée simultanément. Cela permet d'apprendre les relations contextuelles entre les mots.
2. Réseau de rétroaction entièrement connecté : La deuxième sous-couche est un réseau feed-forward entièrement connecté qui applique une transformation linéaire aux entrées.

Chacune des deux sous-couches est entourée d'une connexion résiduelle, permettant à l'architecture d'éviter les problèmes de dégradation profonde en permettant à l'information de se propager plus facilement à travers les couches. Chaque sous-couche est également suivie d'une normalisation de couche (LayerNorm). Ainsi, la sortie de chaque sous-couche est calculée comme suit : **LayerNorm(x + Sublayer(x))**, où **Sublayer(x)** représente la sortie de la sous-couche.

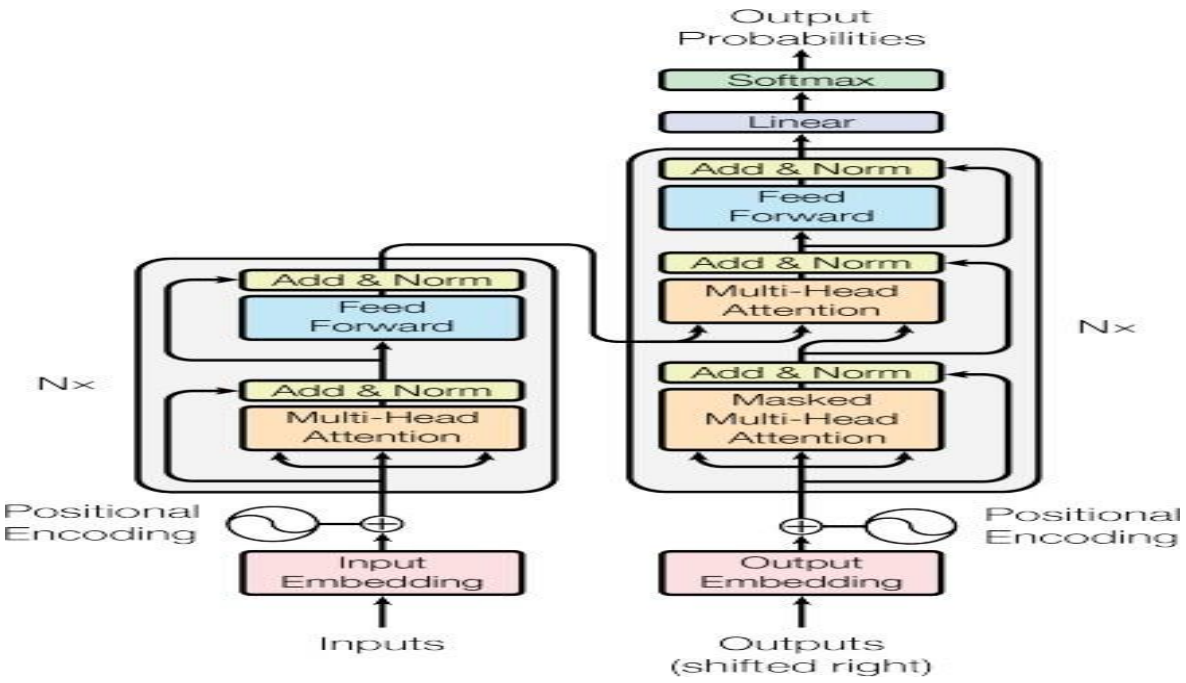


Figure 5: Architecture du modèle de transformateur

I.5.2 Décodeur

Le décodeur est constitué d'un empilement de N = 6 couches identiques. Contrairement à l'encodeur, chaque couche du décodeur comporte une troisième sous-couche, qui applique une attention multi-tête sur la sortie de l'encodeur.

Tout comme dans l'encodeur, des connexions résiduelles entourent chaque sous-couche du décodeur, suivies d'une normalisation de couche. Dans le décodeur, la sous-couche d'auto-attention est modifiée pour éviter que les positions futures ne soient prises en compte dans le calcul (Vaswani, et al., 2017).

I.5.3 Quelques logiciels pour la détection du plagiat

Voici quelques-uns des principaux logiciels de détection du plagiat largement utilisés dans les établissements d'enseignement, les lieux de travail et parmi les professionnels pour garantir le caractère unique des documents :

i. **Urkund :** Il s'agit d'un service Web qui effectue la détection du plagiat côté serveur. Il s'agit d'une solution intégrée et automatisée pour la détection du plagiat. Il s'agit d'un service payant qui utilise un système de courriel standard pour la présentation de documents et l'affichage des résultats. Ce système prétend traiter 300 demandes différentes. Les types de documents soumis et les recherches dans toutes les sources en ligne disponibles. Il donne plus de priorité aux sources éducatives de documents plus lors de la recherche. (Chowdhury & Bhattacharyya, 2018)

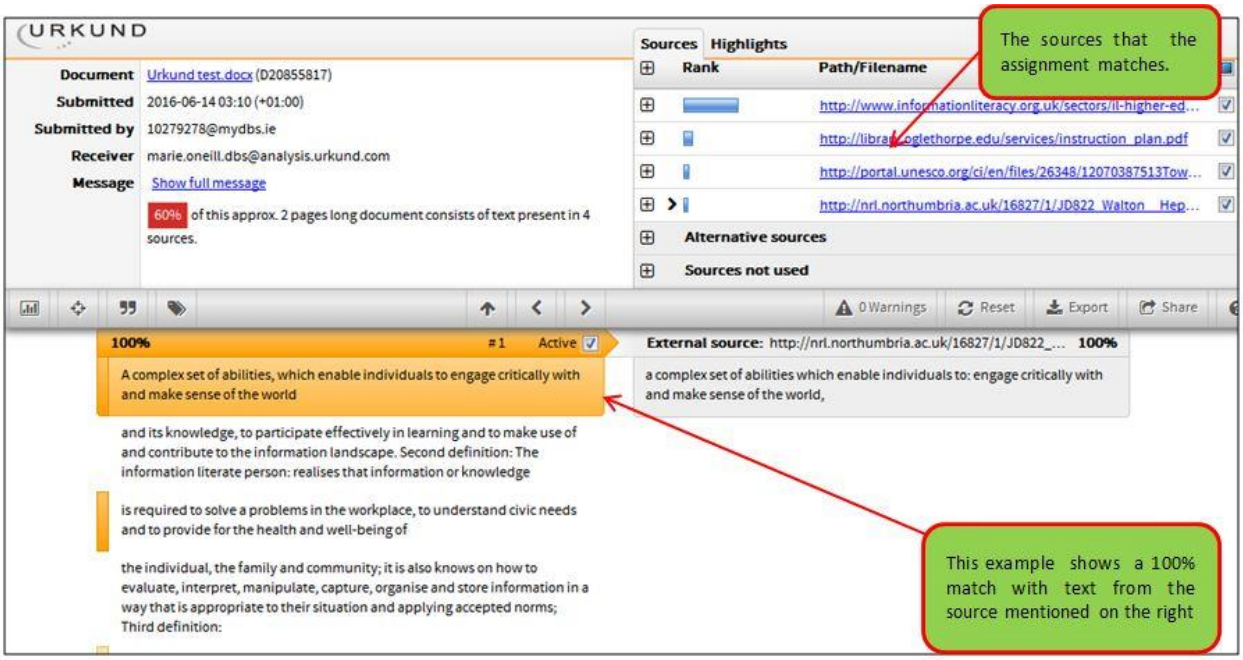


Figure 6:logiciel Urkund

ii. **Copycatch :** Il s'agit d'un outil client qui utilise la base de données locale de documents lors de la comparaison. Il offre des versions « gold » et « campus », offrant des capacités de comparaison avec un grand référentiel de ressources locales. Il a une autre version Web qui utilise les capacités de l'API Google pour la détection du plagiat à travers l'Internet. Pour utiliser la version Web, l'utilisateur a besoin d'une licence Google API personnelle via l'inscription. (Chowdhury & Bhattacharyya, 2018)

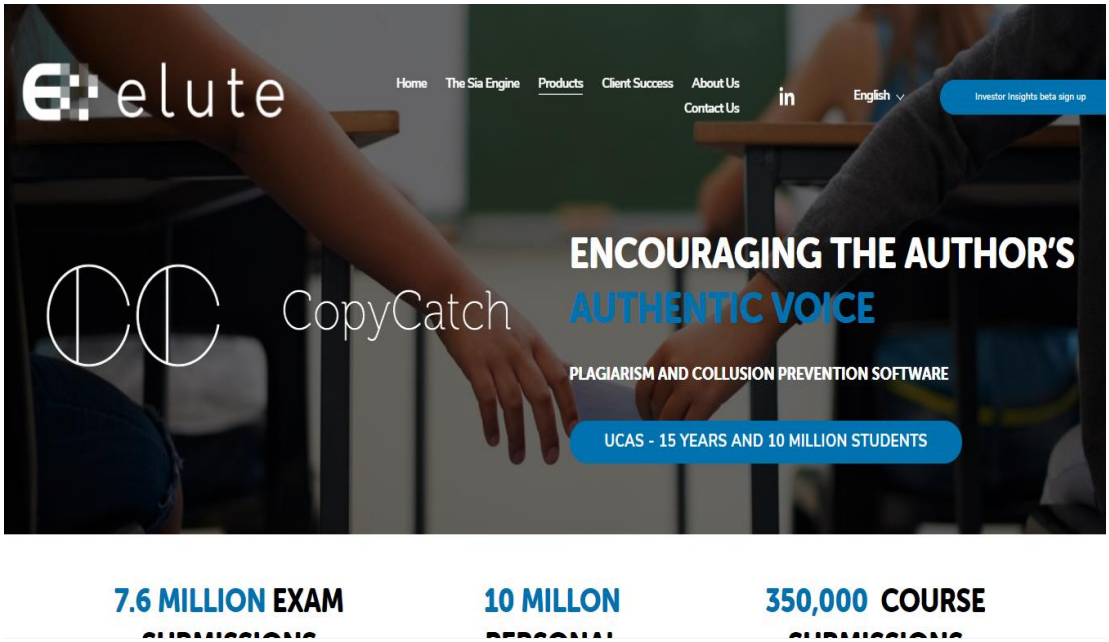


Figure 7: Logiciel Copycatch

iii. **Quetex:** Il utilise le traitement du langage naturel et l'apprentissage automatique pour détecter plagiat. Il effectue d'abord un contrôle interne du plagiat et ensuite il va pour vérification externe. Cet outil gratuit utilise tous les facteurs possibles pour chaque mot plagiat. Il fournit un support à plusieurs langues et on peut rechercher
Un nombre illimité de mots. Pour vérifier le plagiat avec cet outil, il suffit de copier et coller du document texte. Le principal inconvénient de cet outil est qu'il ne fournit pas rapport détaillé. Aussi, il n'est pas convivial. (Chowdhury & Bhattacharyya, 2018)

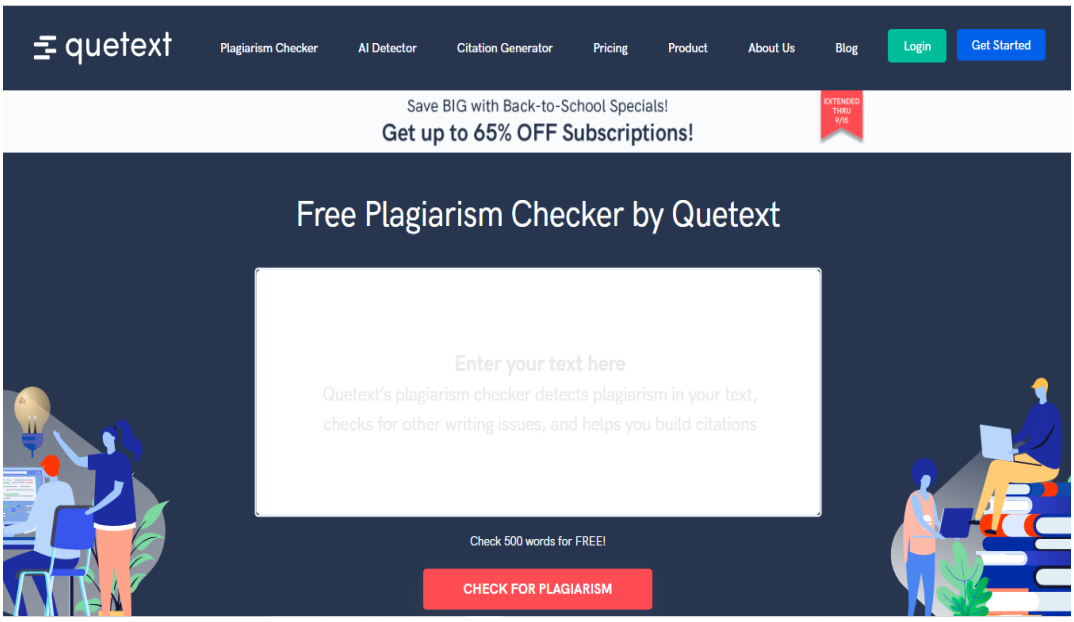


Figure 8: L'interface du logiciel "Quetext"

iv. **DupliChecker** : C'est un vérificateur de plagiat en ligne gratuit. Cet outil ne peut être consulté par un utilisateur non enregistré qu'une seule fois, mais l'utilisateur enregistré peut vérifier le plagiat 50 fois par jour. Le fichier d'entrée doit contenir plus de 1000 mots par recherche de similarité. L'utilisateur peut vérifier l'originalité du contenu par un certain nombre de moyens tels que copier-coller, télécharger un fichier ou en soumettant l'URL. (Chowdhury & Bhattacharyya, 2018)

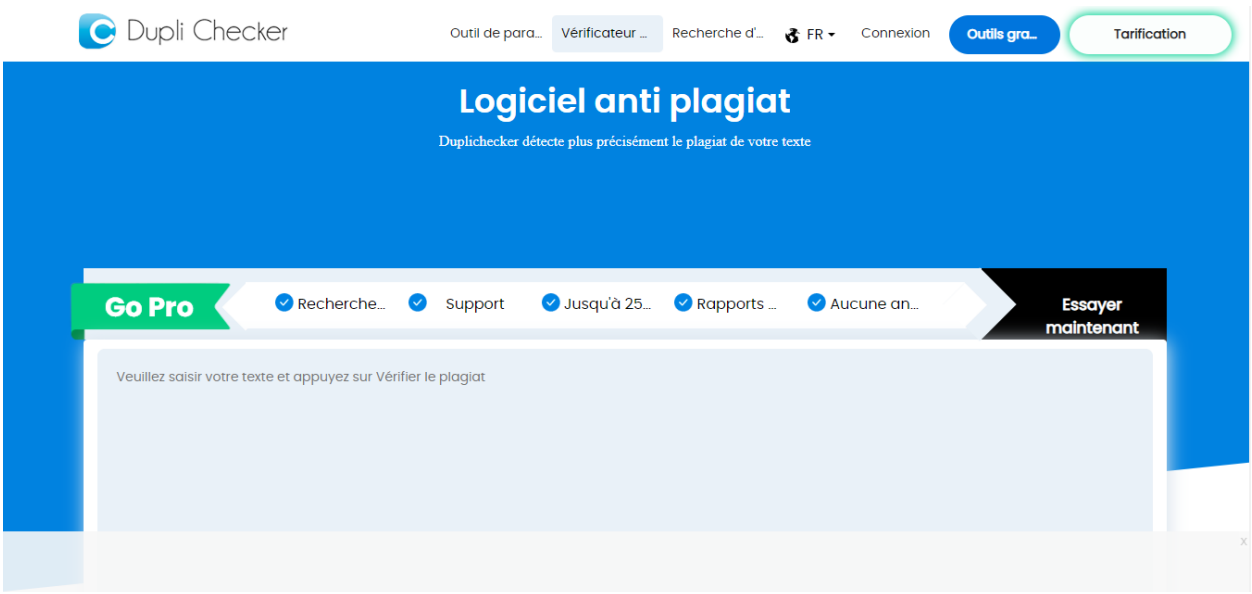


Figure 9: L'interface du logiciel "Duplichecker"

I.6 Définition des bases

I.6.1 Deep Learning ou apprentissage profond

C'est une technique de machine learning reposant sur le modèle des réseaux neurones : des dizaines voire des centaines de couches de neurones sont empilées pour apporter une plus grande complexité à l'établissement des règles.

I.6.2 Réseau neuronal convolutif

Dans le domaine de l'apprentissage profond, le CNN (Convolutional Neural Network) est largement reconnu comme l'algorithme le plus populaire et le plus utilisé. Conçu pour analyser des données structurées telles que des tableaux d'images, il représente un modèle avancé de programmation particulièrement efficace dans la reconnaissance visuelle, comme la classification d'images. Le CNN est appliqué dans divers domaines comme la vision par ordinateur, le traitement du langage naturel et la reconnaissance faciale, où il est devenu la norme pour de nombreuses applications.

Inspiré par les neurones du cerveau humain et des animaux, le CNN excelle à identifier des motifs dans les images, tels que des lignes, des formes géométriques, et même des caractéristiques complexes comme les visages et les yeux. Cette capacité en fait un outil puissant pour la vision par ordinateur. Contrairement aux méthodes précédentes en vision par ordinateur, les CNN peuvent travailler directement avec des images non traitées, sans besoin de prétraitement.

Les CNN sont composés de multiples couches convolutives empilées les unes sur les autres (voir Figure 6), chacune spécialisée dans la reconnaissance de motifs de plus en plus complexes.

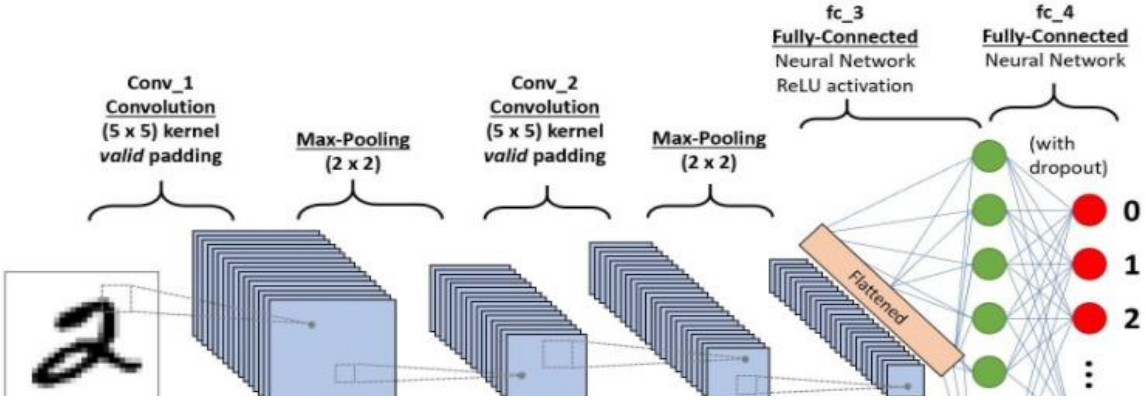


Figure 10 : Les différentes couches du CNN

Par exemple, quelques couches convolutives suffisent pour reconnaître des chiffres manuscrits, tandis qu'un réseau avec une vingtaine de couches peut identifier des visages humains. Ce modèle de couches superposées rappelle le processus de traitement visuel du cerveau humain, où les informations visuelles sont traitées progressivement pour identifier des caractéristiques de plus en plus abstraites.

I.6.3 Glove

GloVe est un modèle d'incorporation de mots basé sur l'apprentissage non supervisé qui a été proposé par (Pennington, Socher, & Manning, 2014). Ce modèle vise à créer des représentations vectorielles des mots, connues sous le nom de Global Vectors (GloVe). Contrairement à word2vec, qui se concentre sur les fenêtres de contexte locales, GloVe capture directement les statistiques du corpus global.

Une caractéristique distinctive de GloVe est son efficacité dans l'utilisation des statistiques en entraînant le modèle sur l'ensemble des cooccurrences mot à mot à l'échelle globale. Son processus se divise en deux étapes principales : d'abord, l'extraction de la matrice de cooccurrence X à partir du corpus d'apprentissage, où chaque élément Xij représente la fréquence d'occurrence des mots i et j ensemble ; puis, la construction des vecteurs de mots en factorisant cette matrice X. (Suleiman, 2018)

Cette approche permet à GloVe de capturer des relations sémantiques et syntaxiques entre les mots, rendant ses représentations particulièrement utiles dans diverses applications de traitement du langage naturel et de recherche d'informations.

I.6.4 TF-IDF

Le TF-IDF, ou terme frequency-inverse document frequency, est une méthode de pondération largement utilisée en recherche d'information, notamment dans l'analyse de texte. Cette approche statistique permet d'évaluer l'importance d'un terme au sein d'un document par rapport à l'ensemble d'une collection ou d'un corpus.

Plus précisément, le TF (terme frequency) mesure la fréquence d'apparition d'un mot dans un document spécifique. Un terme qui apparaît fréquemment dans un document est considéré comme important pour ce document. En revanche, l'IDF (inverse document frequency) évalue l'importance relative du terme dans l'ensemble du corpus. Les termes qui apparaissent fréquemment dans tout le corpus ont un IDF faible, tandis que ceux qui sont rares ont un IDF élevé.

En combinant TF et IDF, le poids d'un terme est calculé de manière à augmenter proportionnellement à sa fréquence dans le document tout en étant ajusté en fonction de sa fréquence dans l'ensemble du corpus. Ainsi, le TF-IDF permet de déterminer quels mots sont les plus représentatifs et discriminants pour un document spécifique par rapport à l'ensemble des documents traités.

I.7 Langue Arabe

La langue arabe, l'une des plus anciennes au monde, est reconnue comme la langue du Coran pour les musulmans et appartient au groupe des langues afro-asiatiques (Menai, 2012). Elle se distingue fortement des langues indo-européennes par ses spécificités linguistiques uniques.

Avec plus de 200 millions de locuteurs natifs et plus de 450 millions de personnes la parlant dans le monde (Zrigui, Ayadi, Zouaghi, & et Zrigui, 2016), l'arabe occupe la cinquième place parmi les langues les plus utilisées.

Cependant, la langue arabe présente plusieurs défis significatifs, tels qu'un vaste lexique et de nombreux synonymes (Zaher, Shehab, Elhoseny, & Osman, 2018).Chaque mot peut avoir plusieurs significations en fonction de sa position dans la phrase et de son diacritique, ce qui complique sa compréhension (Mahmoud, Zrigui, & Zrigui, 2018). Malgré cela, l'arabe demeure un sujet de recherche et d'expérimentation en raison de sa complexité morphologique et typographique (Meddeb, Maraoui, & Aljawarneh, 2016).

Les caractéristiques principales de la langue arabe incluent son appartenance au groupe sémitique, son écriture de droite à gauche, et son alphabet de vingt-huit lettres, comprenant trois voyelles longues et des consonnes (Menai, 2012). Une particularité notable est la variation de forme des lettres en fonction de leur position dans le mot (isolée, initiale, médiane ou finale).

Contrairement à d'autres alphabets, l'alphabet arabe ne comporte pas de lettres majuscules, simplifiant ainsi l'orthographe et la typographie. Cependant, les voyelles courtes sont indiquées par des signes diacritiques souvent omis dans les textes, ce qui peut rendre la compréhension du sens parfois complexe.

I.8 Le AWN (Arabic Word Net)

L'Arabic WordNet (AWN) est une base de données lexicale conçue pour la langue arabe, inspirée du WordNet développé à Princeton pour l'anglais. En gros, c'est un outil qui aide à mieux comprendre les relations entre les mots arabes, comme les synonymes, les contraires, et les mots qui appartiennent à la même catégorie. Cela facilite des applications pratiques comme la traduction automatique ou l'analyse sémantique.

L'idée derrière AWN, c'est de rendre l'arabe plus accessible pour le traitement automatique des langues (TAL), un domaine qui utilise des ordinateurs pour traiter les langues humaines. Comme beaucoup de ressources linguistiques sont principalement disponibles pour les langues comme l'anglais ou le français, l'AWN aide à combler ce manque pour l'arabe.

L'AWN a été développé dans le cadre de plusieurs projets de recherche internationaux, en collaboration avec des équipes travaillant sur le Global WordNet Association. Cette ressource permet non seulement d'organiser les mots en arabe en ensembles de synonymes (appelés "synsets"), mais elle offre aussi une meilleure compréhension de la langue pour les chercheurs, étudiants et développeurs d'applications qui travaillent avec l'arabe. Elle est particulièrement utile dans des domaines comme la traduction et l'intelligence artificielle, où la précision des mots est cruciale (Elkateb, et al., 2006)

Dr/terms	Example of Synset Corresponding	index choice
حدوث	{حَدَث, حُصُول, حُدُوث, ظُهُور, وُقُوع} {حُدُوث, حُصُول, حَادِثَة, حَدَث, وَاقِع}	حُصُول
استدعاء	{ذِكْرَى, اِسْتِدْعاء, تَذَكُّر} {اِسْتِدْعاء, طَلَب حُضُور}	تَذَكُّر
تذكر	{ذاكِرَة, تَذَكُّر} {ذِكْرَى, اِسْتِدْعاء, تَذَكُّر}	ذَاكِرَة
جاء	{أَتَى, جَاءَ} {جَاءَ, ظَهَرَ} {أتى,حَضَرَ,جَاءَ,قَدِم}	أَتَى
ذاكرة	{ذاكِرَة, فِكْر} {ذَاكِرَة,تَذَكُّر}	تَذَكُّر

Figure 11: Exemple de Le AWN

I.9 Conclusion

Le premier chapitre a introduit le phénomène du plagiat ainsi que les méthodes utilisées pour sa détection automatique, à savoir les approches extrinsèques et intrinsèques. Nous avons également exploré les différents logiciels disponibles dans ce domaine. Le chapitre a couvert les concepts clés liés aux incorporations de mots, aux technologies des transformateurs, et a abordé quelques aspects spécifiques de la langue arabe.

II. Chapitre 2 : Étude Expérimentale

II.1 Introduction

Le valet de la détection du plagiat dans les documents académiques et professionnels ne peut être surestimé. L'augmentation du nombre de sources numériques rend de plus en plus crucial le besoin de systèmes automatisés. Ce mémoire offre une étude expérimentale sur la détection de plagiat à l'aide de techniques de traitement automatique des langues naturelles et d'apprentissage en profondeur. L'objectif principal de cette étude est de construire et d'évaluer un système de détection de plagiat en utilisant un modèle de similarité textuelle via le modèle Convolution Neural Network (CNN).

II.2 Environnement de développement

II.2.1 Langage de Programmation

Python : C'est un langage de programmation open source multi-plateformes et orienté objet. Grâce à des bibliothèques spécialisées, Python s'utilise pour de nombreuses situations comme le développement logiciel, l'analyse de données, ou la gestion d'infrastructures.

II.2.2 Bibliothèques Python

- **os :** Utilisée pour l'interaction avec le système de fichiers, notamment pour charger les documents sources et suspects.

- **re :** Utilisée pour la manipulation de chaînes de caractères, notamment pour extraire des mots de textes.

- **math :** Utilisée pour des calculs mathématiques, notamment dans la fonction de similarité cosinus.

- **collections.Counter :** Utilisée pour compter la fréquence des mots dans un texte.

- **nltk :** Pour le tokenization et le traitement du langage naturel.
- **nltk.tokenize.sent_tokenize :** Utilisée pour segmenter les textes en phrases.
- **nltk.download('stopwords'):** télécharger le corpus de stopwords (mots vides) fourni par la bibliothèque Natural Language Toolkit (NLTK).
- **arabic_stopwords = set(stopwords.words('arabic')) :** utilisée pour charger et préparer une liste de stopwords spécifiques à la langue arabe, afin de les exploiter dans un processus de traitement de texte en arabe.
- **numpy :** Utilisée pour les calculs numériques et la manipulation de tableaux.
- **tensorflow et tensorflow.keras :** Utilisées pour construire et entraîner un modèle de réseau de neurones convolutif (CNN) pour la détection de plagiat.
- **json**: Pour la gestion de l'historique et l'enregistrement des résultats de détection.

II.2.3 Interface graphique

- **PyQt5** : Pour créer l'interface graphique de l'application (UI), y compris les éléments comme les boutons, les barres de progression, et les fenêtres de dialogue.
- **Matplotlib** : Pour l'affichage des graphiques circulaires montrant le pourcentage de plagiat et d'originalité.

II.2.4 VS Code

Idéal pour le développement Python avec gestion intégrée des environnements virtuels et des bibliothèques , utilise virtualenv pour isoler ton environnement de développement, gérer les dépendances et les versions de tes bibliothèques.

II.2.5 Structure de fichiers

- **Documents source:** Contient les fichiers .txt utilisés pour détecter le plagiat.
- **AWN (Arabic WordNet)** : Un fichier texte contenant des synonymes, utilisé pour l'expansion des textes avec des synonymes dans le cadre de la détection de plagiat.

- **Historique des résultats :** Un fichier « plagiarism_history. json » est utilisé pour stocker les résultats passés.

II.3 Dataset

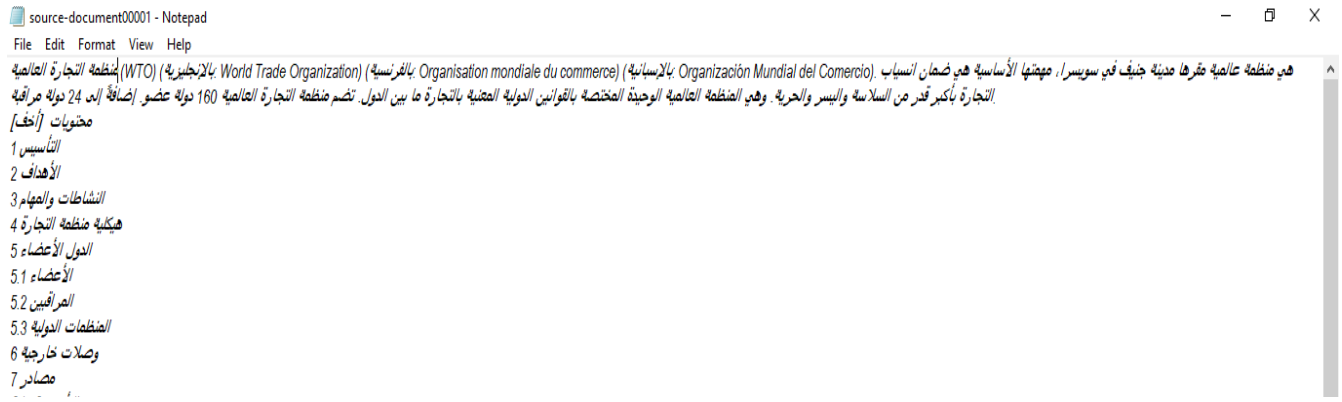
Pour notre évaluation nous avons utilisé le corpus ExAra (Bensalem, et al., 2015).

Récapitulation :

Le corpus ExAra comprend 2345 docs où près de la moitié d'entre eux- "documents suspicieux" incluent déjà des passages empruntés de l'autre moitié - "des documents source afin de simuler des documents contenant réellement des fragments plagiés. Le corpus comprend 2 parties : Training et test.

Description :

Chaque partie du corpus (training et test) se compose principalement de 3 jeux de données : 2 ensembles de fichiers textuels et 1 ensemble de fichiers XML. Les deux ensembles de fichiers textuels sont les documents suspects (c.-à-d. les documents qui contiennent du plagiat artificiel) et les documents sources (c.-à-d. les documents à partir desquels les passages suspects ont été plagiés). Le 3ème ensemble de documents contient des fichiers XML, qui sont l'annotation de plagiat, c'est-à-dire qu'ils fournissent pour chaque passage plagié son décalage de départ et sa longueur dans les documents suspects et sources (l'offset et la longueur ont tous deux étés exprimés en caractères). Un fichier de document suspect (.txt) et son fichier d'annotation de plagiat (.xml) partagent le même nom.


source-document00001 - Notepad

File Edit Format View Help

هي منظمة عالمية مقرها مدينة جنيف في سويسرا، مهمتها الأساسية هي ضمان انسياب (بالإنجليزية: World Trade Organization) (بالفرنسية: Organisation mondiale du commerce) (بالإسبانية: Organización Mundial del Comercio). المنظمة التجارة العالمية (WTO) التجارة بأكبر قدر من السلاسة والييسر والحرية. وهي المنظمة العالمية الوحيدة المختصة بالقوانين الدولية المعنية بالتجارة ما بين الدول. تضم منظمة التجارة العالمية 160 دولة عضو. إضافة إلى 24 دولة مراقبة

محتويات [أخف]

1 التأسيس

2 الأهداف

3 النشاطات والمهام

4 هيكلية منظمة التجارة

5 الدول الأعضاء

5.1 الأعضاء

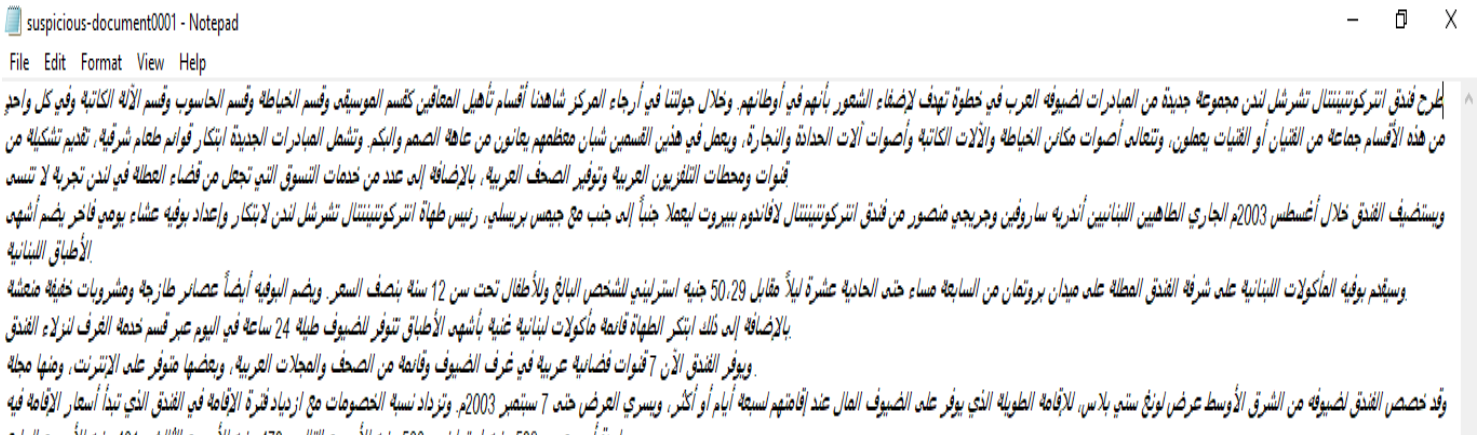
5.2 المراقبين

5.3 المنظمات الدولية

6 وصلات خارجية

7 مصادر

Figure 12 : Partie d'un document source


suspicious-document0001 - Notepad

File Edit Format View Help

يطرح فندق اتركوتينتال تشرشل لندن مجموعة جديدة من المبادرات لضيوفه العرب في خطوة تهدف لإضفاء الشعور بأنهم في أوطانهم. وخلال جولتنا في أرجاء المركز شاهدنا أقسام تأهيل المعاقين كقسم الموسيقى وقسم الخياطة وقسم الحاسوب وقسم الآلة الكاتبة وفي كل واحدٍ من هذه الأقسام جماعة من القيان أو القيات يعملون، وتتعالى أصوات مكائن الخياطة والآلات الكاتبة وأصوات آلات الحدادة والنجارة. ويعمل في هذين القسمين شبان معظمهم يعانون من عاهة الصمم والبكم. وتشمل المبادرات الجديدة ابتكار قوائم طعام شرقية، تقديم تشكيلة من قنوات ومحطات التلفزيون العربية وتوفير الصحف العربية، بالإضافة إلى عدد من خدمات التسوق التي تجعل من قضاء العطلة في لندن تجربة لا تنسى

ويستضيف الفندق خلال أغسطس 2003م الجاري الطاهيين اللبنانيين أندريه سارofين وجريجي منصور من فندق اتركوتينتال لافاندوم بيروت ليعملا جنباً إلى جنب مع جيمس بريسلي، رئيس طهاة اتركوتينتال تشرشل لندن لابتكار وإعداد بوفيه عشاء يومي فاخر يضم أشهى الأطباق اللبنانية.

وسيقدم بوفيه المأكولات اللبنانية على شرفة الفندق العطلة على ميدان بروتمان من السابعة مساء حتى الحادية عشرة ليلاً مقابل 50،29 جنيه استرليني للشخص البالغ ولأطفال تحت سن 12 سنة بنصف السعر. ويضم البوفيه أيضاً عصائر طازجة ومشروبات خفيفة منعشة بالإضافة إلى ذلك ابتكر الطهاة قائمة مأكولات لبنانية غنية بأشهى الأطباق تتوفر للضيوف طيلة 24 ساعة في اليوم عبر قسم خدمة الغرف لنزلاء الفندق.

ويوفر الفندق الآن 7 قنوات فضائية عربية في غرف الضيوف وقائمة من الصحف والمجلات العربية، وبعضها متوفر على الإنترنت، ومنها مجلة

وقد خصص الفندق لضيوفه من الشرق الأوسط عرض لونغ ستي بلاس، للإقامة الطويلة الذي يوفر على الضيوف المال عند إقامتهم لسبعة أيام أو أكثر، ويسري العرض حتى 7 سبتمبر 2003م. وتزداد نسبة الخصومات مع ازدياد فترة الإقامة في الفندق الذي تبدأ أسعار الإقامة فيه

Figure 13 : Partie d'un document texte suspect

La Figure 10 contient une annotation XML sous le nom "suspicious-document0012.xml" décrivant le document texte "suspicious-document0012.txt". En effet, il révèle l'absence du plagiat dans ce dernier.

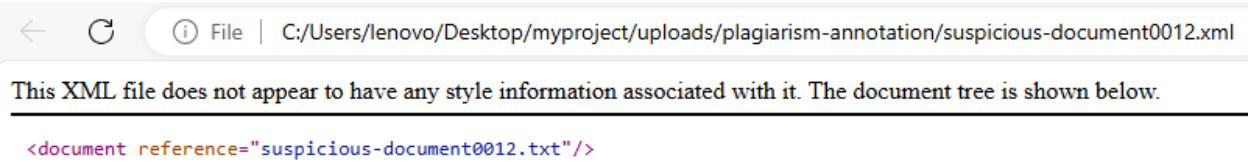


Figure 14 : Exemple de fichier XML révélant l'absence du plagiat dans le document "suspicious-document0012.txt"

Par contre, la Figure 11 montre un fichier XML qui indique que le fichier suspect "suspicious-document0001.txt" est plagié du document source "source-document00402.txt" avec le type du plagiat artificiel sans obscurcissement. Il est à préciser que la partie plagiée est de longueur 316, sa position dans le document source est 19259 et dans le document suspect est 119.

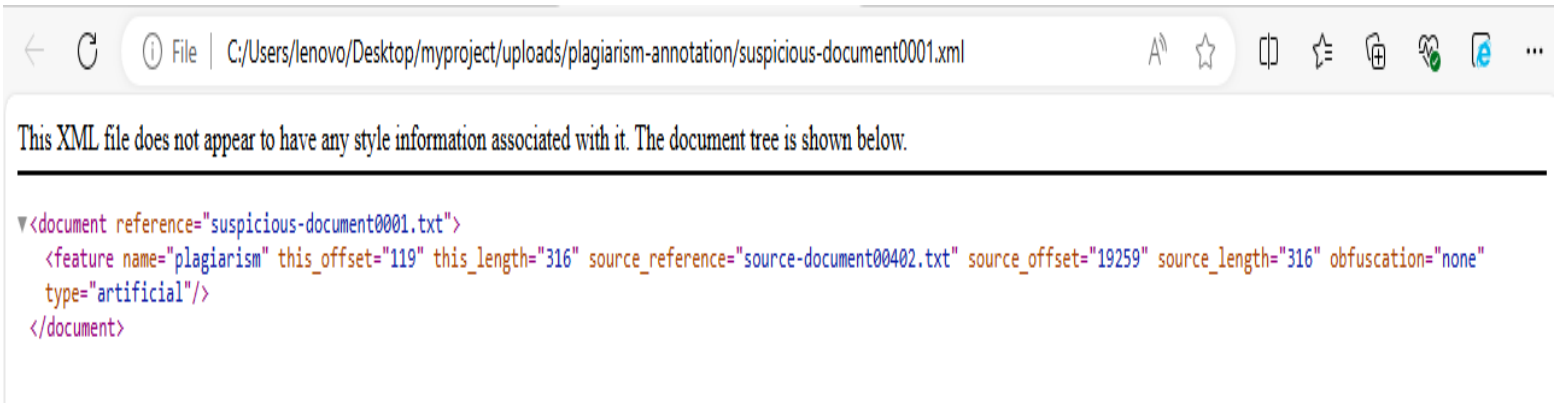


Figure 15 : Exemple de fichier XML indiquant l'existence du plagiat dans le document "suspicious-document0001.txt"

II.4 Modèle Proposé

Le modèle proposé est basé sur une architecture de Réseau de Neurones Convolutifs (CNN) adaptée pour le traitement du texte. L'architecture du modèle peut être divisée en plusieurs étapes :

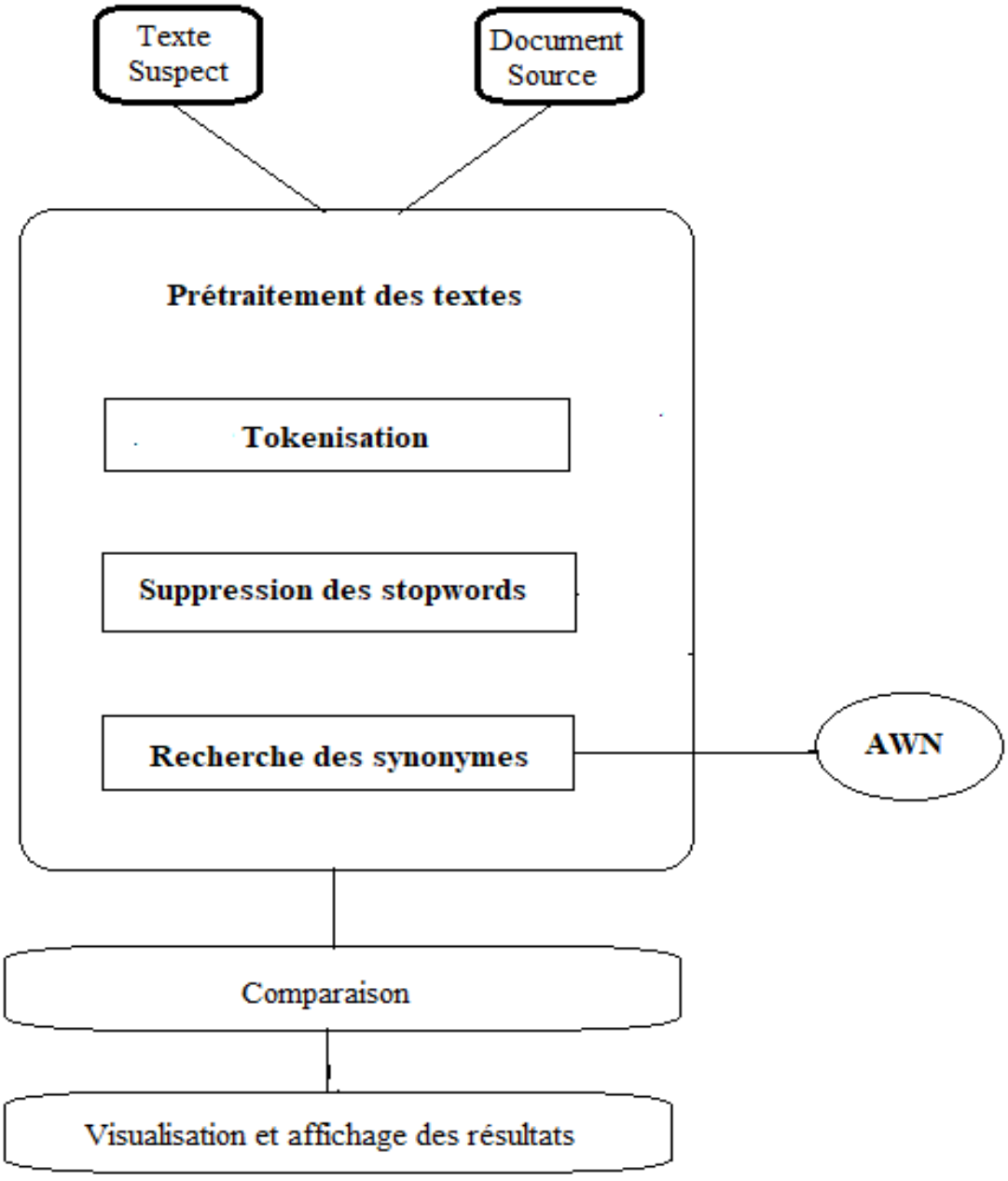


Figure 16:Architecture proposée pour le systeme

Prétraitement des données

L'objectif du prétraitement est de préparer les documents à être comparés en normalisant le texte et en le transformant en une représentation numérique exploitable.

Étapes du Prétraitement :

- **Tokenisation** :

  Décomposer le texte en tokens (mots individuels) à l'aide de bibliothèques comme NLTK ou spaCy.

- **Filtrage des mots vides (Stopwords)** :

  Supprimer les mots vides (comme "من", "إلى", "هي", "هذا"") qui n'apportent aucune information sémantique.

- **Expansion des synonymes** :

  Utiliser une base de données de synonymes comme **AWN** (Arabic WordNet) pour remplacer certains mots par leurs synonymes, ce qui permet de détecter le plagiat même lorsque des mots différents sont utilisés.

La figure 17 montre un phrase plagiée avant l'utilisation d'AWN

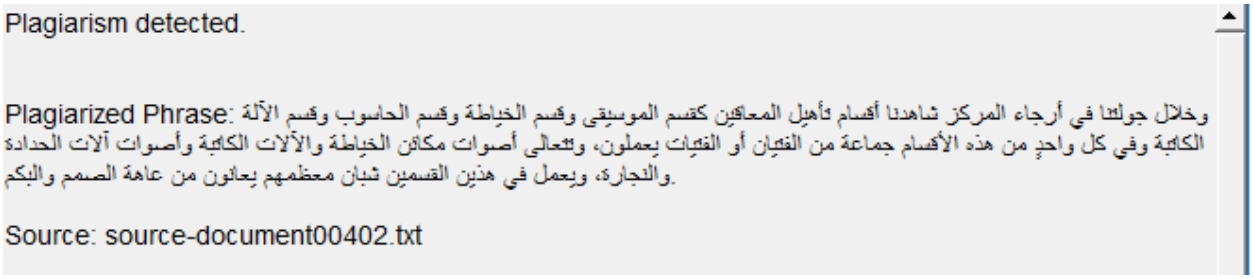


Figure 17:exemple de phrase plagiée avant l'utilisation d'AWN

Si on change le mot 'تأهيل' par le synonyme 'تهيئة' le système détecte toujours un plagiat comme le montre dans la figure 18.

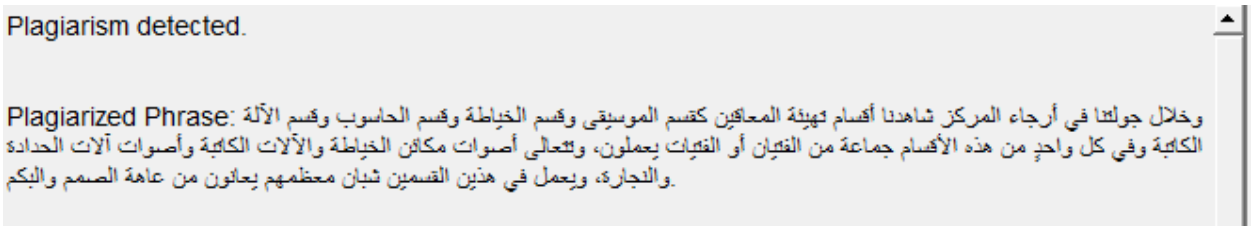


Figure 18::exemple de phrase plagiée après l'utilisation d'AWN

- **Vectorisation des textes** :

  Transformer le texte en vecteurs numériques à l'aide de techniques comme **TF-IDF**, **Word2Vec**, ou **embeddings** pré-entraînés (ex. **GloVe** ou **BERT**).

II.4.1 Comparaison des textes

Cette étape consiste à calculer la similarité entre le texte suspect et le texte de référence en utilisant des méthodes de similarité sémantique et syntaxique.

Techniques de Comparaison :

- **Similarité Cosinus** :

  Calcul de la similarité entre les vecteurs des deux textes en utilisant la **similarité cosinus**. Cela permet de mesurer le degré de proximité entre deux textes après leur vectorisation.

- **Modèle CNN (Convolutional Neural Network)** :

  Utiliser un modèle de réseau neuronal convolutionnel pour analyser les structures sémantiques profondes dans le texte. Ce modèle prend les séquences de texte comme entrée et génère des représentations vectorielles profondes permettant de mieux capturer la similarité entre deux phrases même si des synonymes ou des paraphrases sont utilisés.

- **Comparaison sémantique via expansion des synonymes** :

À partir du texte prétraité, l'application compare non seulement les mots exacts mais aussi les synonymes de ces mots pour détecter des similarités sémantiques même en présence de réécritures ou de reformulations.

- **Flux de Comparaison :**

1. Expansion des synonymes dans les deux textes.
2. Conversion en vecteurs numériques.
3. Calcul de la similarité cosinus pour mesurer la similarité textuelle.
4. Utilisation du modèle CNN pour affiner la détection de plagiat en capturant les relations sémantiques.

II.4.2 Évaluation et Visualisation

Une fois la comparaison effectuée, les résultats doivent être évalués et présentés à l'utilisateur.

II.4.2.1 Calcul de la similarité globale *:*

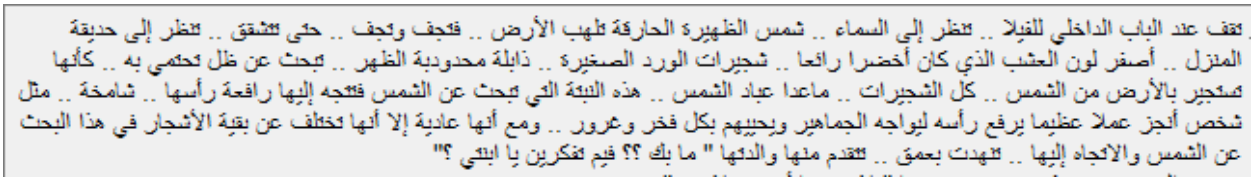
- Le score de similarité globale est calculé sur la base des similarités locales (au niveau des phrases ou des paragraphes).
- Un seuil de similarité peut être défini pour détecter automatiquement les zones suspectes.

II.4.2.2 Détection des zones suspectes *:*

- Identifier les phrases ou paragraphes spécifiques dans le texte suspect ayant un score de similarité élevé avec un texte de référence.

Le résultat de ce score peut être assimilé au taux de plagiat entre le texte source et le texte suspect. Nous considérons, expérimentalement, un seuil $\alpha = 0,65$ pour différencier entre plagiat et non plagiat. Si score $\geq \alpha$ alors le texte suspect est plagié.

Text 1:



Text 2:

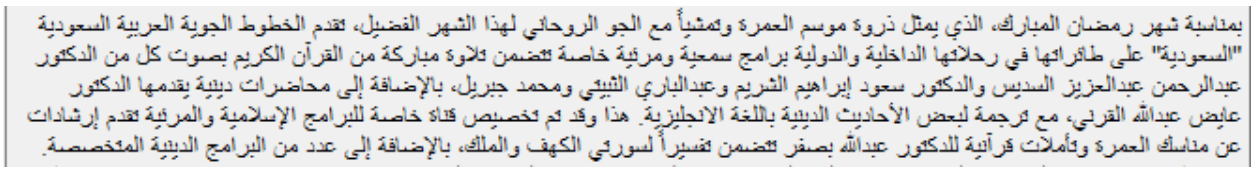


Figure 19: Séquence de texte non plagiées

Pour les deux séquences non plagiés de la Figure 19, notre modèle donne une mesure de similarité < 0.65, le premier texte le score est 0.28 et pour le deuxième c'est 0.34.

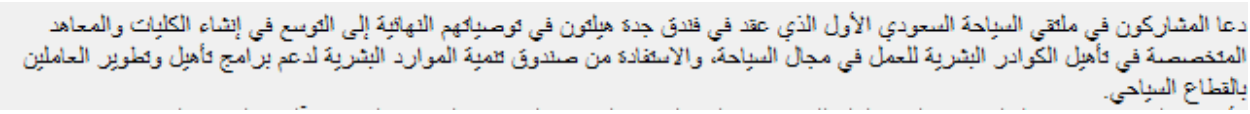

دعا المشاركون في ملتقى السياحة السعودي الأول الذي عقد في فندق جدة هيلتون في توصياتهم النهائية إلى التوسع في إنشاء الكليات والمعاهد المتخصصة في تأهيل الكوادر البشرية للعمل في مجال السياحة، والاستفادة من صندوق تنمية الموارد البشرية لدعم برامج تأهيل وتطوير العاملين بالقطاع السياحي.

Figure 20: Séquence de texte plagiées

Aussi, le modèle a calculé une similarité de 0,65 détecte non plagiat dans les séquences de la Figure 20.

À titre d'exemple, notre système a détecté le plagiat dans le document nommé « suspicious-document0001 », où la source de la phrase plagiée est le document « source-document00402.txt »

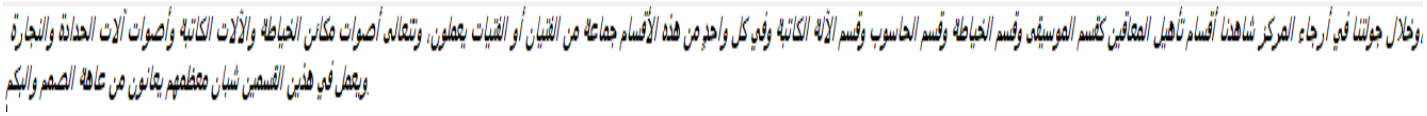

وخلال جولتنا في أرجاء المركز شاهدنا أقسام تأهيل المعاقين كقسم الموسيقى وقسم الخياطة وقسم الحاسوب وقسم الآلة الكاتبة وفي كل واحدٍ من هذه الأقسام جماعة من الفتيان أو الفتيات يعملون، وتتعالى أصوات مكائن الخياطة والآلات الكاتبة وأصوات آلات الحدادة والنجارة ويعمل في هذين القسمين شبان معظمهم يعانون من عاهة الصمم والبكم.

Figure 21: La phrase plagiée dans le suspicious-document0001

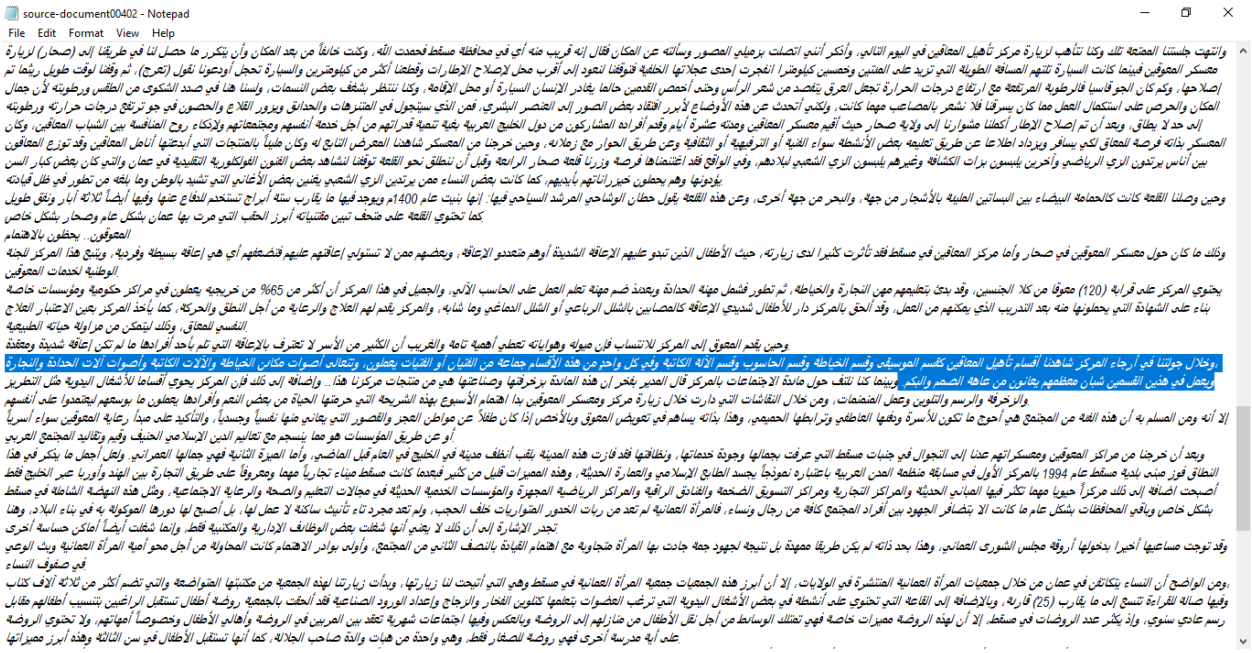


Figure 22: La phrase original dans le source-document00402

Génération de rapports :

- Générer un rapport détaillé pour l'utilisateur, qui montre les parties du texte suspect qui sont similaires à des sources existantes.
- Inclure des informations sur le pourcentage de plagiat détecté.

II.4.2.3 Visualisation graphique :

- Générer des graphiques pour visualiser les résultats, comme un **camembert** représentant la proportion de texte original et plagié.
- Une barre de progression ou des couleurs peuvent être utilisées pour indiquer les zones suspectes directement dans le texte.

II.5 Travaux Future

❖ **Rapports automatisés** : Développer un système de génération automatique de rapports détaillés, intégrant non seulement des pourcentages de plagiat, mais aussi des graphiques interactifs, des visualisations de similitudes textuelles, et des suggestions d'améliorations pour les utilisateurs.

- **Optimisation des temps de traitement** : Utiliser des techniques d'optimisation pour accélérer les performances du modèle sur de grandes quantités de texte.

II.6 Conclusion

Dans ce chapitre, nous avons présenté une approche dédiée à la détection du plagiat dans les textes arabes. Nous avons détaillé les différentes étapes de traitement, en commençant par le prétraitement des données, incluant notamment l'expansion textuelle via les synonymes et l'élimination des stopwords arabes. Ensuite, nous avons mis en œuvre un réseau de neurones convolutif (CNN) pour analyser les similarités sémantiques entre les documents. Ces étapes permettent de détecter des correspondances profondes, même lorsque le plagiat est masqué par des reformulations. Nous avons également intégré une interface conviviale pour rendre le système accessible à un public plus large.

Conclusion Générale

Dans ce travail, nous avons développé une application performante dédiée à la détection du plagiat en langue arabe, en exploitant des techniques telles que l'expansion textuelle via les synonymes et des méthodes d'analyse basées sur la similarité cosinus, couplées à l'apprentissage profond. En intégrant des ressources comme Arabic WordNet pour identifier les synonymes et en prenant en compte les stopwords propres à la langue arabe, nous avons réussi à améliorer la précision du système, notamment face aux tentatives de dissimulation du plagiat par des reformulations ou l'utilisation de termes similaires.

Le recours à un modèle de réseau de neurones convolutif (CNN) a permis d'entraîner le système sur des corpus variés, en mettant l'accent sur les similarités sémantiques profondes entre les textes suspects et leurs sources potentielles. L'application a été conçue pour être facile à utiliser, avec une interface intuitive qui rend la détection du plagiat accessible même pour les personnes non spécialisées en informatique. De plus, nous avons enrichi l'outil avec des fonctionnalités comme la gestion des historiques et des options de personnalisation visuelle, telles que la possibilité de choisir des thèmes ou de visualiser les résultats sous forme de graphiques interactifs. Cela permet une meilleure compréhension des résultats obtenus, ainsi qu'une traçabilité des analyses effectuées.

Néanmoins, notre projet peut encore évoluer. Par exemple, il serait intéressant d'élargir la base de synonymes arabes pour affiner davantage la détection. De plus, l'intégration de techniques d'apprentissage non supervisé permettrait de traiter de nouveaux textes sans avoir besoin d'un corpus de référence prédéfini. Il serait aussi pertinent d'étendre l'application à d'autres langues, et d'améliorer la détection de paraphrases pour traiter des formes de plagiat encore plus subtiles.

En résumé, cette étude offre des perspectives prometteuses dans le domaine du traitement automatique des langues (TAL) et de la détection du plagiat, tout en proposant un outil pratique et accessible, particulièrement utile dans le milieu académique et pour les chercheurs.

Bibliographie

Abbas, J., & Hamdad, A. (2020). *Apprentissage automatique du dialecte algérien.*

Abdelrahman, Y. A. (2017). A method for Arabic documents plagiarism detection. *International Journal of Computer Science and Information Security*, 79.

Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2)*, 133–149.

Balikas, G. (2017). *Mining and learning from multilingual text collections using topic models and word embeddings.* Grenoble 1 UGA-Université Grenoble Alpes.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*.

Bensalem, I., Boukhalfa, I., Rosso, P., Abouenour, L., Darwish, K., & Chikhi, S. (2015). Overview of the AraPlagDet PAN@FIRE2015 Shared Task on Arabic Plagiarism Detection. *FIRE workshops*, (pp. 111–122).

Bourgeois, B., Giroux-Bougard, X., Winegardner, A., Chrétien, E., Granados, M., & Braga., P. H. (2021, octobre). *types-de-coefficients-de-distance.* Récupéré sur r.qcbs.ca: https://r.qcbs.ca/workshop09/book-fr/types-de-coefficients-de-distance.html

Chowdhury, A., & Bhattacharyya, D. K. (2018). *arXiv preprint arXiv :1801.06323.* Récupéré sur Plagiarism :Taxonomy, tools and detection techniques: https://arxiv.org/pdf/1801.06323

Eissen, S., & Stein, B. (2006). Intrinsic plagiarism detection., *3936*, pp. 565–569.

Elkateb, S., Black, W., Rodriguez, H., Pease, A., Alkhalifa, M., Vossen, P., & Fellbaum, C. (2006). Building a WordNet for Arabic. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).* Gênes, Italie: European Language Resources Association (ELRA).

Farghaly, A., & Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*.

Fishman, T. (2009). *"We know it when we see it" is not good enough: Toward a standard definition of plagiarism that transcends theft, fraud, and copyright.*

Gupta, D., & Awasthi, R. (2016). Study on extrinsic text plagiarism detection techniques and tools. *Journal of Engineering Science & Technology Review*.

Harris, Z. S. (1954). Distributional structure. *word*, 146–162.

Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: a survey. Dans Springer, *Guide to big data applications* (pp. 83–104).

Mahmoud, A., Zrigui, A., & Zrigui, M. (2018). A text semantic similarity approach for Arabic paraphrase detection. Dans A. Gelbukh (Éd.), *Computational Linguistics and Intelligent Text Processing*, (pp. 338–349).

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism-a survey. *Journal of Universal Computer Science, 12(8)*, 1050–1084.

Meddeb, O., Maraoui, M., & Aljawarneh, S. (2016). Hybrid modeling of an offline Arabic handwriting recognition system AHRS., (pp. 1–8).

Menai, M. E. (2012). Detection of plagiarism in Arabic documents. *International Journal of Information Technology and Computer Science, 10(10)*, 80-89.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *arXiv preprint.* Récupéré sur Efficient estimation of word representations in vector space: http://arxiv.org/abs/1301.3781

Muhr, Markus, Kern, Roman, Zechner, Mario, . . . Michael. (2010). External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System. *CLEF 2010 Conference on Multilingual and Multimodal Information Access Evaluation.* Italy.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove : Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. *CLEF 2009* (pp. 1-9). CEUR proceedings.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. *CLEF 2011 Evaluation Labs and Workshops – Working Notes papers.*

Stamatatos, E. (2009). Intrinsic Plagiarism Detection Using Character n-gram Profiles. Dans P. R. Benno Stein (Éd.), *Proceedings of the SEPLN'09 workshop on uncovering plagiarism, authorship and social software misuse (PAN 09)* (pp. 38–46). CEUR Workshop Proceedings (CEURWS.org).

Suleiman, D. e. (2018). Comparative study of word embeddings models and their usage in Arabic language applications. *2018 International Arab Conference on Information Technology (ACIT)* (pp. 1-7). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems.*

Zaher, M., Shehab, A., Elhoseny, M., & Osman, L. (2018). A new model for detecting similarity in Arabic documents., (pp. 488–499).

Zrigui, S., Ayadi, R., Zouaghi, A., & et Zrigui, S. (2016). Isao: An intelligent system of opinions analysis. *Research in Computer Science, 110*, 21–30.

Résumé

Le développement technologique et la prolifération des données ont exacerbé le problème du plagiat, notamment avec l'augmentation de l'utilisation de la langue arabe sur Internet. Dans ce contexte, nous avons développé un système de détection du plagiat pour les textes arabes, basé sur l'architecture des réseaux de neurones convolutionnels (CNN). Ce système analyse les motifs linguistiques tout en intégrant des représentations issues de l'Arabic WordNet (AWN). Nous avons évalué notre modèle en utilisant le corpus ExAra, et les résultats préliminaires montrent son efficacité, avec un potentiel d'amélioration grâce à des données supplémentaires et un ajustement du modèle.

Mots-clés : Détection du plagiat, Textes arabes, AWN, ExAra, CNN.

Abstract

Technological development and the proliferation of data have exacerbated the problem of plagiarism, particularly with the increased use of the Arabic language on the Internet. In this context, we developed a plagiarism detection system for Arabic texts based on the convolutional neural network architecture (CNN). This system analyses linguistic patterns while integrating representations from the Arabic WordNet (AWN). We evaluated our model using the ExAra corpus, and preliminary results show its effectiveness, with potential for improvement through additional data and model adjustment.

Keywords: Plagiarism detection, Arabic texts, AWN, ExAra, CNN.

ملخص

أدى التطور التكنولوجي والزيادة الهائلة في البيانات إلى تفاقم مشكلة الانتحال، خاصةً مع تزايد استخدام اللغة العربية على الإنترنت. في هذا السياق، قمنا بتطوير نظام للكشف عن الانتحال في النصوص العربية باستخدام بنية الشبكة العصبية التلافيفية .(CNN) يعتمد نظامنا على تحليل الأنماط اللغوية مع دمج التمثيلات من WordNet العربية(AWN).استخدمنا مجموعة بيانات ExAra لتقييم النظام، وأظهرت النتائج الأولية فعاليته مع إمكانية تحسينه بمزيد من البيانات وتحسين النموذج.

الكلمات المفتاحية: اكتشاف الانتحال، النصوص العربية، AWN، ExAra، CNN.