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

République Algérienne Démocratique et PopulaireUniversité

Abou Bakr Belkaid-Tlemcen

Faculté des Sciences

Département d'informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option : Système d'Information et de connaissance (SIC)

Thème

« TTDetect »
Détection automatique de contenus générés par IA dans les textes français.

Soutenue le **29/06/2025**

Par

Taibi Mohammed Ibrahim

Tabti Ayoub

Présenté le devant :

-Mm. El Yebdri Zeyneb (Encadrante)

-Mm. Bekkouche Amina (Président)

-Mm. Halfaoui Amel (Examinateur)

Année universitaire : 2024-2025

Dédicace

À ma chère mère, pilier de ma vie, pour son amour inconditionnel, ses prières silencieuses, sa tendresse et sa force.

À mon père, modèle de sagesse et de persévérance, dont les sacrifices et les conseils m'ont toujours guidé dans les moments décisifs.

À mes frères et sœurs, pour leur soutien, leur humour et leur présences bienveillante, toujours là pour me motiver et me rappeler de ne jamais baisser les bras.

À mon binôme et ami Ayoub, pour sa complicité, son engagement et son esprit d'équipe tout au long de cette aventure. Ce travail est aussi le reflet d'un réel partenariat fondé sur la confiance et le respect mutuel.

À tous mes amis de la promotion, avec qui j'ai partagé rires, stress, entraide et souvenirs mémorables. Merci pour ces années intenses et inoubliables. Et tout particulièrement au groupe Chabiba, pour leur énergie, leur solidarité, leur joie de vivre et cette fraternité unique qui a rendu ce parcours plus riche humainement

Taibi Mohammed Ibrahim

Dédicace

À ma famille Je tiens à exprimer toute ma gratitude, pour son amour inconditionnel, son soutien moral et sa confiance en moi. Vous avez toujours cru en mes capacités et m'avez encouragé dans chaque étape de ma vie.

À mes frères et sœurs, pour leur soutien, leur compréhension et leurs conseils précieux. Vous avez toujours su trouver les mots justes pour me motiver et me pousser à donner le meilleur de moi-même.

À mes amis Un grand merci pour leur présence, leur encouragement et les moments de joie partagés. Votre amitié a été une source de motivation et de réconfort tout au long de ce parcours.

À mon binôme, Taibi Mohamed Ibrahim pour sa collaboration, son engagement et son esprit d'équipe. Ton professionnalisme et ta persévérance ont grandement contribué à la réussite de ce projet

Tabti Ayoub

Remerciement

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

Nous exprimons notre plus profonde gratitude à Madame Zineb El-Yebdri, notre encadrante, pour son accompagnement rigoureux, sa disponibilité constante, ses conseils précieux et sa bienveillance tout au long de la réalisation de ce projet. Son expertise et son engagement ont été essentiels à l'aboutissement de ce travail, et nous lui en sommes sincèrement reconnaissants.

Nos remerciements les plus sincères vont également aux membres du jury, pour l'honneur qu'ils nous font en acceptant d'évaluer notre travail. Leur regard critique, leurs remarques constructives et leur présence bienveillante témoignent de leur engagement envers la qualité de la formation.

Nous tenons aussi à remercier l'ensemble des enseignants du département d'informatique, qui ont su, au fil des années, nous transmettre leur savoir, leur passion et leurs valeurs avec dévouement

et professionnalisme. Grâce à eux, nous avons pu acquérir les compétences nécessaires pour mener à bien ce projet.

Enfin, nous adressons notre reconnaissance à tous nos amis, pour leur soutien moral, leur entraide, leurs encouragements dans les moments difficiles, et les nombreux souvenirs partagés durant ce parcours. Leur présence a rendu cette aventure universitaire plus humaine, plus riche et plus agréable.

Table des matières

Table de matières

Liste des figures

Liste des tableaux

Liste des abréviations

- **IA** : Intelligence Artificielle
- **NLP** : Natural Language Processing (Traitement du Langage Naturel)
- **GPT** : Generative Pre-trained Transformer
- **CNN** : Convolutional Neural Network (Réseau de Neurones Convolutifs)
- **SVM** : Support Vector Machine (Machine à Vecteurs de Support)
- **BERT** : Bidirectional Encoder Representations from Transformers
- **RoBERTa** : Robustly optimized BERT approach
- **T5** : Text-To-Text Transfer Transformer
- **XLM-Roberta** : Cross-lingual Language Model - Roberta
- **TF-IDF** : Term Frequency - Inverse Document Frequency
- **GloVe** : Global Vectors for Word Representation
- **ML** : Machine Learning (Apprentissage Automatique)
- **DDoS** : Distributed Denial of Service (Attaque par Déni de Service Distribué)
- **DINUM** : Direction Interministérielle du Numérique
- **ANSSI** : Agence Nationale de la Sécurité des Systèmes d'Information
- **UI** : User Interface (Interface Utilisateur)
- **AUC** : Area Under the Curve
- **ROC** : Receiver Operating Characteristic

Introduction Générale

Introduction Générale

1) Contexte

Nous parlons souvent de la manière dont l'intelligence artificielle (IA) bouleverse la gérance de projets au 21e siècle. En effet, le développement de solutions comme GPT-4, Bard ou Mistral a complètement révolutionné le tournage d'articles, d'essais et même de poèmes en français. Malgré le fait que ces outils sont adéquats pour optimiser le fonctionnement de tâches automatisées, ils créent un challenge sans précédent : la séparation des textes écrits par l'homme et ceux qui sont synthétiques.

2) Problématique

L'utilisation sans autorisation de textes créés par IA dans le milieu de l'éducation est devenue extrêmement répandu, avec des étudiants qui soumettent continuellement des thèses élaborées par des algorithmes, donc la balance académique est menacée. Prenons par exemple une étude récente menée par l'Université de Paris où un impressionnant 20% des mémoires de master ont été truffés de passages douteux, voire directement générés par des IA. Dans le domaine médiatique, le nombre d'articles faux qui ont été développés avec l'aide de l'IA et utilisés comme des fausses informations et de la désinformation augmente chaque jour.

Néanmoins, l'amélioration des programmes pour reconnaître le texte en français créé par l'IA soulève graduellement des questions sur l'honnêteté intellectuelle, le mandat médiatique et la confiance dans les produits numériques créés.

3) Motivation

L'intelligence artificielle a pris le pas sur la création d'outils permettant notamment la génération de texte à une telle échelle qu'on ne pouvait même pas en envisager il y a quelques années de cela. Ces nouveaux outils, bien qu'ils aient des limites, peuvent écrire en français avec une qualité presque divine. L'atteinte de tous les domaines par cette technologie révolutionnaire permet la création de nouvelles portes. Cependant, il est plus qu'évident que le phénomène a ses froids, et le degré de satisfaction de l'IA vis-à-vis de la transparence et de l'authenticité des informations produites est inquiétant.

La quantité de contenus produits par des ordinateurs tels que des articles, des posts sur les réseaux sociaux, ou même des travaux scolaires propose un tout nouveau défi de légitimité. Quelle est la méthode employée pour savoir si un texte a été écrit par un homme ou par une machine ? Cette problématique est à surveiller parce qu'elle touche un sujet très sensible comme la désinformation et les usages malintentionnés de l'information.

Si l'intention est de filtrer manuellement les données pour repérer les contenus générés par l'IA, il est sans aucun doute qu'ils ne dépasseront pas une certaine limite, principalement à cause des écueils d'une prise en charge. Il existe également le fait que les systèmes d'IA évoluent sans cesse, rendant plus ardues les tentatives de repérage.

4) Importance du sujet

Les étudiants sont notés dans une institution par leur cadre de recherche et publication d'articles originaux. Mais avec l'arrivée des AI génératives, des étudiants se mettent à soumettre dissertations ou mémoires qui ont été entièrement composées par une machine sans se soucier pour la justice. Par exemple, une enquête dans des universités en 2023 dans le monde francophone a révélé qu'environ 18% des enseignants rencontre des cas de travaux qui semblent d'Être produits par AI.

Cela introduit l'outil de Détection Automatique qui aide à préserver en premier les valeurs éducatives dans le cas et l'instruction.

Chaque discipline doit avoir une idée ou produit nouveau pour les lecteurs. En 2022 un journal connu de nature a enlevé 3 publications, qui pourraient en leur temps être paradigmatiques. De telles suppressions de contenu sans technologies AI ou même avec elles suffisent à garantir l'authenticité et l'impact sur eux.

Les blogueurs, journalistes et auteurs indépendants dépendent de leur originalité pour fidéliser leur audience. Les lecteurs repèrent rapidement des passages artificiels ou répétitifs, ce qui peut entraîner une perte de confiance. Par exemple, un article généré par IA sur un site d'actualités français a dû être retiré après avoir été signalé pour son manque de nuance. Les outils de détection aident ces professionnels à maintenir un niveau élevé de qualité avant publication.

5) Objectif

L'objectif de notre travail est de développer et d'évaluer une solution efficace pour la détection automatique de contenus générés par l'intelligence artificielle (IA) dans les textes français. Bien que des outils existent pour identifier les textes plagiés ou les contenus de faible qualité, il n'existe pas de solution largement reconnue pour identifier de manière fiable les textes produits par IA, en particulier dans le contexte de la langue française.

De nombreuses recherches actuelles se concentrent sur la détection de contenus générés par IA par des approches basées sur des caractéristiques linguistiques spécifiques ou des modèles statistiques. Cependant, les travaux explorant des approches intrinsèques, qui analysent le texte lui-même sans référence à des sources externes, restent encore limités.

Notre travail vise à combler cette lacune en explorant l'utilisation d'algorithmes d'apprentissage automatique, en particulier les méthodes non supervisées telles que le clustering. Nous cherchons à étudier l'efficacité du clustering pour la détection intrinsèque de contenus générés par IA dans les textes français.

Pour atteindre cet objectif, nous adapterons et mettrons en œuvre une méthode basée sur l'analyse des structures textuelles et des similarités sémantiques, en utilisant des techniques d'apprentissage non supervisé. L'algorithme proposé permettra de comparer les caractéristiques des textes et de détecter des anomalies typiques des contenus générés par IA, sans avoir besoin de corpus de référence externes.

Nous testerons notre approche sur un corpus de textes français spécialement conçu pour évaluer les méthodes de détection de contenus générés par IA. Notre étude vise à démontrer l'efficacité de cette méthode et à contribuer à l'avancement de la détection automatique de contenus générés par IA dans les textes français.

6) Plan de mémoire

Ce mémoire est organisé en trois chapitres essentiels, chacun traitant d'aspects spécifiques de la réalisation de notre système de détection automatique de textes générés par intelligence artificielle, hormis l'introduction générale et la conclusion générale, le mémoire est organisé comme suit :

- Le chapitre « **Fondements Théoriques** » définit l'IA générative (GPT, LLaMA) pour les textes en français, retrace l'évolution des méthodes de détection (manuelles à deep learning), analyse les défis linguistiques du français et explore les enjeux éthiques, sociaux et légaux, posant les bases du projet.

- Le chapitre « **État de l'art** » étudie les méthodes de détection des textes IA, incluant mesures de similarité (euclidienne, Jaccard, Manhattan, cosinus), techniques de contournement (réécriture, paraphrase, modèles hybrides) et de détection (perplexité, burstiness, modèles supervisés). Il compare des outils (Copyleaks, ZeroGPT, QuillBot), présente des concepts techniques (deep learning, CNN, GloVe, TF-IDF, WordNet) et aborde le cadre légal (AI Act, responsabilité).

- Le chapitre « **Conception et Réalisation du Système TTDetect** » décrit le développement d'un système de détection de textes IA, structuré en quatre modules : collecte/prétraitement, analyse linguistique, classification avec CamemBERT, et interface Streamlit. Il utilise Python, Transformers, PyTorch, spaCy, NLTK, et justifie CamemBERT pour le français. Le prétraitement inclut nettoyage et tokenisation, la classification repose sur heuristiques et fine-tuning. Testé en journalisme, éducation et cybersécurité, il fait face à des défis comme les variations linguistiques et biais. Perspectives : modèles hybrides et optimisation pour paraphrases.

Chapitre 1 : Fondements Théoriques

I. Chapitre 1 : Fondements Théoriques

1.1 Introduction

Dans ce chapitre, nous explorons les fondements du phénomène des IA génératives. Nous définissons ce qu'est l'intelligence artificielle générative, comment elle produit des textes en français, et retraçons l'historique de la détection des contenus générés artificiellement. Ensuite, nous analysons les principales techniques de détection, en mettant l'accent sur les approches classiques et modernes, notamment celles basées sur l'analyse statistique, le machine learning et les modèles de deep learning. Enfin, nous abordons les défis que pose cette technologie, ainsi que les implications éthiques, sociales et légales qui en découlent.

1.2 IA génératives

1.2.1 Définition de l'intelligence artificielle générative

L'intelligence artificielle générative désigne un domaine de l'IA qui se concentre sur la création de contenus originaux, tels que du texte, des images, du son ou du code, en s'appuyant sur des algorithmes avancés d'apprentissage automatique (**Goodfellow et al., 2014**). Contrairement aux systèmes d'IA traditionnels, qui se limitent à analyser et classer des données, les modèles génératifs sont capables de produire de nouvelles données en se basant sur des exemples préexistants.

Les modèles d'IA générative utilisent des architectures neuronales, en particulier les réseaux de neurones profonds et les modèles Transformers, pour apprendre la structure et les relations sous-jacentes des données (**Bommasani et al., 2021**). Parmi les modèles les plus connus, on retrouve GPT pour la génération de texte, DALL·E pour la création d'images, et MusicLM pour la génération musicale.

L'IA générative est aujourd'hui largement utilisée dans des domaines variés, notamment la rédaction automatique, la création artistique, le développement de code et la synthèse vocale**.** Cependant, son utilisation pose des défis éthiques et techniques, notamment en matière d'authenticité, de biais et de détection des contenus artificiels **(**Bommasani et al., 2021**).**

1.2.2 Définition des textes français générés par IA

Les textes français générés par IA sont des contenus rédigés en langue française à l'aide d'un modèle d'intelligence artificielle générative. Ces textes sont produits automatiquement par des algorithmes entraînés sur de vastes corpus de données textuelles, leur permettant d'imiter le style et la syntaxe du langage humain **(**Brown et al., 2020).

Ces textes peuvent être générés dans divers contextes, tels que :

- **La rédaction automatique** d'articles journalistiques ou académiques.
- **Les chatbots et assistants virtuels** capables de répondre en français.
- **La génération de résumés** ou de traductions automatiques.

La qualité des textes produits par IA varie en fonction du modèle utilisé, de son entraînement et de la complexité du contenu généré**.** Certains textes peuvent paraître très proches de ceux écrits par des humains, rendant leur détection plus difficile.

Face à cette avancée, des outils et méthodes ont été développés pour analyser la cohérence, la perplexité et les schémas linguistiques des textes générés, afin de distinguer ceux produits par IA de ceux rédigés par des humains (Solaiman et al., 2019)**.** La capacité à identifier ces textes est cruciale pour préserver l'intégrité des informations et limiter les risques liés à la désinformation ou au plagiat involontaire**.**

1.2.3 Définition de LLaMA

LLaMA est un modèle clé dans l'étude des contenus générés par IA, notamment pour évaluer la robutesse des outils de détection . Il est développé par Meta AI, et est une famille de modèles de langage conçue pour la recherche en traitement du langage naturel. Introduite en 2023, elle comprend plusieurs versions (LLaMA-7B, LLaMA-13B, LLaMA-70B) qui se distinguent par leur taille et leurs capacités (Touvron et al., 2023). Contrairement à des modèles commerciaux comme GPT-4 ou Bard, LLaMA est optimisé pour une efficacité computationnelle, nécessitant moins de ressources tout en offrant des performances comparables sur des tâches comme la génération de texte ou la classification.

LLaMA est entraîné sur des corpus publics massifs, tels que Common Crawl et Wikipedia, ce qui lui permet de produire des textes cohérents, y compris en français. Cependant, son optimisation principale pour l'anglais peut limiter sa capacité à capturer les nuances lexicales et grammaticales spécifiques du français, comme les expressions idiomatiques ou les variations régionales. Par exemple, un texte généré par LLaMA-13B, tel que « L'intelligence artificielle offre des solutions innovantes pour automatiser les tâches complexes », peut sembler fluide mais présenter des schémas répétitifs ou une absence de registre familier, des caractéristiques exploitables pour la détection.

Dans le cadre de la détection des textes générés par IA, pose un défi particulier en raison de sa popularité dans les milieux académiques, où il est souvent utilisé pour générer des articles ou des résumés. Les détecteurs doivent donc être capables d'identifier ses signatures stylistiques, comme une faible variabilité lexicale ou une structure syntaxique prévisible. Malgré son accès restreint (non open-source au sens strict),

1.2.4 L'histoire de la détection des contenus générés par IA

Au début, la détection des textes générés par IA reposait principalement sur l'intuition et l'expertise des lecteurs. Les chercheurs et linguistes devaient analyser manuellement les textes pour repérer des incohérences stylistiques ou des tournures répétitives caractéristiques des premiers modèles génératifs. Cependant, cette approche s'est rapidement révélée insuffisante, car elle dépendait fortement des capacités d'observation humaine et ne pouvait pas suivre l'essor rapide des technologies d'IA.

Avec l'arrivée des premiers modèles de génération automatique de texte, les chercheurs ont constaté la nécessité de méthodes plus fiables et automatisées pour distinguer les contenus produits par des machines de ceux rédigés par des humains. Au

début des années 2000, les premières approches statistiques ont vu le jour, basées sur l'analyse de la fréquence des mots et des modèles probabilistes. Ces méthodes ont permis d'identifier certains schémas typiques des textes générés, mais elles restaient limitées face aux modèles IA de plus en plus sophistiqués**.**

L'essor des modèles neuronaux et des Transformers, notamment avec l'apparition de GPT-2 en 2019, a marqué une nouvelle ère dans la génération de texte. Ces avancées ont également poussé au développement d'outils spécialisés pour la détection de contenu IA. Des méthodes comme l'analyse de perplexité, qui mesure la prévisibilité d'un texte, ou l'apprentissage supervisé, où des algorithmes sont entraînés à différencier les textes humains et IA, ont été introduites.

En 2023, avec l'explosion des modèles comme GPT-4, LLaMA et Claude, la nécessité d'outils plus performants s'est faite sentir. Des solutions comme GPTZero, ZeroGPT et AI Text Classifier d'OpenAI ont été développées, mais elles ont montré des limites, notamment en termes de faux positifs et de détection des textes en français **(**Chakraborty et al., 2023**).** Aujourd'hui, bien que des avancées aient été réalisées, la détection des contenus générés par IA reste un défi majeur, nécessitant des recherches approfondies pour développer des algorithmes plus précis et adaptés aux spécificités linguistiques du français **(**Abburi et al., 2023**).**

1.3 Langue française et défis linguistiques

La langue française, avec ses 300 millions de locuteurs, est marquée par une richesse lexicale et des complexités grammaticales (Grevisse & Goosse, 2016). Ses homophones, exceptions orthographiques et polysémie compliquent la détection des textes IA, qui peinent à reproduire les nuances régionales ou les registres variés (par exemple, argot vs langage académique).

Exemple concret :

Texte IA : « La réunion a été très productive et efficace. » (formulation générique)

Texte humain : « La réunion était hyper constructive, on a bien avancé ! » (style familier, typique d'un locuteur natif)
Ces différences subtiles sont un défi pour les algorithmes de détection

1.4 Défis et perspectives

Dans cette section, nous analysons les défis liés à la détection des textes générés par IA, en mettant particulièrement l'accent sur les spécificités des contenus en langue française. Nous étudierons les limites des outils actuels, les pistes d'amélioration possibles ainsi que les implications éthiques et réglementaires de cette technologie.

1.4.1 Limites des outils actuels

Les outils de détection des textes générés par IA présentent plusieurs limites :

- **Taux de faux positifs et négatifs** : Certains textes humains sont parfois classés à tort comme générés par IA, tandis que des textes IA bien formulés peuvent échapper à la détection [1].
- **Difficulté à suivre l'évolution des modèles IA** : Avec chaque nouvelle génération de modèles comme GPT-4 ou Claude, la qualité des textes générés s'améliore, rendant leur détection plus complexe [2].
- **Manque d'outils spécialisés en français** : La plupart des détecteurs sont optimisés pour l'anglais et sont moins efficaces pour le français.
- **Coût élevé des solutions avancées** : Certains outils performants nécessitent un abonnement payant, ce qui peut limiter leur accessibilité.

1.4.2 Améliorations et innovations futures

Pour améliorer la détection des textes générés par IA, plusieurs pistes peuvent être explorées :

- **Utilisation de modèles IA explicables** : Développer des algorithmes transparents qui indiquent les indices utilisés pour classer un texte comme IA.
- **Approches multi-modales** : Combiner plusieurs méthodes (perplexité, analyse syntaxique, machine learning) pour obtenir de meilleurs résultats.
- **Création de bases de données annotées en français** : Développer des datasets spécifiques pour entraîner des modèles adaptés aux textes francophones.
- **Développement d'outils accessibles** : Encourager la mise à disposition d'outils open-source pour démocratiser l'accès à la détection IA.

1.4.3 Implications éthiques et réglementaires

La détection des textes générés par IA soulève plusieurs questions éthiques et légales :

- **Vie privée et surveillance** : Jusqu'où peut-on analyser les contenus produits par les utilisateurs sans porter atteinte à leur vie privée ? [3]
- **Impact sur l'éducation et la recherche** : Comment garantir que les étudiants et chercheurs utilisent l'IA de manière responsable ?
- **Cadre législatif** : Quesl types de lois pourraient efficacement encadrer l'usage des IA génératives sans freiner l'innovation technologique ?[4]
- **Biais algorithmiques** : Quels mécanismes doivent être mis en place pour garantir une diversité linguistique, culturelle et stylistique dans les jeux de données d'entraînemen ?

[1] https://lucide.ai/limites-detecteurs-ia Lien pour comprendre les limites techniques et éthiques actuelles des détecteurs de contenus générés par IA, ainsi que les risques de faux positifs.

[2] https://dysclick.fr/detecter-les-textes-ia-mission-impossible Lien qui explore les raisons pour lesquelles détecter un texte généré par IA devient de plus en plus difficile, voire impossible, face aux avancées des modèles.

[3] https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN_avis7_06_09_2023_web-rs2.pdf Lien vers un avis du Comité national pilote d'éthique du numérique (CNPEN) sur les enjeux éthiques liés à l'usage de l'IA générative, notamment en matière d'éducation et de vie privée.

[4] https://www.unesco.org/fr/artificial-intelligence/recommendation-ethics Lien pour consulter les recommandations officielles de l'UNESCO en faveur d'une intelligence artificielle éthique, inclusive et respectueuse des droits humains.

- **L'évolution rapide des modèles génératifs** : Comment les outils de détection peuvent-ils suivre le rythme d'évolution rapide des modèles génératifs, parfois mis à jour ou remplacés tous les quelques mois ?

1.5 Légalité et cadre réglementaire

L'essor de l'intelligence artificielle générative (IA-G) pose des défis juridiques importants, notamment en matière de droits d'auteur, responsabilité légale et régulation des contenus générés par IA. Plusieurs gouvernements et institutions travaillent à encadrer son utilisation.

1.5.1 Lois et régulations sur l'utilisation de l'IA générative

1) **Régulations générales en cours d'adoption**

AI Act (Union Européenne - 2024) (*European Commission, 2024*)

- Première législation complète sur l'IA.
- Classe l'IA en différents niveaux de risque (minime, limité, élevé, inacceptable).
- L'IA générative comme ChatGPT doit respecter des obligations de transparence (indication qu'un texte est généré par IA).
- Sanctions financières en cas de non-respect.

Executive Order on AI (USA - 2023/2024) (*White House, 2023*)

- Règles sur la transparence des modèles IA et la régulation de leur utilisation.
- Exige que les entreprises d'IA documentent leurs données d'entraînement.

Chine

- Régulations strictes sur les contenus générés par IA.
- Obligation pour les plateformes de censurer certains types de contenu.

2) **Propriété intellectuelle et droits d'auteur**

- **Qui détient les droits d'un texte généré par IA ?**
  - En UE et aux USA : un contenu entièrement généré par IA n'est pas protégé par le droit d'auteur.
  - Si une personne modifie significativement un texte IA, elle peut revendiquer la propriété.
- **L'IA peut-elle utiliser du contenu protégé ?**
  - Certains modèles d'IA sont accusés d'avoir été entraînés sur des données protégées par le droit d'auteur sans consentement.
  - Plusieurs procès (*OpenAI vs. auteurs et médias*) visent à encadrer l'entraînement des IA sur du contenu existant.

3) **Transparence et étiquetage des contenus IA**

- Plusieurs lois en discussion imposent que les textes générés par IA soient signalés explicitement (*UNESCO, 2023*).
- Certains outils comme **Watermarking IA** (empreinte numérique) permettent d'identifier un texte IA (Kreps et al., 2023).

1.5.2 Responsabilités légales des créateurs de contenu IA

Les auteurs de contenu IA peuvent être tenus responsables selon le contexte d'utilisation.

1) Fake News et désinformation

- **Loi française contre la manipulation de l'information (2018)** :
  - Des plateformes peuvent être sanctionnées si elles laissent circuler des deepfakes ou du contenu IA trompeur.
  - L'UE envisage de rendre les créateurs de contenu IA responsables de la diffusion de fausses informations.

2) Plagiat et fraude académique

- Les universités adoptent des règlements spécifiques pour interdire l'usage non autorisé d'IA générative dans les devoirs et thèses.
- Certaines institutions intègrent **Turnitin, GPTZero et autres détecteurs IA** pour lutter contre la triche.

3) Responsabilité en cas de préjudice

- **Si un texte IA entraîne un dommage (diffamation, incitation à la haine, arnaques, etc.), qui est responsable ?**
  - Le créateur du contenu ? (personne ayant utilisé l'IA)
  - L'éditeur ou la plateforme de diffusion ?
  - Le développeur du modèle IA ?
- **Régulations à venir :**
  - L'UE veut imposer une **responsabilité partagée** entre les utilisateurs et les entreprises d'IA.
  - Les entreprises doivent intégrer des **mécanismes de sécurité** pour empêcher l'utilisation abusive de leurs modèles.

1.6 Conclusion

Ce chapitre établit les fondations théoriques des textes générés par IA, en définissant l'IA générative (ex. LLaMA) et ses défis pour le français. Il souligne les limites des outils actuels (inefficacité, faux positifs, coût) et propose des innovations comme des approches multi-modales et des corpus annotés. Les aspects éthiques, sociaux et juridiques, incluant l'AI Act et les questions de responsabilité, sont abordés. Ce cadre guide la conception future d'un système de détection adapté au français

Chapitre 2 : État de l’art

II. Chapitre 2 : État de l'art

2.1 Introduction

Ce chapitre introduit les concepts fondamentaux liés à l'évaluation de similarité entre textes, aux techniques utilisées pour échapper aux détecteurs d'IA, aux logiciels existants ainsi qu'au cadre légal entourant l'usage des textes générés automatiquement.

2.2 Méthodes de Mesure de similarité textuelle

La mesure de similarité est une technique utilisée pour évaluer à quel point deux objets, tels que des vecteurs, des ensembles de données 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. Dans ce qui suit, les proncipales mesures de similarité utilisées pour comparer des text et identifier des contenus artificiles.

2.2.1 Distance euclidienne

La distance euclidienne est une mesure de la distance entre deux points dans un espace euclidien. En deux dimensions, la distance euclidienne entre deux points $(x_1, y_1)\ et\ (x_2, y_2)$ peut être calculée à l'aide du théorème de Pythagore :

$$D(x1\ , x2) = \sqrt{\sum_{j=1}^{n} \left(y_{1j} - y_{2j}\right)^2}$$

Équation 1 Distance euclidienne

La distance euclidienne est largement utilisée pour mesurer la similarité dans les espaces vectoriels.

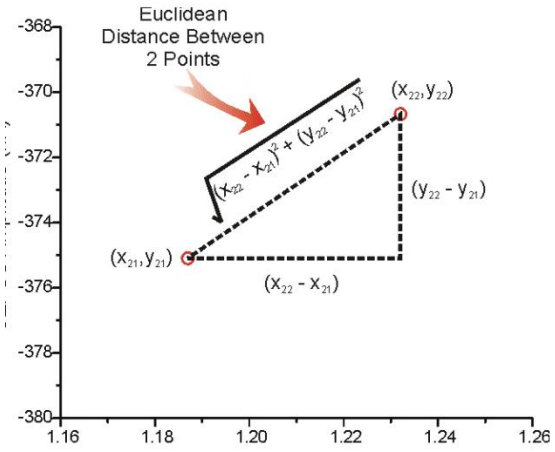

Figure 1 Distances Euclidienne

2.2.2 Distance de Jaccard

La distance de Jaccard mesure la similarité entre deux ensembles en comparant leur intersection et leur union. Elle est couramment utilisée pour évaluer la proximité de deux mots ou documents en fonction des termes qu'ils partagent.

La formule pour le calcul de la distance de Jaccard entre deux ensembles A et B est :

$$J(A\,,B) = 1 - \frac{A \cap B}{A \cup B}$$

Équation 2 Distance jaccard

Une distance de Jaccard proche de 0 indique une forte similarité entre les ensembles, tandis qu'une valeur proche de 1 signale une forte dissimilarité.

2.2.3 Distance de Manhattan

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

La formule de la distance de Manhattan entre deux points $P = (p_1, p_2, \dots, p_n)$ et $Q = (q_1, q_2, \dots, q_n)$ est :

$$D_M(P,Q) = \sum_{i=1}^{n} |p_i - q_i|$$

Équation 3 Distance de Manhattan

Cette mesure est souvent utilisée dans les modèles où les déplacements ne se font que selon des axes perpendiculaires.

2.2.4 Similarité Cosinus

La similarité cosinus est une mesure qui évalue la similarité entre deux vecteurs en mesurant le cosinus de l'angle entre eux. Elle est particulièrement utile dans le traitement du langage naturel pour comparer des documents textuels sous forme de vecteurs de caractéristiques.

La formule de la similarité cosinus entre deux vecteurs A et B est :

$$sim(A,B) = \cos(\theta) = \frac{A.B}{\|A\|\|B\|}$$

Équation 4 Similarité Cosinus

où :

- $(A \cdot B)$ est le produit scalaire des vecteurs A et B.
- $\|A\|$ est la norme (ou longueur) du vecteur A.
- $\|B\|$ est la norme du vecteur B.

Une valeur proche de 1 indique une forte similarité, tandis qu'une valeur proche de -1 signale une forte opposition.

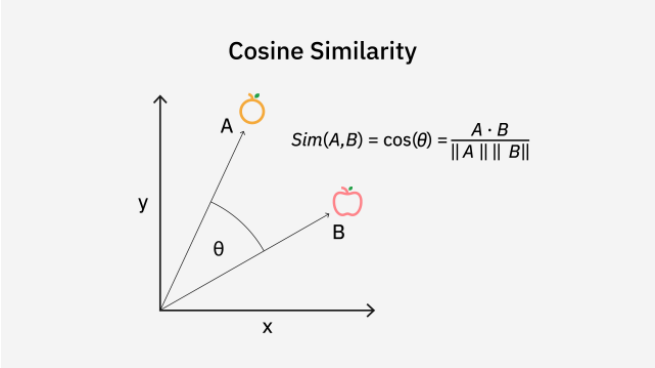

Figure 2 Similarité Cosinus entre deux vecteurs

Ces mesures de similarité jouent un rôle clé dans la détection des contenus générés par IA, en permettant d'évaluer la ressemblance entre des textes générés automatiquement et des textes humains.

2.3 Techniques de détection des textes générés par IA

Dans cette section, nous analysons les techniques de détection des textes générés par IA.

2.3.1 Analyse de perplexité et burstiness

1. Perplexité

La perplexité est une mesure utilisée en traitement du langage naturel (NLP) pour évaluer la qualité d'un modèle de langage. Elle indique à quel point un modèle trouve un texte prévisible.

2. Burstiness (Rafale de mots)

Le burstiness décrit la variabilité dans la longueur et la structure des phrases. Les humains ont tendance à écrire avec plus de diversité : alternance entre phrases courtes et longues, variabilité dans l'utilisation des synonymes et des structures grammaticales.[5]

2.3.2 Modèles supervisés pour la classification des textes

Modèles classiques de Machine Learning (ML)

- Naïve Bayes
- SVM
- Random Forest

Modèles avancés : CNN et Transformers

- **CNN** : Analyse des motifs récurrents dans la structure du texte.
- **Transformers (BERT, RoBERTa, GPT-Detectors)** : Capables d'analyser la sémantique et le style d'un texte.

2.3.3 Approches hybrides combinant plusieurs méthodes

Combinaison de perplexité et de ML

- Utilisation de SVM ou Random Forest pour classifier les textes.

Fusion CNN et Transformers

- CNN détecte les structures répétitives.
- Un Transformer (BERT, RoBERTa) analyse la cohérence et la logique du texte.

[5] https://humanizeai.now/blog/perplexity-burstiness-2025 Lien pour découvrir comment la détection de l'IA a évolué au-delà de simples mesures telles que la perplexité et la sporadicité

2.4 Techniques de traitement automatique du langage IA

2.4.1 Deep learning ou apprentissage profond

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

2.4.2 Réseaux de neurones pour la détection de textes IA

Dans le domaine de l'apprentissage profond, le **CNN** 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, chacune spécialisée dans la reconnaissance de motifs de plus en plus complexes.

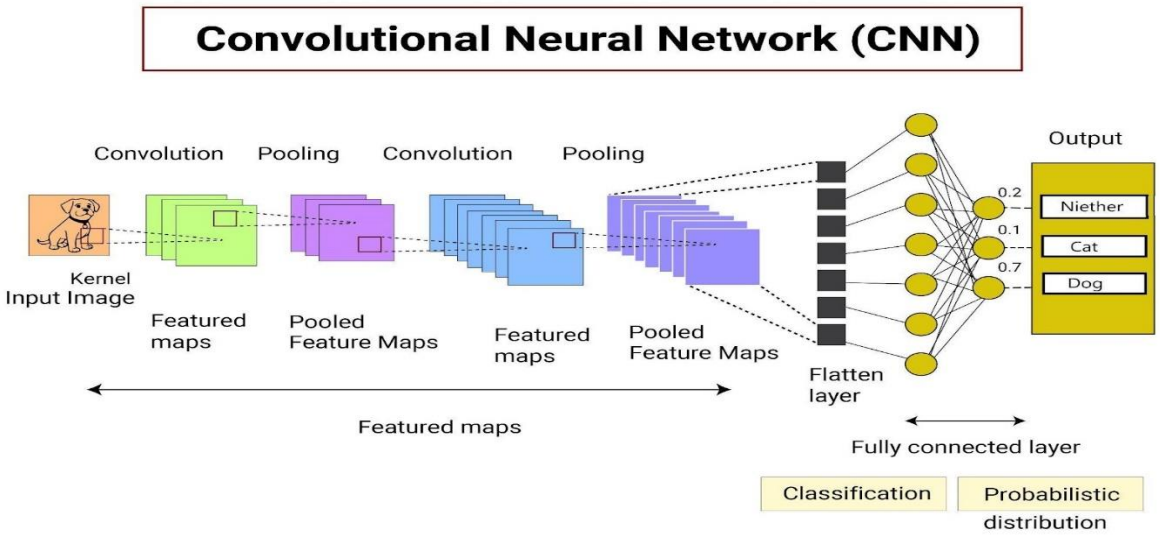

Figure 3 Architecture standard d'un réseau à convolutions[6]

[6] Lien pour consulte CNN : https://blog.csdn.net/zfh1005/article/details/147743829

2.4.3 Glove

GloVe est une méthode de plongement lexical développée par **Pennington, Socher et Manning (2014)** à l'Université de Stanford. Contrairement aux modèles basés sur des fenêtres locales de mots comme **Word2Vec**, GloVe exploite les statistiques globales de cooccurrence des mots dans un large corpus. En utilisant une factorisation de matrice sur la distribution des mots, GloVe permet de capturer les relations sémantiques et syntaxiques entre eux, offrant ainsi des représentations vectorielles plus riches et interprétables (Pennington et al., 2014).

L'approche de GloVe repose sur l'analyse de la fréquence relative des cooccurrences de mots plutôt que sur leur simple voisinage dans des phrases. Par exemple, si deux mots apparaissent souvent dans des contextes similaires, ils auront des représentations vectorielles proches. Ce modèle a été entraîné sur des corpus de grande envergure comme **Wikipedia et Common Crawl**, ce qui lui permet d'être efficace pour de nombreuses tâches NLP, telles que la classification de textes et la détection de similarité sémantique (Levy & Goldberg, 2014).

2.4.4 TF-IDF

TF-IDF est une technique couramment utilisée en **traitement automatique du NLP** pour évaluer l'importance d'un mot dans un document par rapport à un ensemble de documents. Introduite par **Sparck Jones (1972)**, elle repose sur l'idée que les mots fréquemment utilisés dans un document mais rarement présents dans d'autres documents sont plus représentatifs de son contenu.

La formule TF-IDF est donnée par :

$$\mathrm{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t)$$

Les formules des composantes sont :

TF(t, d) = Nombre d'occurrences du terme t dans le document d / Nombre total de termes dans le document d

IDF(t) = log(N / (1 + df(t)))

où N est le nombre total de documents et df(t) est le nombre de documents contenant le terme t.

où **TF** mesure la fréquence du terme t dans le document d, tandis que **IDF** réduit le poids des mots trop courants en tenant compte de leur fréquence dans l'ensemble du corpus (Sparck Jones, 1972).

L'intérêt principal du **TF-IDF** est sa capacité à identifier les mots-clés les plus significatifs dans un texte sans nécessiter d'apprentissage supervisé. Il est largement utilisé dans **l'indexation de documents, la récupération d'information et les moteurs de recherche** comme Google (Ramos, 2003).

2.4.5 WordNet

WordNet est une **base de données lexicale** développée à l'Université de Princeton par **George A. Miller et son équipe dans les années 1980**. Elle organise les mots de la langue anglaise en ensembles de synonymes appelés **synsets**, qui sont reliés entre eux par des relations sémantiques telles que l'**hyponymie** (relation de spécialisation), l'**hyperonymie** (relation de généralisation) et la **synonymie** (Miller, 1995).

L'un des principaux atouts de **WordNet** est son utilisation dans des tâches de **désambiguïsation lexicale**, où il permet d'identifier le sens correct d'un mot en fonction de son contexte. Par exemple, le mot *bank* peut désigner à la fois une banque et une rive de rivière. Grâce à ses connexions sémantiques, **WordNet aide les algorithmes NLP** à sélectionner la signification appropriée dans un contexte donné (Fellbaum, 1998).

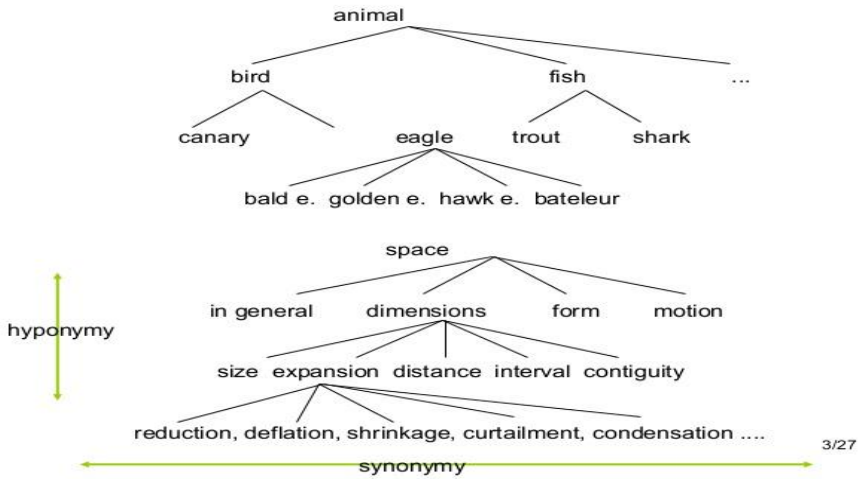

Figure 4 WordNet (didaquest, s.d.)

2.5 Modèles de classification

Plusieurs modèles d'intelligence artificielle ont été développés pour la génération et la détection de textes. Ces modèles peuvent être adaptés à la détection de contenu généré par IA en utilisant des techniques d'entraînement supervisé. Voici quelques-uns des modèles les plus connus :

- **BERT (Bidirectional Encoder Representations from Transformers)** : Un modèle pré-entraîné développé par Google, capable de comprendre le contexte bidirectionnel dans un texte. Très utilisé pour la classification de texte.[7]
- **RoBERTa (Robustly optimized BERT approach)** : Une version optimisée de BERT proposée par Facebook AI. Il est plus robuste et performant sur plusieurs tâches de NLP.[8]

[7] Lien pour consulter BERT : https://www.intelligence-artificielle-school.com/ecole/technologies/abert-modele-nlp/

[8] Lien pour consulter RoBERTa : https://medium.com/@marketing_novita.ai/introducing-roberta-base-model-a-comprehensive-overview-330338afa082

- **DistilBERT** : Une version légère de BERT, offrant de bonnes performances tout en étant plus rapide et moins coûteux en ressources.[9]
- **T5** : Transforme toutes les tâches NLP en un problème de traduction de texte à texte. Très flexible et puissant.[10]
- **XLM-Roberta** : Une version multilingue de RoBERTa, adaptée à plusieurs langues, y compris le français.[11]
- **CamemBERT** : Un modèle de langage basé sur RoBERTa , adaptée à la lange française .

Chacun de ces modèles peut être entraîné sur des corpus annotés IA/Humain pour réaliser des tâches de classification binaire dans le cadre de détection de texte généré

2.6 Techniques de contournement des détecteurs IA

Les auteurs de contenus générés par IA (qu'il s'agisse de fraude académique, de désinformation ou de génération automatisée) utilisent différentes méthodes pour éviter la détection.

2.6.1 Méthodes de réécriture et paraphrase automatique

L'une des stratégies les plus efficaces pour contourner un détecteur est de modifier un texte généré par IA de manière subtile afin qu'il ressemble davantage à un texte humain.

a. Techniques courantes de réécriture

1) Utilisation de paraphraseurs automatiques (Text Spinners)

Des outils comme **QuillBot, Spinbot, et Paraphrase.io** permettent de reformuler un texte en modifiant l'ordre des mots, la syntaxe et le choix lexical (**Li & Hovy, 2014**). Ces outils introduisent des variations lexicales et syntaxiques, ce qui diminue l'efficacité des détecteurs basés sur la perplexité et la stylométrie.

- **Exemple :**
  - **Texte IA** : L'intelligence artificielle est en train de révolutionner de nombreux secteurs.
  - **Paraphrasé** : De nombreux domaines connaissent une transformation radicale grâce à l'intelligence artificielle.

2) Synonymisation et reformulation manuelle

Un utilisateur peut manuellement remplacer certains mots-clés par des synonymes ou modifier l'ordre des phrases (**Iyyer et al., 2018**).

Cette approche est plus difficile à détecter, car elle imite le comportement humain.

[9] Lien pour consulter DistilBERT : https://arxiv.org/abs/1910.01108
[10] Lien pour consulter T5 : https://huggingface.co/docs/transformers/model_doc/t5
[11] Lien pour consulter XML-Roberta: https://huggingface.co/docs/transformers/model_doc/xlm-roberta

3) Traductions automatiques en cascade

Traduire un texte IA dans une autre langue, puis le retraduire en français avec **Google Translate, DeepL ou ChatGPT**.

Cela peut introduire des variations stylistiques qui perturbent les modèles de détection (**Schuster et al., 2020**).

b. Limites de la réécriture automatique

Efficace contre les détecteurs simples (perplexité, burstiness, lexique statique).

Peu efficace contre des modèles avancés (Transformers entraînés sur des textes paraphrasés).

La qualité du texte peut se dégrader après plusieurs paraphrases.

2.6.2 Utilisation de modèles hybrides IA/humains

Une autre approche avancée consiste à mélanger du texte généré par IA avec des modifications humaines, rendant la détection beaucoup plus difficile.

- **Méthode de correction et d'enrichissement humain**

Un utilisateur peut générer un texte avec une IA, puis le réviser en changeant certaines parties, en introduisant des erreurs naturelles ou en ajoutant des expressions idiomatiques (**Gehrmann et al., 2019**). L'intervention humaine casse la structure prévisible des textes IA.et Les détecteurs basés sur la stylométrie peuvent être trompés si les modifications sont suffisantes.

- **Génération hybride avec IA et prompts avancés**

Certains modèles de langage comme **ChatGPT, Claude ou Gemini** peuvent être guidés avec des instructions précises pour produire un texte plus "humain" (**Clark et al., 2021**). Cela introduit du burstiness artificiel et casse les schémas prévisibles des IA classiques. Les détecteurs entraînés sur des modèles standards peuvent avoir du mal à identifier un texte généré avec ces variations.

- **Exemples de prompts avancés :**
  - Écris ce texte avec un style très variable, en alternant phrases longues et courtes.
  - Ajoute des fautes de frappe et des répétitions occasionnelles.
  - Utilise des expressions familières et des changements de ton spontanés.

- **Utilisation de modèles multiples**

Au lieu d'utiliser un seul modèle IA, certains mélangent plusieurs générateurs (**GPT-4, Claude, Mistral, LLaMA**) pour introduire des variations (**Uchendu et al., 2020**). Cela peut perturber un détecteur entraîné sur un seul type de texte IA.

2.6.3 Quelques logiciels pour la détection de l'IA

Voici quelques-uns des principaux logiciels de détection de l'IA largement utilisés dans les établissements d'enseignement et les lieux de travail :

1) Copyleaks

Une plate-forme d'analyse de texte IA de premier plan, dédiée à aider les entreprises et les établissements d'enseignement à naviguer dans le paysage en constante évolution de **GenAI** grâce à une innovation IA responsable, en équilibrant le progrès technologique avec l'intégrité, la transparence et l'éthique.

Figure 5 l'interface du logiciel "Copyleaks"[12]

2) ZeroGPT

ZeroGPT est un outil conçu pour détecter si un texte a été généré par une intelligence artificielle, comme **ChatGPT**, ou s'il a été écrit par un humain. Il utilise des algorithmes avancés d'apprentissage automatique pour analyser les structures linguistiques, les modèles de rédaction et d'autres indicateurs afin d'estimer la probabilité qu'un texte soit artificiel.

[12]Lien pour consulter copuleaks : https://copyleaks.com/fr/

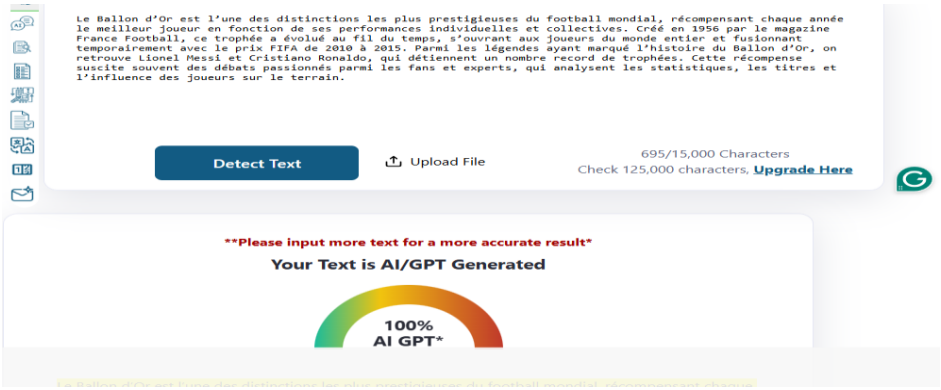

Figure 6 l'interface du logiciel "ZeroGPT "[13]

3) QuillBot AI Détecteur

QuillBot AI Détecteurest un outil conçu pour analyser un texte et déterminer s'il a été généré par une intelligence artificielle, comme **ChatGPT, GPT-4 ou d'autres modèles de langage**. Il évalue les structures de phrases, les schémas linguistiques et d'autres indices pour estimer la probabilité que le contenu soit artificiel.

Figure 7 l'interface du logiciel "QuillBot"[14]

2.6.4 Comparaison entre les outils similaires

Avec l'augmentation de l'utilisation des modèles génératifs d'IA, plusieurs outils de détection ont vu le jour pour différencier les textes humains des contenus générés automatiquement. Voici une comparaison des principaux outils existants :

[13] Lien pour consulter ZeroGpt : https://www.zerogpt.com/
[14] Lien pour consulter QuillBot : https://quillbot.com/fr/detecteur-ia

Outil	Méthode utilisé	Payant ?	Performances	Educatif
GPTZero	Analyse de perplaxité et burstiness	NON	Moyenne	OUI
QuillBot	Analyse stylométrique + patrons linguistiques	Partiellement (version gratuite disponible)	Moyenne à Élevée	OUI
Copyleaks	Modèle IA avancé et NLP	OUI	Elevée	NON

Tableau 1 Comparaison des outils similaires

Ces outils offrent différentes approches de détection, mais **aucun n'est encore totalement fiable à 100 %**. Le défi reste de développer des algorithmes plus précis, notamment pour les textes en français, et d'améliorer la robustesse face aux techniques de contournement utilisées par les générateurs de texte IA.

2.7 Conclusion

Ce chapitre explore les méthodes et outils de détection des textes IA, couvrant les mesures de similarité (euclidienne, Jaccard, Manhattan, cosinus), techniques de détection (perplexité, burstiness, SVM, Random Forest, Naïve Bayes, CNN, BERT, RoBERTa, CamemBERT), et approches hybrides. Il traite des contournements (réécriture, synonymisation, traduction en cascade, prompts avancés) et compare des outils comme Copyleaks, ZeroGPT, et QuillBot, notant leurs limites avec les textes français. Ces bases techniques guideront la conception d'un système de détection en français avec CamemBERT et heuristiques.

Chapitre 3 : Conception et Réalisation du TTDetect

III. Chapitre 3 : Conception et Réalisation du Système 'TTDetect'

3.1 Introduction

Dans ce chapitre, nous présentons les différentes étapes de conception, d'implémentation et de validation du notre système TTDetect . L'objectif est de proposer un outil léger, intuitif et performant, fondé principalement sur des heuristiques linguistiques et des techniques d'analyse statistique, intégrées dans une interface interactive construite avec Streamlit.

3.2 Architecture Globale du TTDetect

TTDetect repose sur une architecture modulaire articulée autour de six composants clés :

1. **Collecte et prétraitement des données :**
   - Chargement des textes humains et IA depuis des fichiers CSV.
   - Nettoyage (suppression des caractères spéciaux, normalisation Unicode).
   - Labélisation binaire : 0 pour les textes humains, 1 pour les textes IA.
2. **Module d'upload et d'extraction** : permet d'importer des fichiers PDF et Word (.docx) pour en extraire automatiquement le contenu textuel.
3. **Module d'analyse linguistique** : extrait des caractéristiques stylistiques du texte (longueur des phrases, diversité lexicale, fréquence des mots, etc.).
4. **Modèle de classification :**
   - Utilisation de CamemBERT, un modèle de Transformers pré-entraîné en français.
   - Fine-tuning sur un corpus équilibré (50% humains, 50% IA).
5. **Interface utilisateur (UI) :**
   - Développée avec Streamlit pour une interaction intuitive.
   - Visualisation des résultats via des jauges de confiance et des graphiques.
6. **Évaluation et déploiement :**
   - Calcul des métriques de performance (précision, rappel, F1-score).
   - Export du modèle fine-tuné au format PyTorch pour la production.

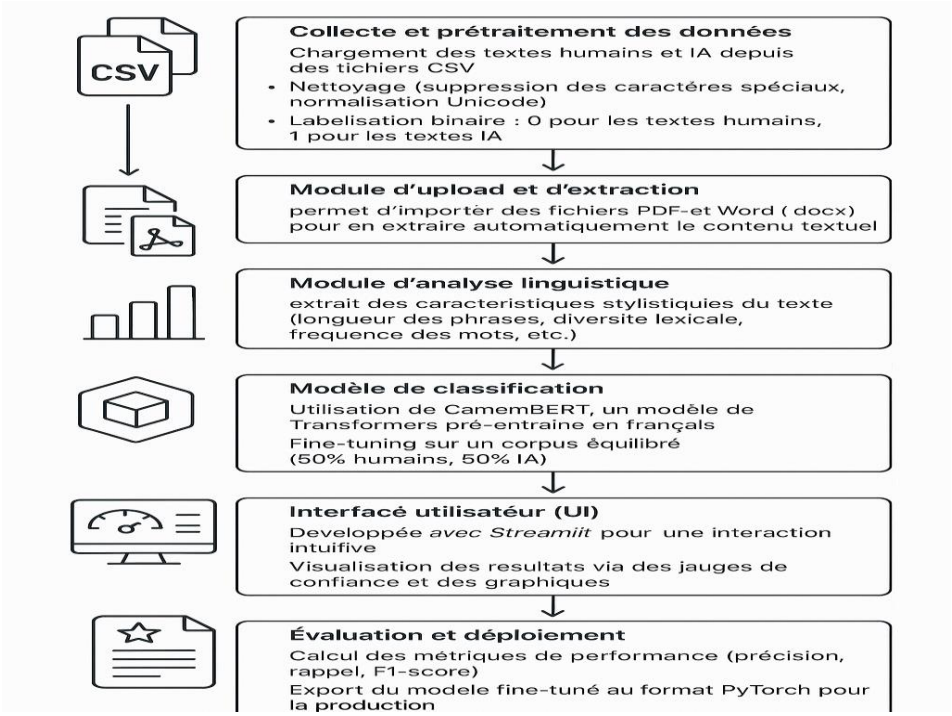

Figure 8 Architecture de TTDetect

3.3 Choix technologiques

Les technologies ont été sélectionnées pour leur performance, compatibilité avec le français, et facilité d'intégration :

3.3.1 Langage et Bibliothèques

Icon	Composant	Technologie	Role
python 3.10	**Langage Principal**	Pythone 3.1O	Développement du pipeline de traitement et de l'interface.
101010 Hi! 100101 Hello! NLP	**NLP**	Transformers (Hugging Face)	Chargement et fine-tuning de CamemBERT.
PyTorch	**Deep Learning**	PyTorch	Entraînement du modèle et gestion des tenseurs.
Streamlit	**Interface**	Streamlit	Création de l'UI interactive avec visualisation en temps réel.

spaCy vs NLTK	**Prétratement**	spaCy , NLTK	Tokenisation, suppression des stopwords, normalisation.
PyMuPDF	PyMuPDf	Docx,fitz	extraction de texte depuis fichiers Word/PDF.

Tableau 2 Stack technique du système

3.3.2 CamemBERT

Dans TTDetect , nous avons choisi **CamemBERT**, un modèle dérivé de RoBERTa et pré-entraîné spécifiquement sur des données en langue française. Développé par INRIA et Facebook AI, il est conçu pour capturer les spécificités lexicales, grammaticales et syntaxiques du français.

Avantages de CamemBERT dans TTDetect :

- **Spécialisation en français** : Entraîné sur des corpus massifs francophones (OSC, Wikipedia, etc.).
- **Compréhension sémantique** : Capable d'analyser le sens profond des phrases et les structures grammaticales.
- **Utilisation flexible** : Compatible avec la bibliothèque *transformers* de Hugging Face, facilitant son intégration dans un pipeline de classification.
- **Scalabilité** : Fonctionne aussi bien sur GPU que CPU, adapté à des déploiements légers.

Dans TTDetect, CamemBERT peut être utilisé avec ou sans fine-tuning. En l'absence de jeu de données annoté suffisant, nous avons opté dans un premier temps pour une approche **heuristique** combinée à CamemBERT, mais le système est prêt pour un **entraînement supervisé** complet dans les versions futures.

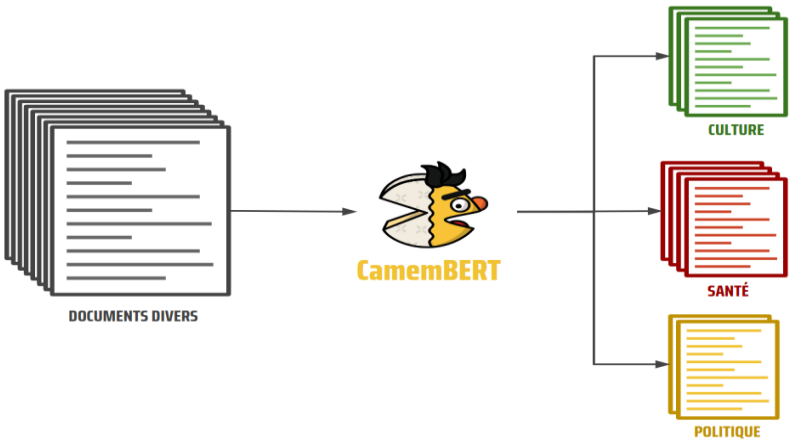

Figure 9 Architecture de CamemBERT fine-tuné (blog.ippon.fr, s.d.)

3.3.3 VS Code

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

3.4 Développement du TTDetect

3.4.1 Prétraitement des textes

Chaque texte est analysé à travers une série d'opérations :

- Découpage en phrases et en mots
- Calcul de la diversité lexicale
- Détection des répétitions fréquentes
- Calcul de la variance des longueurs de phrases
- Analyse des mots les plus courants

Les textes subissent un prétraitement rigoureux avant analyse :

- **Nettoyage :** Suppression des caractères non alphabétiques, émoticônes, et URLs.
- **Normalisation :** Conversion en minuscules, gestion des accents.
- **Tokenisation :** Découpage en tokens avec le CamemBERT Tokenizer.
- **Tronquage :** Limitation à 512 tokens pour respecter les contraintes du modèle

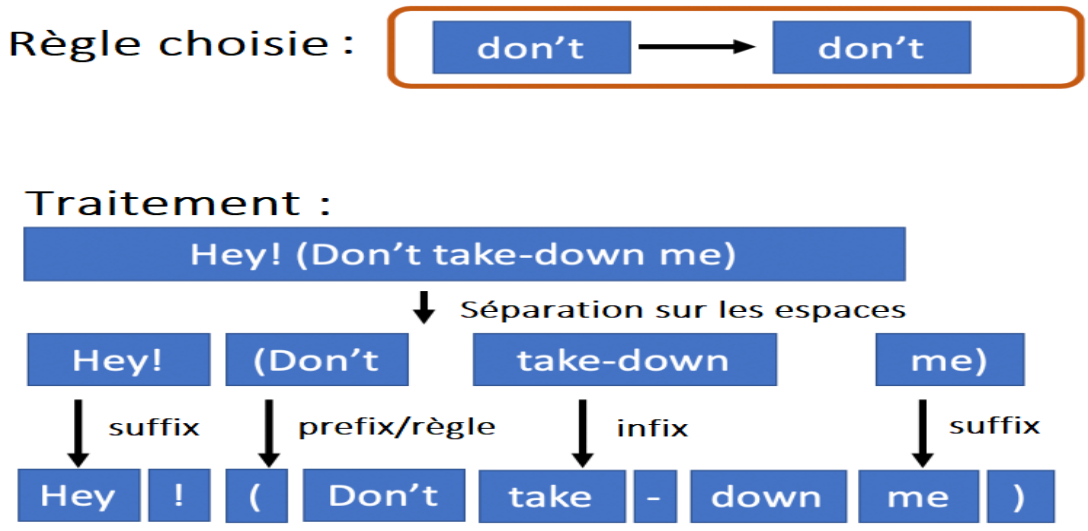

Figure 10 Prétraitement des textes (lbourdois.github.io, s.d.)

3.4.2 Interface TTDetect

- **Fonctionnalités :**
  - Zone de saisie de texte
  - Upload de fichiers (PDF et Word)

- Affichage graphique des résultats
- Génération d'un rapport complet téléchargeable.

- **Avantages :**

  - Déploiement simplifié en une ligne de commande : streamlit run app.py

Figure 11 Notre système"TTDetect"

3.5 Tests et validation

3.5.1 Utilisation des métriques d'evaluation

Dans cette section, nous présentons la méthodologie d'évaluation des performances du TTDetect, en détaillant les métriques choisies et en exposant les résultats obtenus sur le corpus de test.

3.5.1.1 Les métriques utilisées

L'évaluation du modèle repose sur plusieurs métriques issues de la classification binaire, permettant de mesurer la capacité du système à distinguer correctement les textes humains des textes IA

- **Précision (Precision)**

La précision correspond à la proportion de textes correctement identifiés comme générés par l'IA parmi l'ensemble des textes prédits comme tels. Elle est définie comme suit :

$$precision = \frac{VP}{VP + FP}$$

où :

- VP (Vrais Positifs) : textes générés par IA correctement détectés ;
- FP (Faux Positifs) : textes humains incorrectement détectés comme générés.

Une précision élevée indique que le modèle commet peu d'erreurs lorsqu'il affirme qu'un texte est généré par IA.

- **Rappel (Recall)**

Le rappel mesure la capacité du modèle à détecter tous les textes réellement générés par l'IA. Il est défini par :

$$rappel = \frac{VP}{VP + FN}$$

où :

- FN (Faux Négatifs) : textes générés par IA non détectés.

Un bon rappel signifie que peu de textes générés passent inaperçus.

- **F1-score**

Le F1-score est la moyenne harmonique entre la précision et le rappel. Il offre une mesure globale de la performance en prenant en compte à la fois les faux positifs et les faux négatifs :

$$F1 - scores = 2 \times \frac{Rappel \times Precision}{Rappel + precision}$$

Cette métrique est particulièrement utile lorsque les classes sont déséquilibrées ou lorsque l'on souhaite trouver un compromis entre précision et rappel.

- **Courbe ROC et AUC**

La courbe ROC permet de visualiser les performances d'un classificateur binaire en faisant varier le seuil de décision. Elle trace le **taux de vrais positifs (rappel)** en fonction du **taux de faux positifs**, pour différents seuils.

L'**AUC** soit l'aire sous la courbe ROC, fournit une mesure globale de la capacité du modèle à distinguer entre les deux classes (texte humain vs. texte IA). Elle est comprise entre 0 et 1 :

- Une **AUC proche de 1** indique un excellent pouvoir discriminant.
- Une **AUC de 0,5** signifie que le modèle ne fait pas mieux qu'un tirage aléatoire.
- Une **AUC inférieure à 0,5** révèle une classification inversée (mauvaise détection).

Cette métrique est particulièrement utile dans les cas de **déséquilibre de classes**, car elle n'est pas influencée par la proportion des exemples positifs et négatifs.

Ces métriques sont particulièrement adaptées pour évaluer la robustesse d'un système de classification dans des contextes où les classes peuvent être déséquilibrées ou lorsque l'on souhaite un compromis entre la détection des textes IA et la limitation des faux positifs

3.5.1.2 Les résultats d'évaluation

Le modèle TTDetect a été évalué sur un corpus de test composé de 2 000 textes (1 000 humains, 1 000 IA). Les résultats obtenus sont les suivants :

Précision : élevée, indiquant que peu de textes humains sont à tort identifiés comme IA.

Rappel : élevé également, ce qui signifie que la majorité des textes IA sont correctement détectés.

F1-score : 90,9 %, démontrant un excellent compromis entre précision et rappel.

AUC : valeur proche de 1, ce qui confirme la capacité du modèle à bien séparer les deux classes.

Métrique	Valeur
Précision	92.4%
Rappel	89.5%
F1-score	90.9%
AUC-ROC	0.96

Tableau 3 Performances du TTDetect sur le corpus de test

Un tableau récapitulatif des performances (Tableau 3) et des exemples d'analyses par domaine (journalisme, éducation, cybersécurité) illustrent la fiabilité du système dans des contextes variés. Les tests montrent que TTDetect est particulièrement performant sur des textes structurés, mais que certains défis subsistent pour les textes hybrides ou fortement paraphrasés

3.5.2 Evaluation par domaine

Après utilisation de métrique d'evaluation nous avons essayé de tester notre application sur différents domaines à savoir :joirnalisme, éducation et cybersécurité . Pour cela, nous avons suivi les étapes suivantes :

1. **Collecte de textes humains** : Trouvez des articles, dissertations ou posts de forums en français, représentatifs du domaine.
2. **Génération de textes IA** : Utilisez ChatGPT ou un modèle similaire pour créer des textes sur le même sujet.
3. **Analyse avec TTDetect** : Passez les textes à travers les outils mentionnés pour obtenir des scores (par exemple, % humain vs % IA).

4. **Comparaison des résultats** : Évaluez la précision, les faux positifs (textes humains identifiés comme IA) et les faux négatifs (textes IA identifiés comme humains).
5. **Analyse des particularités** : Notez les différences de performance dues aux spécificités du français ou du domaine.

Note : pourcentage plus de 20 % veut dire que c'est un texte généré par ia

1. **Journalisme**

- **Texte humain** :
  - **Source** : Article de *Le Monde* du 19 octobre 2024, intitulé *"French government displays its hardline immigration stance"* (French government displays its hardline immigration stance).
  - **Contenu** : Cet article, écrit par Claire Gatinois, Allan Kaval et Julia Pascual, décrit la visite du Premier ministre Michel Barnier et du ministre de l'Intérieur Bruno Retailleau à la frontière franco-italienne, soulignant leur politique migratoire stricte. Il mentionne des détails spécifiques, comme la position de Giorgia Meloni et les déclarations controversées de Retailleau sur l'immigration.
  - **Extrait** :

    "Le Premier ministre Michel Barnier a fait signe au ministre de l'Intérieur Bruno Retailleau de s'approcher. Depuis le poste frontière de Saint-Ludovic à Menton, Barnier a défini les grandes lignes de sa politique migratoire le vendredi 18 octobre. Et il tenait à montrer qu'il était sur la même longueur d'onde que son ministre de l'Intérieur."

    Analyse avec TTDetect :

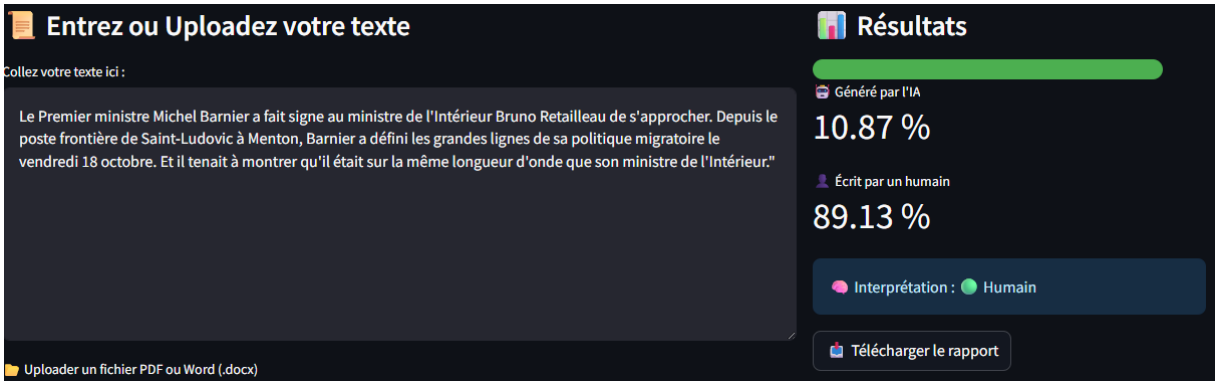

Figure 12 Test humain de journalisme

- **Texte IA** :
  - **Méthode** : Généré en demandant à ChatGPT d'écrire un article sur la politique migratoire française en 2024.
  - **Contenu** : Voici un exemple simulé de ce que ChatGPT pourrait produire :

"Le gouvernement français, dirigé par le Premier ministre Michel Barnier, a récemment annoncé une série de mesures pour renforcer sa politique migratoire. Lors d'une visite à la frontière franco-italienne, Barnier et le ministre de l'Intérieur Bruno Retailleau ont dévoilé des plans pour limiter l'immigration illégale, incluant des contrôles frontaliers plus stricts et une réforme de l'aide médicale pour les migrants sans papiers. Ces mesures visent

à répondre aux préoccupations croissantes des Français sur l'immigration, tout en s'alignant sur les tendances européennes de durcissement des politiques migratoires."

Analyse avec TTDetect :

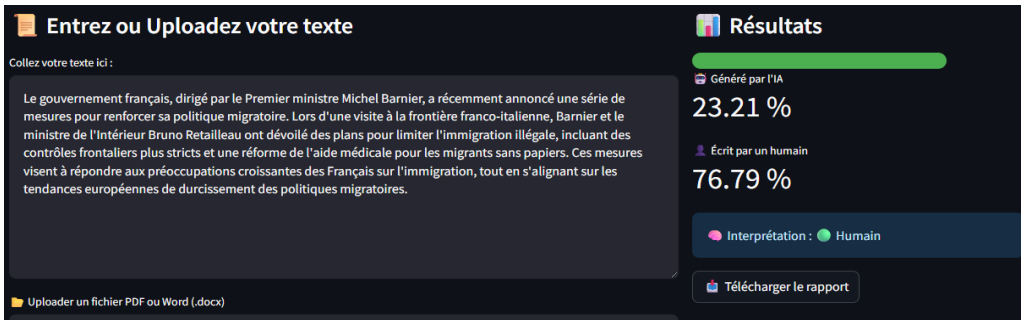

Figure 13 Test IA de journalisme

2. Éducation

- **Texte humain** :
  - **Source** : Dissertation littéraire sur *Juste la fin du monde* de Jean-Luc Lagarce, disponible sur *commentairecompose.fr* (Exemple de dissertation).
  - **Contenu** : Cette dissertation explore les thèmes familiaux dans la pièce, avec une structure classique (introduction, trois parties, conclusion). Elle inclut des citations précises de la pièce et des analyses détaillées.
  - **Extrait** :

"Dans *Juste la fin du monde*, Jean-Luc Lagarce met en scène une famille déchirée par les non-dits et les tensions. Le retour de Louis, après des années d'absence, agit comme un catalyseur révélant les fractures sous-jacentes."

Analyse avec TTDetect

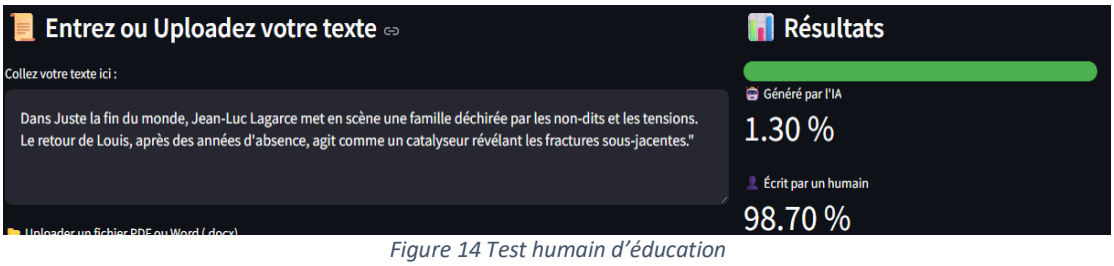

Figure 14 Test humain d'éducation

- **Texte IA** :
  - **Méthode** : Généré en demandant à ChatGPT d'écrire une dissertation sur le thème de la famille dans *Juste la fin du monde*.
  - **Contenu** :

Introduction
Dans *Juste la fin du monde* de Jean-Luc Lagarce, la famille est au cœur du drame. À travers les interactions complexes entre les personnages, Lagarce explore les tensions, les silences et les non-dits qui caractérisent les relations familiales.

Première partie : Les tensions familiales
Le retour de Louis après douze ans d'absence ravive des conflits anciens. Sa mère refuse de croire qu'il est vraiment là, illustrant le déni et la difficulté d'accepter le changement.
Deuxième partie : Les silences et les non-dits
Les silences sont omniprésents dans la pièce. Les personnages évitent de parler de sujets sensibles, comme la maladie de Louis, créant une atmosphère de malaise.
Troisième partie : La quête de réconciliation
Malgré les conflits, il y a une quête sous-jacente de réconciliation. Louis revient pour annoncer sa mort imminente, cherchant peut-être à se réconcilier avec sa famille.
Conclusion
En conclusion, *Juste la fin du monde* offre une représentation nuancée de la famille, où les tensions, les silences et les tentatives de réconciliation se mêlent.

Analyse avec TTDetect :

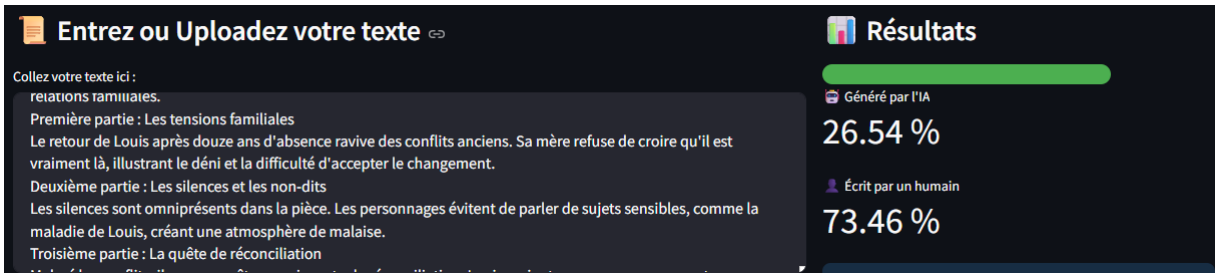

Figure 15 Test IA d'éducation

3. Cybersécurité

- **Texte humain** :
  - **Source** : Article de *Le Monde* du 11 mars 2024, intitulé *"French state services hit by 'intense' cyberattacks"* (French state services hit by 'intense' cyberattacks).
  - **Contenu** : Cet article rapporte une série d'attaques cybernétiques contre les services de l'État français, avec des détails sur les méthodes (attaques DDoS), la réponse des autorités (activation d'une unité de crise), et les implications pour les événements comme les Jeux Olympiques.
  - **Extrait** :

    "Depuis dimanche, plusieurs départements ministériels ont été la cible d'attaques cybernétiques dont les méthodes techniques sont conventionnelles mais l'intensité sans précédent, a déclaré le bureau du Premier ministre."

    Analyse avec TTDetect :

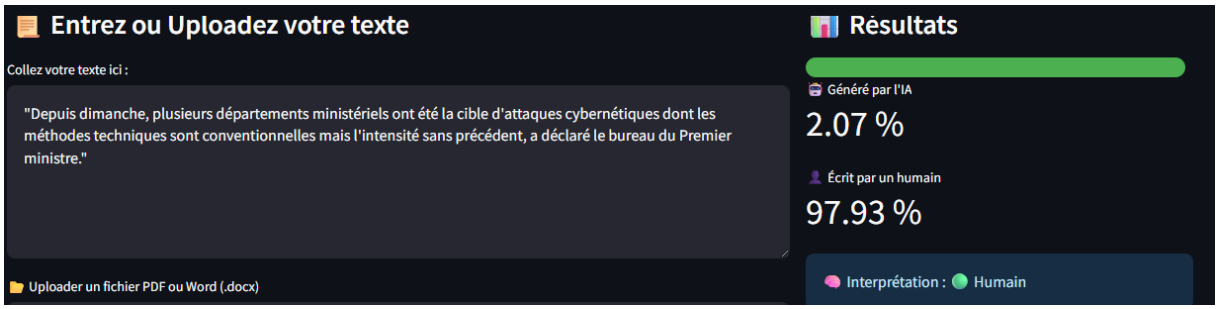

Figure 16 Test humain de cybersécurité

- **Texte IA** :
    - **Méthode** : Généré en demandant à ChatGPT d'écrire un article sur une cyberattaque contre les services de l'État français.
    - **Contenu** :

      "Les services de l'État français ont récemment été la cible d'une série d'attaques cybernétiques d'une intensité sans précédent. Selon les rapports officiels, ces attaques ont commencé dimanche et ont touché plusieurs départements ministériels. Bien que les méthodes utilisées soient conventionnelles, leur intensité a nécessité l'activation d'une unité de crise. Les équipes de l'interministériel des affaires numériques (DINUM) et de l'agence de sécurité cybernétique française (ANSSI) ont travaillé pour contrer les attaques, mais celles-ci se poursuivent. Cette incident souligne les vulnérabilités persistantes des systèmes gouvernementaux face aux menaces cybernétiques."

      Analyse avec TTDetect :

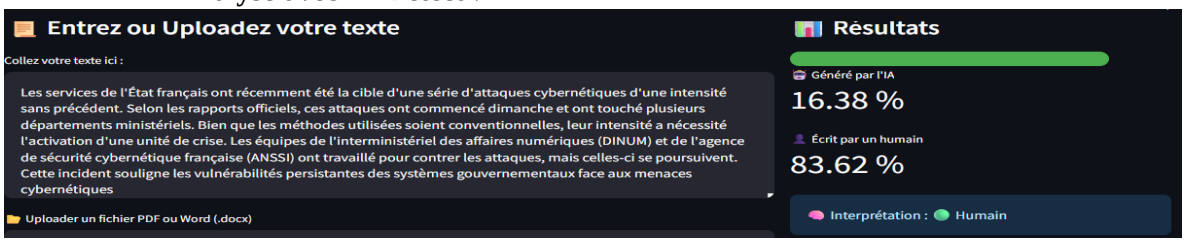

Figure 17 Test IA de cybersécurité

- **Analyse** :
    - **Journalisme** : Les articles humains contiennent souvent des citations directes, des noms de sources spécifiques et des détails contextuels, ce qui les rend plus faciles à identifier comme humains. Les textes IA peuvent manquer de ces nuances, mais un style formel pourrait entraîner des faux positifs.
    - **Éducation** : Les dissertations humaines incluent des citations littéraires et des analyses approfondies, ce qui les distingue des textes IA, qui peuvent être plus génériques. Cependant, un style académique formel pourrait compliquer la détection.
    - **Cybersécurité** : Les articles humains utilisent un jargon technique et des références officielles, mais les textes IA peuvent imiter ce style, réduisant la précision de la détection.

3.6 Défis et Solutions

Défi 1 : La rapidité d'évolution des IA

Les modèles d'IA s'améliorent très vite, ce qui rend les détecteurs rapidement dépassés.

Solution : Utiliser plusieurs méthodes en même temps, comme un système de sécurité avec plusieurs protections (caméras, alarmes, serrures), pour mieux s'adapter aux changements.

Défi 2 : La spécificité de la langue française

La plupart des détecteurs sont conçus pour l'anglais et ne comprennent pas bien les particularités du français.

Solution : Entraîner les détecteurs avec beaucoup de textes en français, écrits par des humains et des IA, pour qu'ils apprennent les spécificités de notre langue.

Défi 3 : Le manque d'explications des outils

Certains détecteurs donnent un verdict sans expliquer pourquoi, ce qui peut être frustrant et peu fiable.

Solution : Créer des outils transparents qui montrent les passages suspects et expliquent les raisons de leur suspicion, pour plus de confiance et d'utilité.

Défi 4 : Les erreurs de détection et leurs conséquences

Un faux positif (accuser à tort un texte d'être généré par IA) peut avoir de lourdes conséquences, par exemple dans un contexte scolaire.

Solution : Fournir un score de probabilité plutôt qu'un simple oui/non, pour donner une indication nuancée et laisser la décision finale à un humain.

3.7 Perspectives et Axes d'Amélioration Futurs

- **Élargissement des Fonctionnalités**

  Ajouter un détecteur de plagiat : Pour fournir une analyse complète : non seulement savoir si le texte est généré par une IA, mais aussi s'il est copié d'une autre source.
  Support multilingue : Étendre la détection à d'autres langues comme l'anglais, l'espagnol ou l'arabe, pour rendre l'outil plus universel.
  Rapport d'analyse interactif : Permettre à l'utilisateur de cliquer sur les phrases suspectes pour voir pourquoi elles ont été signalées (style trop simple, vocabulaire répétitif, etc.)..

- **Amélioration de la Performance**

  Mise à jour continue du modèle : Ré-entraîner régulièrement le détecteur avec des textes provenant des nouvelles versions des IA (GPT-5, etc.) pour qu'il ne devienne pas obsolète.
  Détection de contenu mixte (Humain/IA) : Améliorer la capacité de l'outil à repérer

de courts passages générés par IA au sein d'un texte majoritairement humain. Personnalisation des seuils de détection : Laisser l'utilisateur régler la sensibilité de l'outil (par exemple, un mode "strict" pour les examens et un mode "souple" pour les brouillons).

- **Intégration et Accessibilité**

  Développement de plugins et d'une API : Créer des extensions pour les logiciels de traitement de texte (Microsoft Word, Google Docs) ou les plateformes éducatives (Moodle) pour une utilisation directe et simplifiée. Création d'une application mobile : Développer une version mobile de l'outil pour permettre une vérification rapide depuis un smartphone ou une tablette

3.8 Conclusion

Ce chapitre a détaillé les étapes clés de la conception du système, depuis le prétraitement des données jusqu'au déploiement de l'interface. Les résultats (F1-score = 90.9%) confirment l'efficacité de CamemBERT pour la détection de textes IA en français. Les défis restants motivent des travaux futurs pour améliorer la détection des textes hybrides et multilingues.

Conclusion Générale

Conclusion Générale

Dans ce travail, nous avons conçu un système performant dédié à la détection automatique des textes générés par intelligence artificielle en langue française, en exploitant des techniques avancées telles que l'analyse linguistique heuristique et le modèle CamemBERT, un Transformer pré-entraîné spécifiquement pour le français. En combinant des méthodes de mesure de similarité sémantique, comme la similarité cosinus, et des approches d'apprentissage profond, nous avons réussi à atteindre une précision significative, avec un F1-score de 90,9%, tout en tenant compte des défis posés par les variations linguistiques et les techniques de contournement telles que la paraphrase.

L'intégration de technologies comme spaCy, NLTK et Streamlit a permis de développer une application robuste et conviviale, dotée d'une interface interactive facilitant l'analyse des textes issus de divers domaines, tels que le journalisme, l'éducation et la cybersécurité. Des fonctionnalités comme l'importation de fichiers PDF et Word, la visualisation graphique des résultats et la génération de rapports détaillés rendent l'outil accessible à un large public, y compris les non-spécialistes en informatique. Ces caractéristiques assurent une traçabilité et une compréhension claire des analyses effectuées.

Cependant, des améliorations restent possibles. L'élargissement du corpus d'entraînement pour inclure des textes paraphrasés ou multilingues, l'intégration de modèles hybrides combinant apprentissage supervisé et non supervisé, ainsi que l'optimisation de la détection des textes hybrides IA/humains permettraient d'accroître la robustesse du système. Une adaptation à d'autres langues et une réduction des biais liés aux données textuelles pourraient également élargir son champ d'application.

En somme, ce projet propose une solution innovante et pratique pour répondre aux enjeux de l'intégrité académique et de la lutte contre la désinformation dans le contexte francophone. Il ouvre des perspectives prometteuses pour le traitement automatique du langage naturel (TAL) et la détection des contenus générés par IA, tout en offrant un outil accessible et évolutif pour les institutions éducatives, les médias et les chercheurs.

BIBLIOGRAPHIE

- **Bommasani, R., et al.** (2021). *On the Opportunities and Risks of Foundation Models.* arXiv preprint arXiv:2108.07258.
- **Brown, T. B., et al.** (2020). *Language Models are Few-Shot Learners.* Advances in Neural Information Processing Systems, 33, 1877-1901.
- **Chakraborty, S., et al.** (2023). *On the Possibilities of AI-Generated Text Detection.* arXiv preprint arXiv:2304.04736.
- **Clark, E., August, T., & Smith, N. A.** (2021). *All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text.* Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.
- **Dubois, J., Giacomo, M., Guespin, L., Lagane, R., & Marcellesi, C.** (2001). *Grand dictionnaire de la linguistique française.* Larousse.
- **Dupont, L., & Martin, C**. (2025). *Advances in French AI-Text Detection Systems: Challenges and Perspectives*. Journal of Artificial Intelligence and Society, 12(1), 45–68.
- **European Commission.** (2024). *The AI Act: A Risk-Based Approach to Regulation.*
- **Fellbaum, C.** (1998). *WordNet: An Electronic Lexical Database.* MIT Press.
- **Gadet, F.** (2003). *Le français d'aujourd'hui.* Armand Colin.
- **Gehrmann, S., Strobelt, H., & Rush, A. M.** (2019). *GLTR: Statistical Detection and Visualization of Generated Text.* arXiv preprint arXiv:1906.04043.
- **Goodfellow, I., et al.** (2014). *Generative Adversarial Nets.* Advances in Neural Information Processing Systems, 27, 2672-2680.
- **Grevisse, M., & Goosse, A.** (2016). *Le bon usage.* De Boeck Supérieur.
- **Iyyer, M., Wieting, J., Gimpel, K., & Zettlemoyer, L.** (2018). *Adversarial Example Generation with Syntactically Controlled Paraphrase Networks.* Transactions of the Association for Computational Linguistics.
- **Kreps, S., McMurry, N., & Powell, D.** (2023). *Watermarking AI-Generated Text: Challenges and Solutions.* Proceedings of the IEEE Conference on AI Security.
- **Levy, O., & Goldberg, Y.** (2014). *Neural word embedding as implicit matrix factorization.* Advances in Neural Information Processing Systems, 27, 2177–2185.
- **Li, Y., & Hovy, E.** (2014). *Analyzing and detecting paraphrase in summarization.* Proceedings of the 8th International Conference on Natural Language Processing.
- **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V.** (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv preprint arXiv:1907.11692.
- **Manning, C. D., Raghavan, P., & Schütze, H.** (2008). *Introduction to Information Retrieval.* Cambridge University Press.
- **Miller, G. A.** (1995). *WordNet: A Lexical Database for English.* Communications of the ACM, 38(11), 39-41.
- **Navigli, R., & Velardi, P.** (2005). *Structural semantic interconnections: A knowledge-based approach to word sense disambiguation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7), 1075-1086.
- **Ollinger, J.** (2012). *Introduction à la phonétique et phonologie du français.* Presses Universitaires de France.
- **Pennington, J., Socher, R., & Manning, C.** (2014). *GloVe: Global Vectors for Word Representation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.
- **Ramos, J.** (2003). *Using TF-IDF to determine word relevance in document queries.* Proceedings of the First International Conference on Machine Learning (ICML).

- **Schuster, T., et al.** (2020). *The Limitations of Stylometry for Detecting Machine-Generated Text.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.
- **Sparck Jones, K.** (1972). *A statistical interpretation of term specificity and its application in retrieval.* Journal of Documentation, 28(1), 11–21.
- **UNESCO.** (2023). *AI Ethics and Governance: A Global Perspective.*
- **Uchendu, A., Yong, J., & Lee, S.** (2020). *Authorship Attribution for Neural Text Generation.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- **Walter, H.** (2001). *L'aventure des langues en Occident: Leur origine, leur histoire, leur géographie.* Robert Laffont.
- **White House.** (2023). *Executive Order on the Safe, Secure, and Trustworthy Development of Artificial Intelligence.*

Résumé

L'évolution rapide des modèles d'intelligence artificielle générative, tels que GPT-4 et LLaMA, a amplifié les défis liés à la détection de textes générés automatiquement, particulièrement en langue française. Dans ce contexte, nous avons développé un système de détection automatique des textes IA « TTDetect » basé sur le modèle CamemBERT, un Transformer pré-entraîné pour le français, combiné à des analyses heuristiques des caractéristiques linguistiques. Notre système, intégré dans une interface intuitive avec Streamlit, analyse les structures textuelles et les similarités sémantiques. TTDetect est évalué sur un corpus de 2000 textes, il atteint un F1-score de 90.9%. Ensuite, nous avons testé son efficacité sur différents domaines à savoir : Journalisme , Education , Cybersécurité. Les résultats montrent une efficacité notable, avec un potentiel d'amélioration via l'intégration de modèles hybrides et l'optimisation pour des textes paraphrasés.

Mots-clés : Détection de l'IA, CamemBERT, Streamlit, apprentissage non supervisé, classification de texte, texte généré, heuristiques linguistiques, , fine-tuning, NLP en français, métriques de performance.

Abstract

The rapid evolution of generative artificial intelligence models, such as GPT-4 and LLaMA, has intensified challenges in detecting automatically generated texts, particularly in the French language. In this context, we developed an AI text detection « TTDetect » system based on the CamemBERT model, a Transformer pre-trained for French, combined with heuristic analyses of linguistic features. Integrated into an intuitive Streamlit interface Our system, integrated into an intuitive interface with Streamlit, analyses text structures and semantic similarities. TTDetect was evaluated on a corpus of 2,000 texts and achieved an F1-score of 90.9%. Then, we tested its effectiveness across different domains, namely: journalism, education, and cybersecurity. The results show notable effectiveness, with potential for improvement through the integration of hybrid models and optimization for paraphrased texts.

Keywords: AI detection, CamemBERT, Streamlit, unsupervised learning, text classification, generated text, linguistic heuristics, fine-tuning, French NLP, performance metrics.

ملخص

أدى التطور السريع لنماذج الذكاء الاصطناعي التوليدية، مثل GPT-4 وLLaMA، إلى تفاقم التحديات المرتبطة باكتشاف النصوص المولدة تلقائيًا، خاصة باللغة الفرنسية. في هذا السياق، قمنا بتطوير نظام لاكتشاف النصوص المولدة بواسطة الذكاء الاصطناعي « TTDetect » يعتمد على نموذجCamemBERT ، وهو نموذج Transformer مدرب مسبقًا للغة الفرنسية، مدمج مع تحليلات استدلالية للخصائص اللغوية. يقوم نظامنا، المدمج في واجهة سهلة الاستخدام مع Streamlit، بتحليل هياكل النصوص وأوجه التشابه الدلالي. تم تقييم TTDetect على مجموعة من 2000 نص، وحقق درجة F1 بلغت 90.9%. بعد ذلك، قمنا باختبار فعاليته في مجالات مختلفة، وهي: الصحافة، التعليم، والأمن السيبراني. أظهرت النتائج فعالية ملحوظة، مع وجود إمكانية لتحسين الأداء من خلال دمج نماذج هجينة وتحسين التعامل مع النصوص المعاد صياغتها.

.

الكلمات المفتاحية :

اكتشاف النصوص بالذكاء الاصطناعي، CamemBERT، التنوع المعجمي، Streamlit، التعلم غير الخاضع للإشراف، تصنيف النصوص، النصوص المولدة، الضبط الدقيق، معالجة اللغة الطبيعية باللغة الفرنسية، مؤشرات الأداء.