Évaluation de la pertinence des citations dans les articles scientifiques

Qinyue Liu

Évaluation de la pertinence des citations dans les articles scientifiques

Qinyue
LIU

Sous la direction de Cyril LABBE, Amira BARHOUMI et Claude PONTON

Laboratoire : Laboratoire d'informatique de Grenoble (lig)

UFR LLASIC
Département Sciences du langage

Mémoire de master 2 mention Sciences du langage 30 crédits

Parcours : Industries de la langue

Année universitaire 2022-2023

Évaluation de la pertinence des citations dans les articles scientifiques

Qinyue
LIU

Sous la direction de Cyril LABBE, Amira BARHOUMI et Claude PONTON

Laboratoire : Laboratoire d'informatique de Grenoble (lig)

UFR LLASIC
Département Sciences du langage

Mémoire de master 2 mention Sciences du langage 30 crédits

Parcours : Industries de la langue

Année universitaire 2022-2023

Remerciements

Je souhaite exprimer ma profonde reconnaissance envers mes superviseurs de stage, Cyril Labbé et Amira Barhoumi, pour leur soutien précieux tout au long de la durée de ce stage. Leurs conseils avisés, leur expertise et leurs encouragements constants ont joué un rôle primordial dans la réalisation fructueuse de ce projet de recherche.

Je tiens à adresser mes remerciements à Claude Ponton, qui s'est montré toujours aimable et patient envers les étudiants, et qui m'a apporté son aide tout au long de mon parcours de master.

Je tiens également à remercier l'équipe de Sigma pour m'avoir offert cette opportunité de stage, qui a été une expérience enrichissante et formatrice. Je suis reconnaissante pour leur accueil chaleureux et leur collaboration tout au long de mon séjour.

Enfin, mes remerciements s'adressent aussi à mes proches, amis et collègues qui m'ont apporté un soutien moral et encouragé à persévérer. Leur présence positive a été un moteur précieux dans la réalisation de ce travail.

DÉCLARATION ANTI-PLAGIAT

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

PRENOM : ……QINYUE……

NOM : ……LIU……

DATE : ……25/08/2023……

Table des matières

1 Introduction

Le sujet du stage consiste à évaluer des citations dans les articles scientifiques en utilisant les techniques de Traitement Automatique des Langues (TAL).

Les citations dans les articles scientifiques sont couramment utilisées pour divers objectifs, tels que contextualiser la recherche, référencer les publications précédentes et remettre éventuellement en questions les résultats des travaux existants (Jurgens et al., 2018). Elles facilitent également le suivi des avancées scientifiques et aident les lecteurs à développer un cadre pour la formulation d'hypothèses de recherche (Horbach, Aagaard et Schneider, 2021). Néanmoins, certaines études montrent l'existence de quelques erreurs de citationsdans les articles. Par exemple, (Jergas et Baethge, 2015) ont examiné 28 articles dans des journaux médicaux, et ont trouvé un taux d'erreur de 25,4% dans les citations. Une autre étude (Kristof, 1997) s'est penchée sur l'évaluation des citations dans des articles journalistiques (49 articles). Cette étude révèle un taux d'erreur de 30,1%.

Plusieurs études utilisant le traitement automatique des langues (TAL) ont analysé les citations. Dans l'étude de Liu (2017), une analyse de sentiment sur les citations a été effectuée avec un classifieur basé sur différentes représentations vectorielles. Te et al. (2022) ont classé les contextes de citations en catégorie « critiques »ou « non critiques ». Cependant, il n'existe pas encore beaucoup de recherches évaluant la pertinence des citations. Dans ce contexte, notre étude vise à mesurer automatiquement la pertinence des citations dans leurs contextes afin de différencier les citations fiables et celles erronées.

Dans ce stage, nous posons les questions de recherche suivantes :

— Quelles sont les éventuelles erreurs de citations dans les articles scientifiques ? Comment les catégoriser ?

— Comment vérifier, dans les papiers cités, l'information véhiculée dans un

contexte de citation ? Quelles parties des papiers cités (abstract, introduction, méthodes, résultats ou conclusion) reflètent au mieux l'information du contexte de citation ?

— Comment distinguer, dans un contexte de citation donné, les citations fiables des erronées ?

Sur le côté technique, nous posons les questions suivantes :

— Quelles sont les techniques TAL possibles pour évaluer les citations dans leurs contextes ?

— Quels modèles de langues peuvent être adoptés ?

— Comment construire un corpus pour la tâche de mesure de pertinence des citations ?

Dans ce mémoire, nous élaborons tout d'abord une étude de l'existant pour l'analyse de citations de façon générale et plus particulièrement pour la tâche de mesure de pertinence des citations dans leurs contextes (voir Partie 3). Ensuite, nous décrivons la construction du corpus pour notre tâche (voir Partie 4). Puis, nous proposons deux méthodes pour distinguer les citations fiables de celles erronées (voir Partie 5). Nous évaluons nos méthodes et discuterons de leurs performances (voir Partie 6). Enfin, nous concluons notre travail tout en donnant quelques perspectives (voir Partie 7).

Partie 1
-
Contexte et Sujet du stage

2 Contexte et Sujet du stage

Ce mémoire fait partie de mon stage de fin d'étude de Master Sciences du langage parcours Industries de la langue (IDL). Le stage est encadré par l'équipe SIGMA du Laboratoire d'Informatique de Grenoble (LIG), sous la supervision de Monsieur Cyril Labbé (LIG) et Madame Amira Barhoumi (LIG) dans le cadre du projet NanoBubbles

2.1 Laboratoire d'Informatique de Grenoble

Le Laboratoire d'Informatique de Grenoble (LIG) est un laboratoire centré sur les sciences informatiques. Les recherches concentrent sur 5 axes :

— Génie des logiciels et des systèmes d'information
— Méthodes formelles, modèles et langues
— Systèmes intelligents pour les données, les connaissances et les humains
— Systèmes interactifs et cognitifs
— Systèmes répartis, calcul parallèle et réseaux

Le LIG contribue au développement des aspects fondamentaux de l'informatique (modèles, langages, méthodes, algorithmes) et au développement d'une synergie entre les enjeux conceptuels, technologiques et sociétaux associés à cette discipline.

2.2 Équipe SIGMA

L'équipe SIGMA se concentre sur les Systèmes d'Information (SI). Ces systèmes sont omniprésents dans la vie quotidienne et jouent un rôle central dans les stra-

tégies et l'organisation des entreprises. Dans ce contexte, l'équipe SIGMA mène des recherches approfondies sur les SI en combinant les aspects conceptuels, organisationnels et opérationnels. Ces recherches se focalisent sur la formalisation, la conception et l'infrastructure des systèmes d'information. Elles sont continuellement confrontées à des situations concrètes dans divers domaines tels que la santé, les transports, l'éducation et l'industrie.

2.3 Projet NanoBubbles

C'est un projet (voir le lien https ://nanobubbles.hypotheses.org/)dirigé par 4 chercheurs qui viennent de l'université Paris Sorbonne Nord, de l'université Maastricht, de l'université Grenoble-Alpes et de l'université Radboud, en collaboration avec des chercheurs du CNRS, de l'Université de Twente, de l'IRIT et de l'Ecole des Ponts.

Le projet se concentre sur comment, quand et pourquoi la science n'arrive pas à se corriger. Pour comprendre comment la correction de la science fonctionne ou échoue, le projet NanoBubbles est multidisciplinaire et combine des approches comme les sciences naturelles, l'ingénierie (traitement automatique des langages) et des sciences humaines et sociales (linguistique, sociologie, philosophie et histoire des sciences).

Partie 2
-
État de l'art

3 État de l'art

Cette partie est composée de 4 sous-sections. Tout d'abord, nous abordons l'exactitude des citations et examinerons les recherches existantes sur ce sujet dans la Section 3.1. Ensuite, dans la Section 3.2, nous explorons les représentations vectorielles. Puis, nous nous penchons sur les méthodes de mesure de similarité dans la Section 3.3. Enfin, nous présentons les métriques possibles pour évaluer la similarité dans la Section 3.4.

3.1 Citations

Comme précédemment mentionné, cette étude se concentre sur la précision des citations au sein des documents scientifiques. L'objectif est de comprendre le rôle des citations dans la recherche scientifique, d'identifier les sources d'erreurs et, en fin, de déterminer comment évaluer leur exactitude.

Dans la Section 3.1.1, nous abordons l'impact des citations sur les recherches. Ensuite, dans la Section 3.1.2, nous présentons les recherches existantes portant sur l'analyse statistique des citations, tandis que dans la Section 3.1.3, nous abordons l'analyse des contextes de citations. Enfin, dans la Section 3.1.4, nous proposons différentes classifications pour les citations erronées.

3.1.1 Impacts des citations

Une citation, c'est une référence précise à une source. L'utilisation de citations dans les publications scientifiques constitue une pratique largement répandue, remplissant divers objectifs pour les auteurs. À titre d'exemple, les citations sont uti-

lisées afin d'établir le contexte de la recherche, de renvoyer aux méthodes utilisées ou de remettre en questions les résultats (Jurgens et al., 2018). Les citations permettent également de suivre les avancées de la science et aident les lecteurs à élaborer un cadre pour établir des hypothèses (Horbach, Aagaard et Schneider, 2021). Les citations inexactes peuvent entraîner des interprétations erronées, déformer les intentions de l'auteur original et même avoir des conséquences potentiellement plus graves. Une étude menée par Horbach, Aagaard et Schneider (2021) a évoqué les impacts des citations en prenant pour exemple l'article de John Study (John, Loewenstein et Prelec, 2012) comme exemple. Horbach, Aagaard et Schneider (2021)a révélé qu'avec le temps, les articles les plus cités tendent à devenir un concept conventionnel. De plus, le manque d'engagement critique conduit certaines études à être citées bien plus fréquemment que ce que leur contribution académique justifie. Selon cette recherche, cela porte préjudice non seulement au système de récompense interne de la science, mais conduit également à une réduction de la diversité épistémique.

3.1.2 Études statistiques de citations

Dans l'étude menée par Jergas et Baethge (2015), les erreurs de citation ont d'abord été divisées en deux catégories : majeures et mineures. Ils ont conclu que les erreurs majeures sont celles qui sont « *complètement en contradiction avec les affirmations des auteurs* » (Jergas et Baethge, 2015), tandis que les erreurs mineures sont des « *incohérences et des erreurs factuelles qui ne sont pas suffisamment graves pour contredire une déclaration des auteurs cités* » (Jergas et Baethge, 2015). Ils ont principalement analysé 27 articles (plus un article supplémentaire), dont 15 qui ont été analysés par au moins deux chercheurs indépendamment. Pour ces 27 articles, le taux d'erreur médian des erreurs majeures est de 11,5% et le taux d'erreur médian

des erreurs mineures est de 9,6%. Une autre recherche (KRISTOF, 1997) évoque un taux d'erreur de 30.1% pour les citations dans les papiers journalistes en analysant 49 papiers, dont trois papiers qui n'ont pas d'erreur de citation. Dans une autre étude menée par ARMSTRONG et al. (2018), les chercheurs ont évalué 50 références sélectionnées au hasard et publiées dans la revue OHNS *(Otolaryngology-Head and Neck Surgery)*. Ils ont analysé les erreurs de citation. Les erreurs de citation se sont présenté dans 17% de toutes les références étudiées, dont 34% qui ont été classées comme majeures.

3.1.3 Études contextes de citation

Nous présentons différentes recherches existantes liées à l'analyse des citations. Certaines de ces recherches se concentrent sur la réalisation d'analyses statistiques des citations inexactes dans les articles scientifiques, tandis que d'autres se plongent dans des analyses complexes des contextes de citation. Par exemple, analyse des émotions dans les contextes de citations, classification des contextes des citations.

Les études que nous avons examinées et qui analysent en détail les contextes de citation ont utilisé des méthodes d'apprentissage automatique dans leur travail. LIU (2017) ont effectué une analyse de sentiment sur les citations en utilisant un classifieur basé sur différentes représentations vectorielles. Dans son travail, il a défini trois classes de sentiment pour les citations dans son ensemble de données, qui est extrait du Corpus de Référence de BIRD et al. (2008). Le tableau 1 présente sa classification des contextes de citation : « N »représente la classe de sentiment négatif, « P »représente la classe de sentiment positif, et « O »représente la classe de sentiment objectif. Cette analyse de sentiment vise à distinguer les citations de ces trois classes.

Classe	Nombre d'éléments
Positif	282
Négatif	419
Objectif	2 880

TABLE 1 – Corpus de LIU (2017)

Classe	Nombre d'éléments
Non critique	949
Critique	1 694

TABLE 2 – Corpus CitaNeg (TE et al., 2022)

De même, une étude des chercheurs (TE et al., 2022) vise à détecter les contextes de citation critiquant les sources citées. Les chercheurs ont affiné différents modèles de langage pour classer les contextes critiques et non critiques. En utilisant le corpus CitaNeg et le Corpus des Contextes Critiques, ils ont construit leur propre corpus. Les citations positives et neutres de CitaNeg ont été considérées comme non critiques, tandis que celles du Corpus des Contextes Critiques ont été considérées comme critiques. Comme décrit dans le tableau 2. Cela a donné lieu à 2643 citations, comprenant 949 exemples non critiques et 1694 exemples critiques.

3.1.4 Catégories de citations erronées

Une étude menée par CARLSEN et GLENTON (2019) a évalué les citations de leurs articles précédents (CARLSEN et GLENTON, 2011). Ils ont classé les erreurs de citation en deux catégories :

— Les citations où les auteurs ont utilisé les informations descriptives pour justifier leur *focus group*.

— Les citations dans lesquelles les auteurs ont fait référence à l'étude (CARLSEN et GLENTON, 2011) pour des raisons autres que le *focus group*, ou lorsque

Types d'erreurs	Description d'erreurs
Erreurs triviales	citations dans lesquelles les erreurs de transcription n'ont pas altéré ni obscurci le sens voulu de la source originale.
Erreurs légèrement trompeuses	citations qui ont conduit ou pourraient conduire à des malentendus, bien que les erreurs n'aient pas été assez importantes pour déformer complètement ou changer fondamentalement le sens de la source.
Erreurs graves	citations qui déforment considérablement ou ne présentent aucune similitude avec la source originale.

TABLE 3 – Catégories d'erreurs définies par(DE LACEY, RECORD et WADE, 1985)

l'intention n'était pas clairement évoquée.

Comme présenté dans le Tableau 3, une autre étude (DE LACEY, RECORD et WADE, 1985) a classé les erreurs de citation en trois groupes : erreurs triviales, erreurs légèrement trompeuses et erreurs graves. Nous croyons que la classification des erreurs de citations en différentes catégories pourrait être essentielle pour notre travail à venir.

3.2 Représentations vectorielles

Il existe plusieurs méthodes pour représenter des textes sous forme de vecteurs. Deux types de méthodes de représentation sont couramment utilisés : la représentation discrète et la représentation continue. Dans le cadre de la représentation discrète, chaque mot est considéré comme unique et indépendant des autres. Quelques exemples incluent TF-idf (3.2.1), One-Hot (3.2.2) et CountVectorizer (3.2.3).

En revanche, la représentation continue considère les relations mutuelles entre les mots, contrairement à la représentation discrète. Parmi les méthodes de représentation continue les plus utilisées, nous trouvons Word2Vec (3.2.4), GloVe (3.2.5), FastText (3.2.6), ainsi que des approches comme BERT (3.2.7) et GPT (3.2.8).

3.2.1 TF-idf

Le TF-IDF (Ramos, 2003) fonctionne en évaluant la fréquence relative des mots dans un document spécifique en comparant avec la proportion inverse de ce même mot dans l'ensemble du corpus du document. En d'autres termes, cette méthode détermine la pertinence d'un mot donné dans un document particulier.

Grâce à cette méthode, il devient plus aisé de filtrer les mots très fréquents, qui ont souvent peu d'importance sémantique dans un corpus, tels que les prépositions. Cependant, l'un des inconvénients de cette représentation est qu'elle ne tient pas compte de la position des mots.

3.2.2 One-Hot

Pour « One-Hot », chaque mot unique est représenté par un vecteur binaire de taille égale à la taille du vocabulaire total. Ce vecteur contient des zéros partout, sauf à une seule position qui correspond au mot spécifique. Cette position est définie comme « 1 »pour indiquer la présence du mot dans le texte et « 0 »pour indiquer son absence. Cette méthode est simple et efficace pour la représentation de mots individuels. Cependant, elle ne prend pas en compte les relations sémantiques ou contextuelles entre les mots, de plus, sa matrice de représentation peut être très coûteuse en mémoire et en efficacité si la taille de mots est énorme.

3.2.3 CountVectorizer

En calculant la fréquence des mots dans un document, cette méthode fournit une matrice qui contient le nombre d'occurrences des mots dans les phrases différentes. Plus précisément, chaque document est traité comme un ensemble de mots, et chaque mot unique est attribué à une colonne dans la matrice. Les valeurs dans la matrice indiquent combien de fois chaque mot apparaît dans chaque document. Avec cette méthode, nous pouvons obtenir une information précise sur la fréquence des mots. Dans une étude sur la détection de discours haineux (Turki et Roy, 2022), les auteurs ont extrait des caractéristiques à partir de données Twitter en utilisant cette technique de CountVectorizer.

Cependant, avec le 'CountVectorizer', il n'est pas possible de déterminer les positions des mots dans le document. De plus, les conjonctions et prépositions (qui peuvent s'appeler « stopwords ») ont souvent des fréquences très élevées dans un document, ce qui peut entraîner des biais. Par conséquent, il est recommandé d'effectuer un prétraitement pour supprimer ces "stopwords" avant d'appliquer cette

méthode.

3.2.4 Word2Vec

Une recherche menée par Mikolov et al. (2013) a proposé un modèle Word2Vec pour calculer les représentations vectorielles continues. L'idée principale de Word2Vec est que des mots similaires ont tendance à apparaître dans des contextes similaires. Word2Vec utilise des réseaux de neurones artificiels pour apprendre les vecteurs de mots. Il existe deux principales approches dans Word2Vec : CBOW *(Continuous Bag of Words)* et Skip-Gram. Dans l'approche CBOW, le modèle prédit un mot cible en fonction de son contexte. Pour Skip-Gram, le modèle prédit le contexte à partir d'un mot cible donné.

Le modèle Word2Vec présente une optimisation bien supérieure aux méthodes de représentation mentionnées précédemment. Il saisit la relation sémantique et syntaxique entre les différents mots, en traduisant ces informations en une dimension vectorielle constante. Cependant, il ne parvient pas à gérer les mots absents de son vocabulaire et la représentation d'un mot demeure invariable, ce qui signifie que cette représentation ne s'ajuste pas en fonction du contexte.

3.2.5 GloVe

GloVe (Pennington, Socher et Manning, 2014) *(Global Vectors for Word Representation)* est un modèle d'incorporation de mots basé sur des principes similaires à Word2Vec. Au lieu d'utiliser le CBOW ou Skip-Gram, GloVe est conçu pour factoriser directement la matrice de cooccurrence des mots, visant à capturer les ratios des probabilités de cooccurrence des mots. L'avantage principal de cette

méthode par rapport à Word2Vec réside dans sa prise en compte des relations entre les paires de mots plutôt que des relations entre les mots individuels. Cela peut conférer davantage de signification aux vecteurs de sortie. Cependant, comme il utilise une matrice de cooccurrence, le coût de la mémoire est plus important que pour Word2Vec.

3.2.6 FastText

Le modèle FastText a été développé par JOULIN et al. (2016) dans le laboratoire de Facebook. FastText est une représentation vectorielle améliorée et basée sur Word2Vec. FastText étend le modèle Word2Vec Skip-Gram en tenant compte des informations sur les sous-mots. Il considère les mots comme des ensembles de n grammes de caractères, ce qui lui permet de capturer les variations morphologiques et de mieux gérer les mots hors vocabulaire. Comme mentionné précédemment, Word2Vec n'arrive pas à représenter les mots qui ne sont pas dans son vocabulaire, alors FastText peut répondre à cette problématique. Une autre avantage de FastText est qu'il a été conçu pour entraîner rapidement les plongements de mots sur CPU.

3.2.7 BERT

BERT, en abréviation de *Bidirectional Encoder Representations from Transformers*, est largement utilisé afin de déterminer les représentations contextuelles des mots. C'est un modèle développé par DEVLIN et al. (2018). Selon eux,

> '*BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers*' (DEVLIN et al., 2018)

Cela veut dire que les plongements des textes dépendent des contextes à gauche et à droite, de ce fait, BERT donne en sortie des plongements contextuels. Le modèle standard est pré-entraîné sur des données provenant de la littérature anglaise et de Wikipédia, afin d'accomplir deux tâches : repérer les jetons masqués et prédire la phrase suivante. Il existe aussi des variantes de BERT en utilisant différentes données d'entraînement pendant le pré-entraînement pour s'adapter aux différentes utilisations.

Le modèle BERT (1) nécessite des vecteurs en entrée. Avant d'utiliser ce modèle, il est essentiel de tokeniser les textes bruts, que ce soit une simple phrase ou une paire de deux phrases. Ensuite, il faut fournir les vecteurs conformes au format requis par BERT. Plus précisément, BERT requiert un vecteur pour représenter les identifiants des tokens, un vecteur pour indiquer le contexte et le remplissage (*padding*), ainsi qu'éventuellement un vecteur pour différencier les différentes phrases.

Dans le processus de la tokenisation, les tokens spéciaux ([CLS] qui représente le début d'une phrase, et [SEP] qui représente la fin d'une phrase, ou qui sépare deux phrases en paire) seront ajoutés pour indiquer le début et la fin des phrases. Après, un vocabulaire de 30,000 tokens *(WordPiece)* sera adopté pour aider à ressortir les tokens. Selon les chercheurs (Devlin et al., 2018), pour un token, sa représentation d'entrée est construite en appliquant le token correspondant, le segment et les plongements de position.

Bert est basé sur l'architecture des Transformers(Vaswani et al., 2017). La figure 2 montre la structure de transformer. C'est une architecture très reconnue, construite avec un encodeur et décodeur, qui repose entièrement sur un mécanisme d'attention pour établir des liens globaux entre les entrées et les sorties. Le mécanisme d'attention est une technique qui modifie l'importance de certaines parties des données d'entrée afin de se concentrer sur les éléments les plus pertinents. Ce mécanisme applique des pondérations aux sorties. Les entrées originales, c'est-à-dire

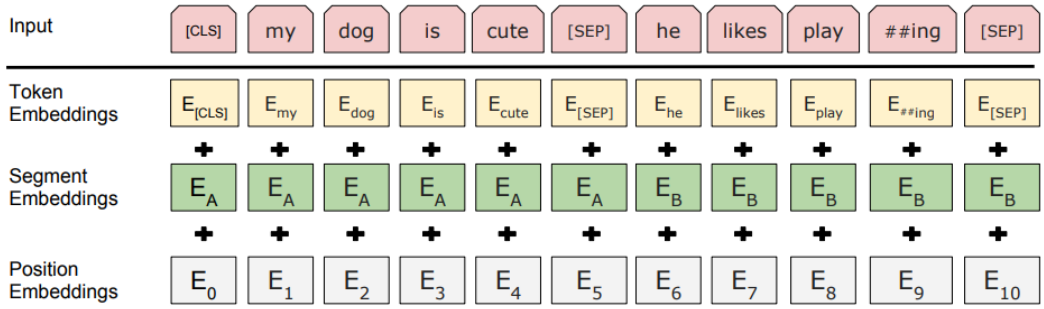

FIGURE 1 – Représentation d'entrée de BERT

les vecteurs représentant initialement les textes, passent à travers trois ensembles de matrices : les matrices de clés, de valeurs et de requêtes. Ensuite, la sortie est calculée comme une somme pondérée des valeurs, où le poids attribué à chaque valeur est calculé en fonction de la compatibilité entre la requête et la clé correspondante.

3.2.8 GPT

Le modèle GPT *(Generative Pre-trained Transformers)* est basé sur les transformers. Le modèle est pré-entraîné sur de grandes quantités de texte. Pendant cette phase, il apprend à prédire le mot suivant dans une séquence de mots donnée. Pour ce faire, le modèle analyse le contexte des mots précédents pour générer des prédictions pertinentes. Pour la variante GPT-2, les chercheurs (RADFORD et al., 2019) ont évalué ce modèle sur différentes tâches, dont la majorité nécessitait la génération de texte. GPT-2 a obtenu de meilleurs résultats pour ces tâches 7 fois sur 8. GPT-2 excelle dans la génération de textes.

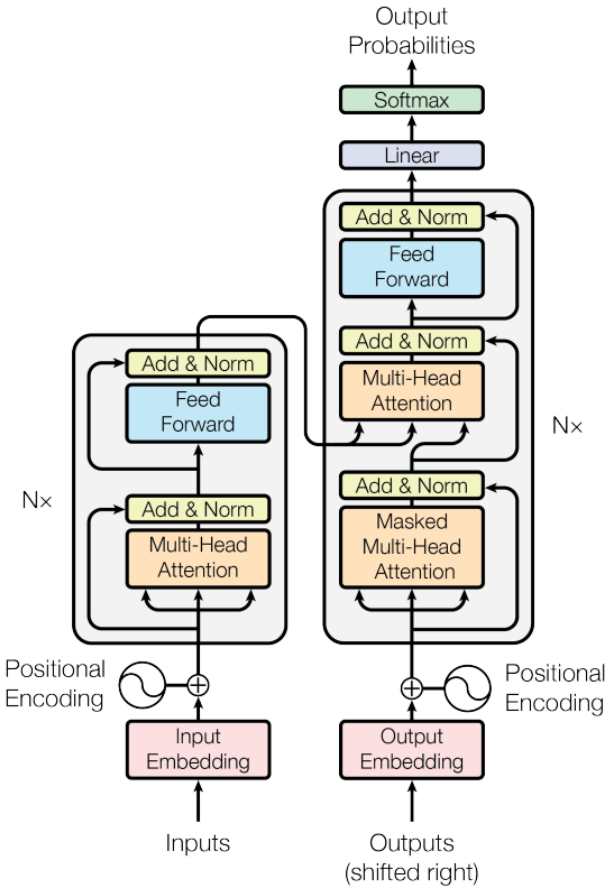

FIGURE 2 – La structure de transformer

3.3 Similarité textuelle

Il existe diverses méthodes pour mesurer la similarité entre des textes. La métrique de similarité cosinus (section 3.3.1) est la plus couramment utilisée et nous l'avons également adoptée. En plus de la similarité cosinus, il existe d'autres métriques telles que la distance euclidienne (section 3.3.2) et le score Jaccard (section 3.3.3). Toutes ces métriques requièrent des vecteurs en entrée. Par conséquent, avant d'appliquer ces métriques, il est nécessaire de représenter les textes sous forme de vecteurs.

3.3.1 Similarité cosinus

Cette métrique est originellement utilisée pour mesurer la similarité entre deux vecteurs, elle concentre sur l'angle entre deux vecteurs, plus précisément, les directions des deux vecteurs.

La définition de la formule :

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

Dans la formule, le $\|A\|$ est la longueur du vecteur A, le $\|B\|$ est la longueur du vecteur B. Le produit des longueurs des deux vecteurs est utilisé comme dénominateur, et le produit de ces vecteurs est utilisé comme numérateur.

Certaines études associent la similarité cosinus avec d'autres méthodes de traitements de textes ou modifient légèrement la formule pour améliorer la performance.

Basée sur la similarité cosinus, une étude (Gunawan, Sembiring et Budiman, 2018) a mesuré la pertinence du texte à partir de mots-clés calculs en utilisant TF-IDF, ainsi qu'à partir des poids de ces mots-clés. Rahutomo, Kitasuka et Aritsugi (2012) a également proposé une amélioration de la similarité cosinus entre deux vecteurs en faisant des opérations comme l'égalisation dimensionnelle, des traitements pour la relation hyperonyme-hyponyme entre vecteurs, et calibrer les valeurs dans les vecteurs par une formule.

3.3.2 Distance euclidienne

La distance euclidienne calcule la distance entre la fin de deux vecteurs. Sa formule est définie ainsi :

$$\|x - y\|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (2)$$

Cette méthode peut surpasser la similarité cosinus quand les vecteurs sont pointés vers la même direction, car la similarité cosinus ne permet pas de comparer les distances entre les vecteurs, mais la distance euclidienne est utilisée pour calculer la distance entre la fin de deux vecteurs.

Cette distance est fréquemment employée dans le domaine de l'apprentissage automatique. Par exemple, dans l'algorithme K-means, elle est utilisée pour effectuer le regroupement des données.

Gulati et Yadav (2019) ont adopté cette métrique pour annoter les images sur la page Web. Ils ont écrit un algorithme qui collecte les blocs de texte les plus proches d'un bloc d'image présent sur la page Web, et puis ils ont utilisé la distance euclidienne pour trouver les blocs les plus sémantiquement proches avec l'image.

3.3.3 Score Jaccard

Il s'agit d'un score qui sert à mesurer la similarité entre deux ensembles. Sa formule est :

$$J(A, B) = \frac{A \cap B}{A \cup B} \quad (3)$$

A, B sont deux ensembles et J représente la distance Jaccard. Il est défini comme la taille de l'intersection des deux ensembles divisée par la taille de l'union des deux

ensembles, plus précisément, ce score juge la similarité par le nombre d'éléments en commun des deux ensembles.

Dans une recherche sur le système de recommandation, une équipe (Bag, Kumar et M. K. Tiwari, 2019) a développé deux nouveaux modèles de similarité basés sur la similarité Jaccard, en considérant tous les vecteurs de notation des utilisateurs pour classer et générer des recommandations dans un temps de calcul réduit. Selon eux, la similarité Jaccard fonctionne plus précisément et efficacement pour générer des recommandations de qualité que les autres modèles de similarité.

En revanche, le Jaccard-score a quand même des limites. Par exemple, la relation sémantique et syntaxique entre les mots ne sera pas capturée dans ce cas, de plus, la taille des textes peut avoir un grand impact sur la similarité Jaccard.

3.4 Métriques

Les métriques sont toujours très importantes pour évaluer l'efficacité de nos approches, avec les métriques, nous pouvons arriver à choisir la meilleure méthode ou à faire la comparaison entre différentes méthodes selon différents critères. Dans cette section, nous représentons les métriques, en commençant par la matrice de confusion dans la Section 3.4.1, ensuite, nous présentons l'exactitude dans la Section 3.4.2, la précision et le rappel dans la Section 3.4.3, puis, f-score dans la Section 3.4.4 et enfin la courbe ROC dans la Section 3.4.5.

3.4.1 Matrice de confusion

Il s'agit d'une mesure de performance qui sert à évaluer les méthodes de classification dans l'apprentissage automatique. Il faut avoir au minimum deux classes pour pouvoir utiliser cette mesure. Elle est utile pour mesurer le rappel, la précision, la spécificité, l'exactitude d'un modèle.

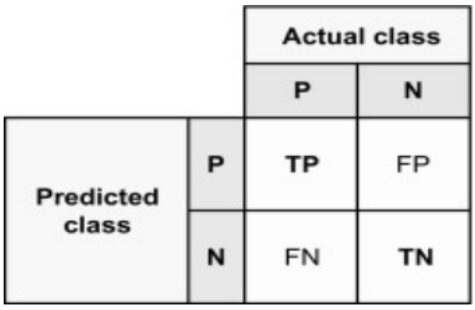

FIGURE 3 – Matrice de confusion (A. TIWARI, 2022)

Dans la figure 3, il y a généralement 4 valeurs cibles dans la matrice, Vrai positif (TP), Faux positif (FP), Vrai négatif (VN) et Faux négatif (FN).

Vrai positif (TP) : l'observation est prédite comme positive et est en réalité positive. Faux positif (FP) : L'observation est prédite positive et est en fait négative. Vrai négatif (TN) : L'observation est prédite négative et est en réalité négative. Faux négatif (FN) : L'observation est prédite négative et est en fait positive.

3.4.2 Exactitude

L'exactitude est calculée avec la formule suivante :

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

Cette métrique peut montrer le nombre d'instances de données correctement classées sur le nombre total d'instances de données. Mais quand notre jeu de données n'est fortement pas équilibré, cette métrique devient invalide, car nous ne pouvons pas connaître le nombre de TN (Vrai négatif) et TP (Vrai positif) séparément.

3.4.3 Précision et rappel

La formule de la précision *(precision)* est définie ainsi :

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

La précision montre le nombre d'éléments positifs bien classés sur le nombre total d'éléments classés comme positifs par le modèle.

La formule de rappel *(recall)* est définie comme :

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

Le rappel montre le nombre d'éléments positifs bien classés sur le nombre total d'éléments positifs. Un modèle sera considéré efficace s'il a une haute précision et un haut rappel en même temps.

3.4.4 F1-score

Pour résoudre le problème de déséquilibre dans un jeu de données, une autre métrique couramment utilisée est le score F1, qui est combiné avec la précision et le rappel. Le score F1 est calculé en utilisant une moyenne harmonique pondérée de la précision et du rappel. Sa formule est définie comme suit :

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

Ce qui peut être également écrit sous cette forme :

$$F1 - score = \frac{TP}{TP + \frac{FN+FP}{2}} \tag{8}$$

Cette métrique sert à classifier les instances positives, il aide à comprendre le compromis entre l'exactitude et la couverture de données. (Kulkarni, Chong et Batarseh, 2020)

3.4.5 Courbe ROC

La courbe ROC (courbe caractéristique de fonctionnement du récepteur) est un graphique qui représente la performance d'un modèle de classification binaire à différents seuils de discrimination (voir 4). Elle illustre la relation entre le taux de vrais positifs (Sensibilité) et le taux de faux positifs (1 - Spécificité) à mesure que

le seuil de décision du modèle varie. L'axe x représente FPR (False Positive Rate), et l'axe y représente TPR (True Positive Rate), ce qui implique le rappel positif. Lorsque le modèle est capable de faire une distinction nette entre les classes et de minimiser les erreurs de classification, la courbe ROC montera rapidement vers le coin supérieur gauche du graphique, indiquant à la fois une sensibilité élevée et une spécificité élevée.

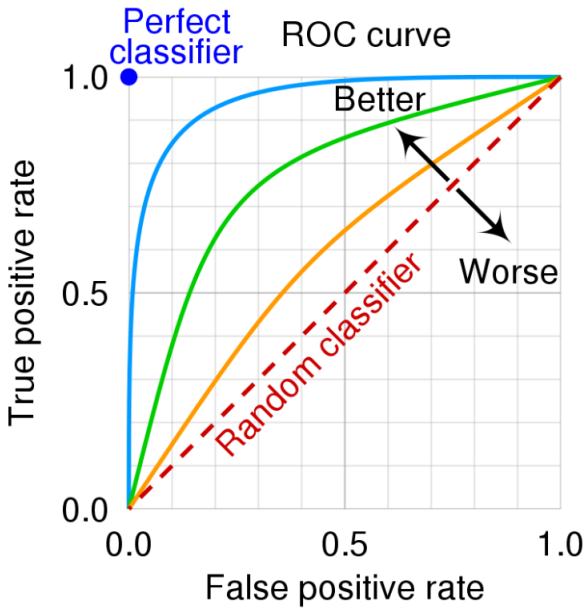

FIGURE 4 – La courbe roc Par cmglee, MartinThoma, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

3.5 Conclusion

Au début de la partie sur l'état de l'art, nous avons exposé les effets des citations et présenté des études statistiques sur la pertinence des citations dans le domaine de la recherche scientifique. De plus, nous avons également poussé notre exploration plus loin en examinant les recherches sur le contexte des citations qui font usage

de l'apprentissage automatique pour nous fournir des idées inspirantes. Ces études ont montré que les modèles de langues sont capables de classifier les textes selon différents critères après les entraînements spécifiques.

Nous avons également exposé diverses méthodes pour obtenir des représentations vectorielles de textes. Par la suite, nous avons introduit différentes métriques pour évaluer la similarité textuelle, ainsi que des métriques permettant d'évaluer différentes techniques d'apprentissage automatique.

Dans les deux parties suivantes, nous introduisons notre corpus. Puis, nous exposons nos méthodes pour évaluer l'exactitude des citations en utilisant l'apprentissage automatique. Après, nous présentons les différentes méthodes pour analyser l'efficacité de ces méthodes.

Partie 3

-

Recueil de données

4 Recueil de données

Cette partie consiste à décrire notre jeu de données principal : Final-test-set dans la Section 4.1, les données pour fine-tuning : MRPC dans la Section 4.2. Nous avons également construit deux autres jeux de données, les détails sont dans la partie annexe.

4.1 Final-test-set

Ce jeu de données est construit afin d'évaluer l'efficacité de nos méthodes pour classifier les citations en tant que fiables ou erronés, ainsi que pour déterminer s'ils sont dans le domaine ou non. Nous avons divisé nos citations en trois catégories : Fiables et dans le domaine, Erronés et hors domaine, et Erronés et dans le domaine (voir exemple dans tableau 5).

— Une citation fiable et dans le domaine fait référence à une source dans le même sujet de recherche et reflète avec précision l'intention originale de l'auteur de l'article cité.

— Les citations erronées et hors domaine représentent des cas où les références citées proviennent d'un sujet de recherche totalement non lié.

— Les citations erronées et dans le domaine citent une référence dans le même sujet, mais ils ne captent pas avec exactitude l'objectif original de l'auteur de l'article cité.

Les contextes de ces citations ont été extraits manuellement des articles citant les références suivantes :Payton et al. (2017), Karthik et al. (2014), Vickers (2017), Portet et al. (2013), Peinelt, Nguyen et Liakata (2020) Carlsen et Glenton (2011). Les résumés ont ensuite été extraits de ces articles de référence.

Dans le Tableau 4, nous avons trouvé au total 64 contextes de citation, parmi lesquels 31 sont annotés comme étant des citations fiables, 12 sont annotés comme erronés et hors domaine, et 21 sont erronés mais dans le domaine.

			Citation erronée	
	Total	Citation fiable	Dans domaine	Hors domaine
Contexte de citation	64	31	21	12

TABLE 4 – Jeu de données Final-test-set

Catégorie	Contexte de citation	Contexte référé
Fiables et dans le domaine	The purpose of the study was to assess parental thoughts on what high school administrators should be doing to reduce the risk of firearm violence in schools	The purpose of this study was to examine what parents thought schools should be doing to reduce the risk of firearm violence in schools.
Erronés et hors domaine	Vacuum assisted closure (VAC) (Kinche, Concepts, Inc, San Antonio, TX, USA) treatment provides a good environment that allows for both open and closed treatment, better wound healing procedures under moist, hygienic, sterile conditions.7	(Pas de contexte référé trouvé, l'article cité est sur violence par arme à feu à l'école aux États-Unis)
Erronés et dans le domaine	Carlsen and Glenton (2011) recommended at least two focus group sessions but due to limited availability of researcher and participants ; only one focus group session occurred with follow-up discussions as necessary to achieve data and theme saturation. Erronés et dans le domaine	The number of focus groups conducted varied greatly (mean 8.4, median 5, range 1 to 96) (Le papier référencé n'a pas recommendé nombre de focus group)

TABLE 5 – Exemples de notre 3 catégories de contextes de citation

4.2 MRPC

Le Corpus de Paraphrases de Microsoft Research (MRPC) est un ensemble de données composé de 5801 paires de phrases extraites d'articles de journaux. Ce corpus a été introduit par W. Dolan et al. (2004). Le but de ce corpus est de fournir des données pour l'évaluation et le développement de modèles de traitement automatique du langage naturel qui visent à détecter les paraphrases. Chaque paire de phrases est étiquetée par des annotateurs humains pour indiquer si elles sont des paraphrases ou non. Les données sont réparties en un sous-ensemble d'apprentissage (composé de 4076 paires de phrases, dont 2753 sont des paraphrases) et un sous-ensemble de test (constitué de 1725 paires, dont 1147 sont des paraphrases).

Exemple des paires paraphrase :

label	sentence1	sentence2
0 (not_equivalent)	The identical rovers will act as robotic geologists , searching for evidence of past water .	The rovers act as robotic geologists , moving on six wheels .
0 (not_equivalent)	Less than 20 percent of Boise 's sales would come from making lumber and paper after the OfficeMax purchase is completed .	Less than 20 percent of Boise 's sales would come from making lumber and paper after the OfficeMax purchase is complete , assuming those businesses aren 't sold .
1 (equivalent)	Spider-Man snatched $ 114.7 million in its debut last year and went on to capture $ 403.7 million .	Spider-Man , rated PG-13 , snatched $ 114.7 million in its first weekend and went on to take in $ 403.7 million .
1 (equivalent)	The 2002 second quarter results don 't include figures from our friends at Compaq .	The year-ago numbers do not include figures from Compaq Computer .

Figure 5 – Exemple de paires paraphrase <https://www.tensorflow.org/datasets/catalog/glue

Partie 4

-

Méthodologies

5 Méthodologies

Dans cette partie, nous exposons diverses méthodes pour évaluer la similarité textuelle. En premier lieu, nous décrivons la constitution de notre corpus (voir la Section 5.1). Nous exposons également notre approche pour obtenir les plongements de mots dans cette même section. Ensuite, nous détaillons le calcul de la similarité cosinus (voir la Section 5.2) ainsi que notre méthode de classification (voir la Section 5.3). Enfin, nous exposons notre démarche pour déterminer les seuils (voir la Section 5.4).

5.1 Construction du corpus

Pour pouvoir faire des futures analyses plus facilement, nous avons défini deux configurations pour notre corpus, le premier, comporte le contexte de citation et l'abstract original du papier cité (voir la Section 5.1.1), et le deuxième comporte le contexte de citation et l'abstract coupés en phrases (voir la Section 5.1.2). Enfin, nous introduisons la méthode pour obtenir la représentation vectorielle (voir la Section 5.1.3).

Nous avons développé également un outil de terminal qui prend en entrée le corpus annoté et qui donne en sortie les deux configurations que nous avons définies selon les besoins.

5.1.1 configuration abstract entier

Dans cette configuration, nous avons inclus les colonnes essentielles et informatives. Les colonnes essentielles englobent le contexte de citation, l'abstract entier

du papier cité correspondant, l'étiquette de similarité, ainsi que le domaine annoté artificiellement. Par ailleurs, les colonnes informatives contiennent des données supplémentaires concernant le papier cité et le papier citant, telles que le nom de l'auteur et le DOI.

Selon cette configuration, nous avons mis en œuvre nos approches sur chaque contexte de citation et sur l'abstract du papier cité correspondant (voir figure 6).

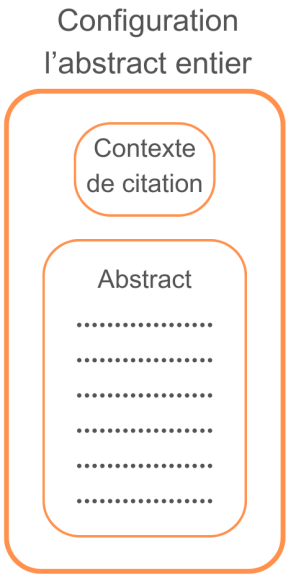

FIGURE 6 – configuration contexte de citation avec l'abstract entier

5.1.2 configuration abstract coupé en phrases

Au lieu de comparer le contexte de citation avec l'abstract entier dans son ensemble, cette configuration a été conçue en optant pour couper l'abstract en phrases distinctes. Cette approche vise à déterminer si de telles modifications apportent des améliorations à nos méthodes.

La décision de découper l'abstract en phrases découle de notre observation selon

laquelle BERT peut traiter un maximum de 512 tokens. Étant donné que certains abstracts sont particulièrement longs, ils ne peuvent pas être intégralement traités par BERT.

Dans ce scénario, le processus de prédiction de similarité se complexifie. Chaque contexte de citation est associé à une phrase entière extraite de l'abstract correspondant. Ainsi, pour un abstract entier, un ensemble de n pairs est généré, où n représente le nombre de phrases entières dans cet abstract.

Par la suite, pour prédire la corrélation entre le contexte de citation et l'abstract, il est nécessaire de développer une méthode permettant de généraliser les similarités établies entre le contexte et les phrases segmentées de l'abstract. Nous avons envisagé l'utilisation de la courbe ROC afin de déterminer des seuils pertinents pour nos diverses méthodes.

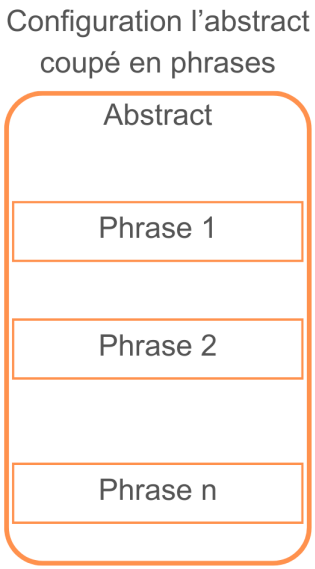

FIGURE 7 – configuration contexte de citation avec l'abstract coupé en phrases

5.1.3 Obtention de plongement de mots

Au lieu de fournir du texte brut directement aux modèles de langage, des étapes de prétraitement sont nécessaires pour adapter ces textes aux configurations requis par les modèles. Ces prétraitements débutent par la tokenisation des phrases selon divers critères.

En ce qui concerne BERT, il dispose de sa propre méthode de tokenisation, qui segmente les phrases en mots complets s'ils sont présents dans son vocabulaire. S'ils ne sont pas présents dans le vocabulaire, les mots sont coupés en sous-unités. La limite maximale de tokens est de 512. Le tokeniseur peut opérer sur une seule phrase ou simultanément sur une paire de phrases.

Une fois la tokenisation effectuée, des matrices contenant les informations originales des textes sont obtenues. Ces matrices incluent, entre autres, un vecteur d'identifiants de tokens *(token ids)*, qui représente les tokens en fonction du vocabulaire du modèle, un vecteur de masques d'attention *(attention mask)*, ainsi qu'un vecteur permettant de différencier les deux phrases dans une paire de phrases *(token type id)*.

Par la suite, ces matrices sont directement transmises au modèle, qui génère automatiquement des représentations vectorielles. Pour évaluer la similarité, l'encodage du "*last hidden state*" est sélectionné afin d'exprimer la sémantique d'une phrase. Cet encodage consiste en une représentation comprenant les états cachés finaux du réseau neuronal pour l'ensemble des tokens de la phrase.

5.2 Similarité cosinus

Comme expliqué dans le précédent chapitre, la similarité cosinus est originellement une mesure mathématique pour évaluer la similitude entre deux vecteurs. Nous avons utilisé cette métrique pour les vecteurs représentant les phrases générées par les modèles de langage.

Nous avons tenté de calculer les similarités dans la section "abstract" des papiers de référence. Pour ce faire, nous commençons par une brève introduction sur la structure générale des articles scientifiques dans la Section 5.2.1. Ensuite, nous expliquons le calcul de la similarité cosinus entre le contexte de citation et l'abstract du papier référencé dans la Section 5.2.2. Enfin, nous abordons le découpage en phrases de l'abstract du papier référencé dans la Section 5.2.3.

Nous fournissons également, en annexe, les détails du calcul de la similarité cosinus entre le contexte de citation et le contexte de référence dans la Section A.4.1, ainsi que l'abstract du papier citant et l'abstract du papier référencé dans la Section A.4.2.

5.2.1 Structure générale des papiers scientifiques

Dans les papiers scientifiques, la structure est globalement similaire. On y trouve souvent les sections suivantes : Titre et auteurs, Résumé (abstract), Introduction, Méthodes, Résultats, Discussion et Références.

La section "Titre et auteurs" comprend le titre du papier ainsi que les auteurs et leurs coordonnées. Dans la section "Résumé", l'auteur présente une synthèse de l'ensemble du papier, offrant ainsi une vue d'ensemble générale. La section "Introduction" expose le contexte de la recherche, ainsi que les travaux antérieurs liés au

sujet. Les détails sur la collecte de données et les méthodes expérimentales sont souvent présentés dans la section "Méthodes". Les résultats sont analysés et résumés dans la section "Résultats", tandis que la pertinence du travail est évaluée dans la section "Discussion". La dernière section contient généralement des informations sur les citations et références.

Bien qu'il puisse exister des références à des contextes en dehors de la section "Résumé" (*abstract*), notre intérêt principal réside dans cette section. Nous effectuons nos comparaisons en utilisant le contexte de citation et l'abstract des papiers référés.

5.2.2 Contexte de citation et l'abstract référencé

Initialement, nous avons obtenu des représentations vectorielles pour l'ensemble de l'abstract du papier cité ainsi que le contexte de citation dans le papier citant. Par la suite, nous avons calculé les similarités cosinus entre ces vecteurs. En utilisant un seuil, nous avons défini la catégorie des contextes de citation. La figure 8 illustre ce processus.

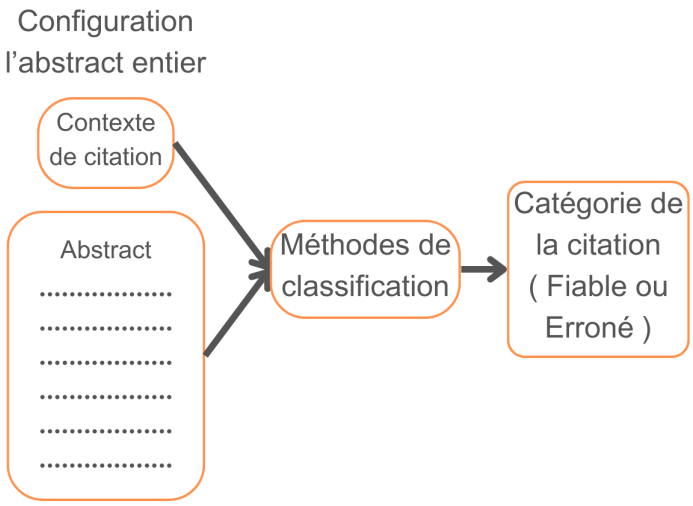

FIGURE 8 – Processus de comparaison entre le contexte de citation et l'abstract

5.2.3 Contexte de citation et l'abstract référencé coupé en phrases

D'abord, en appliquant les modèles de langue, nous obtenons les représentations vectorielles du contexte, puis, nous obtenons les vecteurs de chaque phrase extraite de l'abstract référencé. Nous calculons la similarité entre le contexte et chaque phrase, et puis, nous gardons la valeur maximale de la similarité pour représenter la similarité entre le contexte et l'abstract référencé. La figure 9 illustre ce processus.

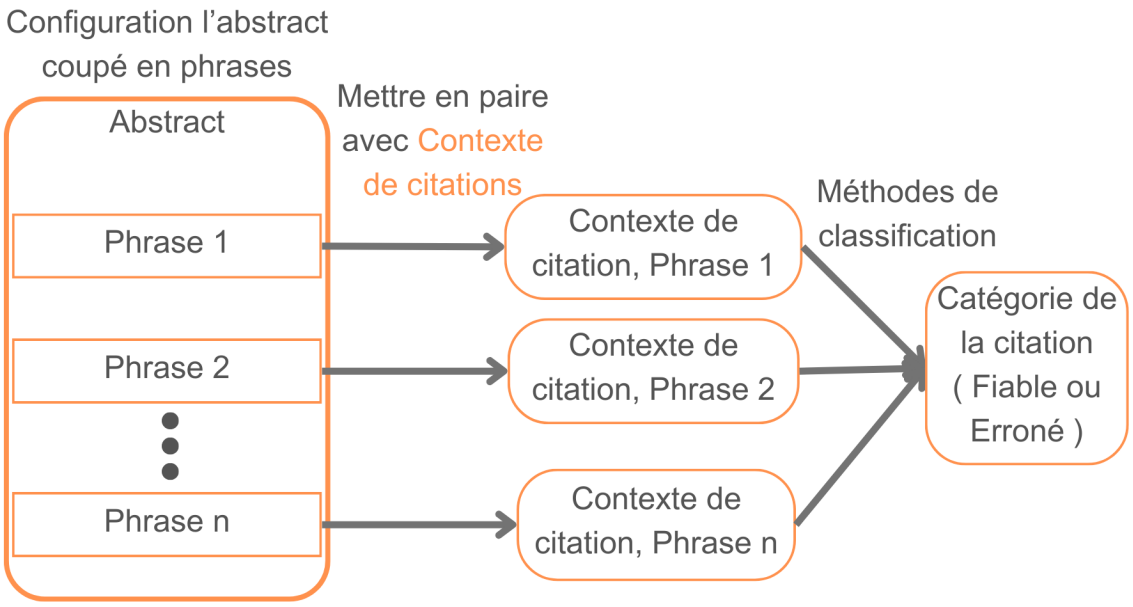

FIGURE 9 – Processus de comparaison entre le contexte de citation et l'abstract coupé en phrases

5.3 Classification avec la technique du fine-tuning

Nous avons essayé d'effectuer un fine tuning sur modèle de BERT en utilisant les paraphrases, avec un classificateur très simple qui applique d'abord un dropout, puis une fonction pour la classification linéaire, et à la fin une fonction d'activation Relu. Nous avons appliqué cette méthode sur nos deux configurations : le contexte de citation et l'abstract entier référencé, ainsi que le contexte de citation avec l'abstract référencé coupé en phrases. Nous introduisons d'abord notre structure détaillée du classificateur dans la Section 5.3.1, puis, la méthode applique sur la configuration : l'abstract entier dans la Section 5.3.2, ainsi que la configuration : l'abstract coupé en phrases dans la Section 5.3.3.

5.3.1 Structure du classificateur

La structure du classifieur de paraphrase :

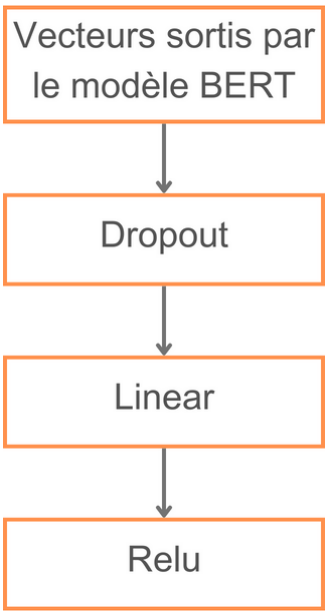

FIGURE 10 – Structure de notre classifieur de paraphrase

Comme indiqué dans le graphique, après avoir obtenu la représentation vectorielle de la paire de contextes de citation avec l'abstract, nous la faisons passer à travers notre structure de classification. Tout d'abord, nous appliquons une régularisation de type dropout pour éviter le surajustement. Ensuite, nous utilisons une fonction linéaire suivie d'une fonction d'activation ReLU pour la classification. Le vecteur résultant sera en une dimension avec deux valeurs qui représentent les deux classes : citation fiable ou erronée. Enfin, nous appliquons la fonction argmax pour déterminer la classe de la paire de contexte et abstract.

5.3.2 La configuration l'abstract entier

Dans ce scénario, nous comparons les contextes de citations dans les articles citants avec l'abstract entier de l'article cité en référence. Nous soumettons d'abord simultanément le contexte et l'abstract entier au modèle BERT. Ensuite, les vecteurs résultants traversent la structure de classification, et en fin de compte, nous utilisons la fonction argmax pour obtenir la prédiction de classe. Le processus est décrit dans la figure 8

5.3.3 La configuration l'abstract coupé en phrases

Pour prédire la similarité entre le contexte de citation et l'abstract entier référencé de manière plus détaillée, dans ce cas, nous associons le contexte de citation avec chaque phrase coupée de l'abstract référencé, et nous effectuons la prédiction de classe pour chaque paire individuellement. La figure 9 décrit ce processus

En fin de compte, nous évaluons la proportion de paires prédites comme appartenant à la classe fiable. Par exemple, si un abstract entier est divisé en 10 phrases, cela signifie que nous aurons 10 paires composées d'un contexte de citation et d'une phrase extraite de l'abstract. Pour chaque paire, nous effectuons une prédiction de catégorie (fiable ou erronée), puis nous examinons la part de paires fiables. Si, par exemple, 5 paires sur ces 10 sont classées comme fiables, la proportion de paires fiables serait de 50% pour ce contexte de citation et son abstract entier référencé.

Ensuite, nous fixons un seuil au-delà duquel le contexte de citation et l'abstract référencé seront considérés comme fiables. Dans le cas contraire, la citation sera considérée comme une citation erronée. Par exemple, si notre seuil est de 20%, cela signifie que s'il y a plus de 20% de paires considérées comme fiables, alors cette citation et son abstract sont considérés comme fiables.

5.4 Définition du seuil

Nous avons adopté 2 méthodes pour définir les seuils, une méthode manuelle (voir la Section 11) et une autre avec la courbe ROC (voir la Section 12). Dans cette Section, nous mettons 2 exemples pour expliquer le calcul du seuil pour la similarité cosinus et pour le classifieur de paraphrase.

5.4.1 Méthode Manuelle

En se basant sur les scores de similarité cosinus et les annotations humaines, il est possible de définir manuellement un seuil pour distinguer les contextes de citation fiables et erronées. Par exemple, pour définir le seuil de la configuration l'abstract entier (voir figure 6), dans la figure 11, les contextes de citation annotés comme fiables sont représentées par les points oranges avec l'étiquette 1, tandis que les points bleus représentent les citations erronées, avec l'étiquette 0.

FIGURE 11 – Le seuil défini manuellement

Théoriquement, les contextes de citations fiables devraient afficher une similarité plus élevée. En effet, les points oranges présentent une tendance à se situer à des valeurs plus élevées que les points bleus, ce qui concorde avec notre hypothèse. Dans ce graphique, il est possible d'identifier un seuil autour de 0.15 permettant de distinguer les citations fiables des citations erronées.

5.4.2 Méthode courbe ROC

Comme mentionné dans l'état de l'art, la courbe roc nous permet de définir un seuil d'une façon statistique pour notre classifieur de paraphrase. Elle présente la relation entre le taux de vrais positifs (TPR) et le taux de faux positifs (FPR) à mesure que le seuil de classification varie.

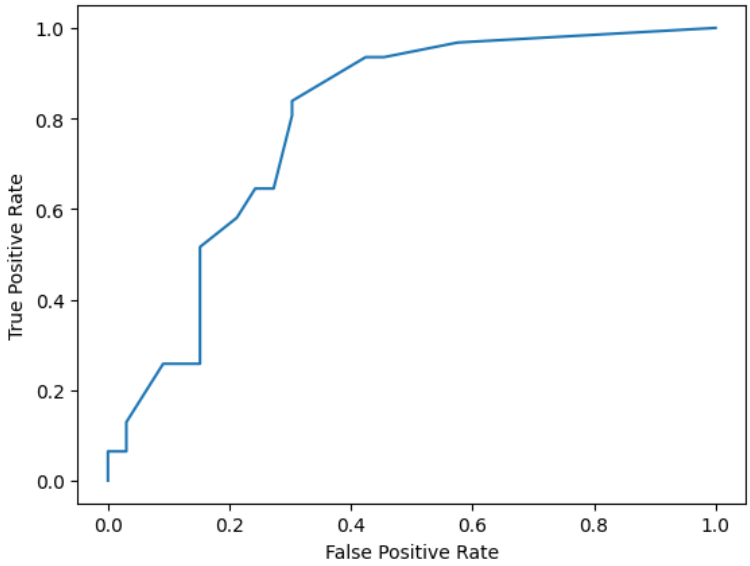

FIGURE 12 – La courbe roc du classifieur de paraphrase

Dans le graphique, l'axe des x représente mesure le pourcentage de faux positifs (éléments négatifs incorrectement classés comme positifs) parmi tous les éléments négatifs réels. L'axe y représente le pourcentage de vrais positifs (éléments positifs correctement classés) parmi tous les éléments positifs réels, donc plus la courbe se rapproche du coin supérieur gauche du graphique, meilleure est la performance de classification du modèle.

Nous avons utilisé la fonction roc intégrée dans scikit-learn pour calculer le meilleur seuil pour notre configuration l'abstract coupé en phrases (voir figure 7).

5.5 Conclusion

Dans cette partie, nous avons expliqué la construction de notre corpus, ainsi que notre deux configurations : Contexte de citation avec l'abstract entier (voir figure 6) et Contexte de citation avec l'abstract coupé en phrases (voir figure 7). Nous avons également détaillé l'application de nos deux méthodes :

— **Similarité Cosinus avec la configuration l'abstract entier du papier cité** : Dans un premier temps, nous avons obtenu des représentations vectorielles pour l'ensemble de l'abstract du papier cité ainsi que le contexte de citation dans le papier citant. Ensuite, nous avons calculé les similarités cosinus entre ces vecteurs.

— **Classisieur de paraphrase avec la configuration l'abstract entier du papier cité** : Nous avons fine-tuned un classifieur en utilisant les plongements BERT et le corpus MRPC (W. B. Dolan et Brockett, 2005). Ce classifieur a ensuite été appliqué pour catégoriser les paires de contextes de citation et d'abstracts provenant des papiers cités.

— **Similarité Cosinus avec la configuration l'abstract coupé en phrases du papier cité** : Tout d'abord, nous avons segmenté l'abstract du papier cité en phrases, puis obtenu des représentations vectorielles pour chacune de ces phrases ainsi que le contexte de citation dans le papier citant. Ensuite, nous avons calculé les similarités cosinus entre chaque phrase de l'abstract et le contexte de citation.

— **classifieur de paraphrase avec la configuration l'abstract coupé en phrases du papier cité** : Nous avons fine-tuned un classifieur en utilisant les plongements BERT et le corpus MSRP(W. B. Dolan et Brockett, 2005). Ce classifieur a ensuite été appliqué pour catégoriser les paires de contexte de citation avec chaque phrase extraite de l'abstract dans les papiers cités.

Les figures 8 et 9 décrit précisément ces deux processus. À la fin, nous avons également exposé nos approches pour déterminer les seuils dans les deux méthodes. Pour le classifieur de paraphrase, nous avons utilisé la courbe ROC pour définir le seuil, tandis que pour la méthode de similarité cosinus, nous avons ajusté manuellement le seuil.

Partie 5

-

Résultats et discussions

6 Résultats et discussions

Dans cette partie, nous procédons à une analyse approfondie de nos différentes méthodes sous différentes perspectives, notamment en évaluant leur précision dans la distinction des citations fiables et leur capacité à identifier les citations hors domaine. Ensuite, nous abordons les limites de notre approche et explorons les potentielles améliorations pour aller plus loin dans nos futures recherches.

6.1 Résultats

Nous avons évalué nos méthodes de classification en deux aspects, la compétence sur la classification de citations fiables et erronées, ainsi que la compétence pour distinguer ces citations hors domaine et dans-le-domaine. Le tableau 6 présente les précisions de nos différentes méthodes. Il semble que la méthode de classification avec la configuration l'abstract coupé en phrases du papier cité surpasse les autres. Nos deux méthodes donnent de meilleurs résultats avec l'abstract coupé par rapport à l'abstract entier.

Configuration	Méthode	Exactitude totale	Citation fiable	Citation erronée	
				Dans domaine	Hors domaine
Abstract entier	Similarité cosinus	61%	38.7%	57.6%	24.2%
	Classifieur de paraphrases	61%	71%	18.2%	33.3%
Abstract coupé	Similarité cosinus	67.1%	71%	27.3%	36.4%
	Classifieur de paraphrases	75%	80.6%	36.4%	33.3%

TABLE 6 – Exactitude de prédictions

6.1.1 Méthode de classifieur de paraphrases

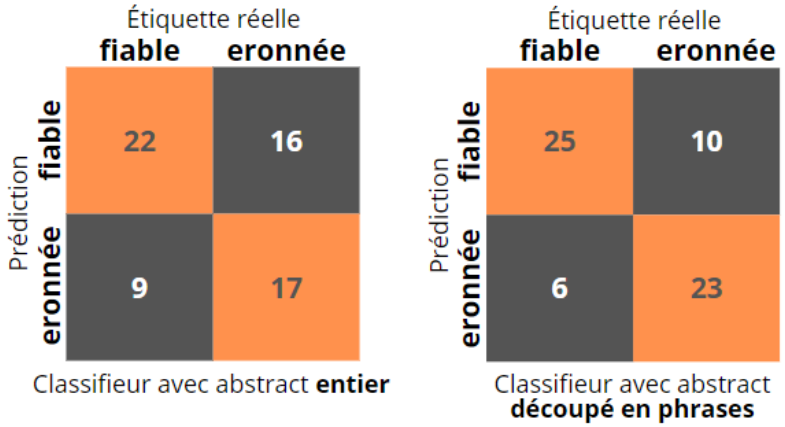

FIGURE 13 – Matrices de confusion de classifieur de paraphrases dans les deux configurations : abstract entier et abstract coupé

À partir de la figure 13, il est évident que pour la configuration basée sur l'abstract entier, notre méthode de classification a correctement prédit 39 citations parmi 64 obtenant ainsi 61% d'exactitude. Cependant, elle n'arrive pas à identifier correctement les citation erronées atteignant seulement une exactitude de 52%. En ce qui concerne la configuration basée sur l'abstract coupé, nous avons observé une amélioration des performances de prédiction avec 75% d'exactitude (48 correctement prédits parmi 64). Nous notons également une amélioration dans l'identification des citation erronées atteignant désormais une exactitude 70%.

6.1.2 Méthode de métrique similarité cosinus

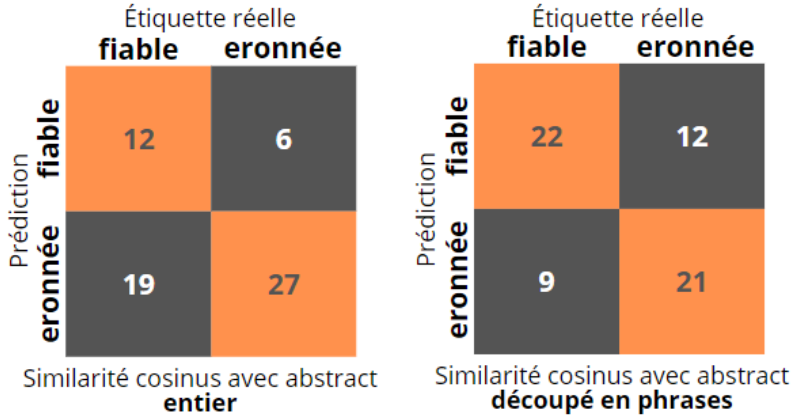

FIGURE 14 – Méthode similarité cosinus sur la configuration abstract entier (voir figure 6) et abstract coupé (voir figure 7)

Dans la figure 14, la méthode de similarité cosinus pour l'ensemble de l'abstract a également prédit correctement 39 contextes de citation sur 64 (61%). Cependant, par rapport à la méthode du classifieur, cette méthode est meilleure pour trouver les contextes de citation erronés, avec une précision de 82% (27 sur 33). Néanmoins, sa capacité à trouver des contextes de citation fiables est très limitée, avec seulement 12 sur 31 (38,7%) correctement identifiés. En utilisant l'abstract coupé, elle a correctement prédit 43 sur 64 contextes de citation (67,1%), avec une amélioration significative dans la prédiction des contextes de citation fiables et une diminution dans la prédiction des contextes de citation erronés. Elle a différencié 22 sur 31 (71%) des contextes fiables et 21 sur 33 (64%) des contextes erronés.

Nos résultats indiquent que les abstracts coupés améliorent à la fois les méthodes du classifieur et de similarité cosinus, notamment dans l'identification des contextes erronés et hors domaine. Bien que travailler avec l'ensemble de l'abstract donne des précisions similaires pour les deux méthodes, le classifieur éprouve des difficultés avec les citations erronées, tandis que la méthode de similarité cosinus peine à identifier

les contextes de citation fiables.

6.2 Discussion

Dans cette section, nous discutons les limites et proposons quelques pistes d'amélioration.

6.2.1 Les limites

Pendant ce stage, nous avons trouvé les limites de modèles de BERT, une des limites les plus graves dans notre cas, c'est que le BERT n'arrive pas à comprendre les nombres, car pendant la tokenisation, il tokenise les chiffres séparément, par exemple, pour un chiffre '24.9%', c'est tokenisé en '2', '4', '.', '9', '%'.

Cette façon de tokenization casse totalement le sens mathématique des chiffres, alors que les chiffres sont très souvent utilisés dans le cadre de citation, donc il est important de trouver une solution pour cette limite dans la recherche suivante. Pour l'instant, nous avons vu une étude (Wallace et al., 2019) qui entraîne les modèles pour qu'ils puissent comprendre les nombres.

BERT n'est pas faible à connaître les termes de négation. Selon une étude de Ettinger (2020), BERT montre une nette insensibilité aux impacts contextuels de la négation.

Il semble que si le modèle ne parvient pas à saisir la négation dans une phrase, cela signifie que s'il y a inversion de sens par l'auteur d'une citation par rapport au papier référencé, le modèle ne sera pas en mesure de détecter cette erreur. À l'heure actuelle, notre direction de recherche consiste à envisager un affinement spécifique

aux négations dans nos futures travails.

À part de BERT, il y a aussi d'autres limites. Tout d'abord, nous nous sommes principalement concentrés sur le résumé des articles cités, en négligeant d'autres sections importantes telles que les conclusions, les résultats et l'ensemble de l'article lui-même. Deuxièmement, en raison de contraintes de temps, nous n'avons pas pu réaliser des comparaisons approfondies entre BERT et divers modèles de langue. Troisièmement, notre classifieur a été affiné en utilisant des paraphrases provenant de journaux. Nous prévoyons d'obtenir des résultats améliorés en utilisant des données issues d'articles scientifiques pour l'affinage.

6.2.2 Améliorations envisageables

Après avoir réalisé toutes les expérimentations, nous pouvons en conclure théoriquement que pour la méthode de similarité cosinus et la méthode du classifieur, la configuration de l'abstract coupé en phrases peut nous donner de meilleurs résultats de classification, cependant, il nous reste encore pas mal de recherches à approfondir.

La configuration de contexte de citation avec contexte référé peut théoriquement surpasser notre résultat, il est nécessaire de grandir notre jeu de données de test et d'aller plus loin avec cette configuration.

En plus de la combinaison, la définition du seuil est aussi très importante. Nous pouvons encore se poser des questions sur les seuils, par exemple, est-ce que les seuils doivent être différents pour trouver les citations dans-le-domaine et hors-domaine ? Est-ce que la taille du texte a un impact sur les seuils aussi ? En plus de la courbe roc, est-ce qu'il existe encore d'autre façon de définir un bon seuil ?

Comme mentionné précédemment, il est également essentiel d'améliorer les modèles de langage tels que BERT en utilisant des techniques de fine-tuning et de

mener des comparaisons approfondies entre BERT et différentes autres modèles de langage.

Partie 6

-

Conclusion et perspectives

7 Conclusion et perspectives

Dans ce mémoire, nous nous intéressons à évaluer les citations dans les articles scientifiques. Ceci revient à mesurer la pertinence des citations dans leurs contextes.

Les citations sont utilisées dans les articles scientifiques pour différents objectifs entre autre le référencement des publications précédentes permettant ainsi de suivre les avancées scientifiques et de repérer les publications importantes dans un domaine scientifique particulier.
Des études sur l'exactitude de citations ont montré un taux d'erreur entre 25% et 54% dans différentes disciplines scientifiques. Ces erreurs modifient le sens original du papier cité. Nous distinguons les erreurs mineures et les erreurs majeures.
L'objectif de mon stage est l'évaluation du contexte de citation dans le papier citant et le contenu du papier cité. Autrement dit, ma mission consiste à étudier la corrélation entre la citation et son contexte dans le papier citant afin d'identifier les citations erronées.

Dans le papier citant, une citation dans un contexte donné fait référence au papier cité dans sa globalité. La citation ne précise pas la séquence textuelle du papier cité à laquelle les auteurs du papier citant font exactement référence. Cette séquence textuelle de référence peut être localiser dans une des parties du papier cité : abstract, introduction, méthodes, résultats, conclusion, *etc.* Dans mon stage, nous nous sommes concentrés à mesurer la corrélation entre le contexte de citation dans le papier citant et l'abstract du papier cité. Dans cette perspective, nous avons proposé deux configurations différentes. La première consiste à mesurer la corrélation entre les deux séquences textuelles : contexte de citation et abstract. La deuxième configuration permet de mesurer la corrélation entre le contexte de citation et l'abstract coupé en phrases. La deuxième configuration est motivée par la différence en nombre de phrases entre le contexte de citation dans le papier citant (qui est souvent

représenté par une seule phrase) et l'abstract du papier cité composé de plusieurs phrases (au moins 2 phrases).

Pour mesurer la corrélation entre le contexte de citation et l'abstract du papier cité (quelle que soit la configuration), nous avons proposé deux méthodes. La première se base sur la similarité cosinus et la deuxième se fonde sur un classifieur de paraphrases. La mesure de corrélation se base sur les représentations vectorielles des séquences textuelles. Nous avons utilisé la représentation vectorielle discrète de type *BERT* permettant de prendre en considération le contexte de la séquence textuelle.

Afin d'évaluer nos deux méthodes, nous avons mis en place un corpus composé de 64 contextes de citations dont 33 erronées. Les résultats ont montré que le classifieur de paraphrases est plus performant que la similarité cosinus. Il atteint 75% d'exactitude.

Comme perspectives, nous envisageons tester d'autres représentations vectorielles. Dans ce travail, nous avons utilisé le modèle BERT. Ce dernier ne représente pas convenablement la négation. Étant conscient de cette limite et pour ne pas biaiser les performances obtenues, nous avons convenablement choisi les exemples composant notre corpus de sorte à ce que les contextes de citations ne contiennent pas de termes de négation. Nous comptons tester le modèle *NegBERT* (Khandelwal et Sawant, 2020) permettant de considérer la négation dans les représentations vectorielles.

Une piste d'amélioration consiste à élargir notre corpus. Nous disposons d'autres exemples que nous les intégrerons à notre jeu de données actuel.

Une autre perspective consiste à implémenter des classifieurs sophestiqués. Étant performants pour plusieurs tâches TAL, les réseaux de neurones profonds pourront donner des bonnes performances pour notre tâche de mesure de pertinence de citations dans leurs contextes.

Au lieu de mesurer la pertinence de citations en se basant uniquement sur l'abs-

tract, nous envisageons aller plus loin et mesurer la corrélation entre le contexte de citations et les différentes parties des papiers cités. Une recherche, dans le papier cité, de la séquence textuelle fortement corrélée avec le contexte de citation dans le papier citant peut être menée afin d'évaluer les citations dans leurs contextes.

Références

Andrade, Chittaranjan (2011). "How to write a good abstract for a scientific paper or conference presentation". In : *Indian journal of psychiatry* 53.2, p. 172.

Armstrong, Michael F et al. (2018). "Reference errors in otolaryngology–head and neck surgery literature". In : *Otolaryngology–Head and Neck Surgery* 159.2, p. 249-253.

Bag, Sujoy, Sri Krishna Kumar et Manoj Kumar Tiwari (2019). "An efficient recommendation generation using relevant Jaccard similarity". In : *Information Sciences* 483, p. 53-64.

Bird, Steven et al. (jan. 2008). "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics." In.

Carlsen, Benedicte et Claire Glenton (2011). "What about N? A methodological study of sample-size reporting in focus group studies". In : *BMC medical research methodology* 11.1, p. 1-10.

— (2019). "When "Normal" Becomes Normative: A Case Study of Researchers' Quotation Errors When Referring to a Focus Group Sample Size Study". In : *International Journal of Qualitative Methods* 18, p. 1609406919841251. doi : 10.1177/1609406919841251.

De Lacey, Gerald, Carol Record et Jenny Wade (1985). "How accurate are quotations and references in medical journals?" In : *Br Med J (Clin Res Ed)* 291.6499, p. 884-886.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In : *arXiv preprint arXiv:1810.04805*.

Dolan, William et al. (août 2004). "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources". In : url : https://www.microsoft.com/en-us/research/publication/unsupervised-

construction-of-large-paraphrase-corpora-exploiting-massively-parallel-news-sources/.

Dolan, William B. et Chris Brockett (2005). *Automatically Constructing a Corpus of Sentential Paraphrases*. URL : https://aclanthology.org/I05-5002.

Ettinger, Allyson (2020). "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models". In : *Transactions of the Association for Computational Linguistics* 8, p. 34-48.

Gulati, Payal et Manisha Yadav (2019). *A novel approach for extracting pertinent keywords for web image annotation using semantic distance and euclidean distance*, p. 173-183.

Gunawan, Dani, C Sembiring et Mohammad Budiman (mars 2018). "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents". In : *Journal of Physics: Conference Series* 978, p. 012120. DOI : 10.1088/1742-6596/978/1/012120.

Horbach, Serge, Kaare Aagaard et Jesper W. Schneider (fév. 2021). *Meta-Research: How problematic citing practices distort science*. MetaArXiv aqyhg. Center for Open Science. DOI : 10.31219/osf.io/aqyhg. URL : https://ideas.repec.org/p/osf/metaar/aqyhg.html.

Jergas, Hannah et Christopher Baethge (2015). "Quotation accuracy in medical journal articles—a systematic review and meta-analysis". In : *PeerJ* 3, e1364.

John, Leslie K, George Loewenstein et Drazen Prelec (2012). *Measuring the prevalence of questionable research practices with incentives for truth telling*. T. 23. 5. Sage Publications Sage CA: Los Angeles, CA, p. 524-532.

Joulin, Armand et al. (2016). "Bag of tricks for efficient text classification". In : *arXiv preprint arXiv:1607.01759*.

Jurgens, David et al. (juill. 2018). "Measuring the Evolution of a Scientific Field through Citation Frames". In : *Transactions of the Association for Computational*

Linguistics 6, p. 391-406. ISSN : 2307-387X. DOI : 10.1162/tacl_a_00028. URL : https://doi.org/10.1162/tacl%5C_a%5C_00028.

KARTHIK, L et al. (2014). "Protease inhibitors from marine actinobacteria as a potential source for antimalarial compound". In : *PloS one* 9.3, e90972.

KHANDELWAL, Aditya et Suraj SAWANT (mai 2020). "NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution". English. In : *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France : European Language Resources Association, p. 5739-5748. ISBN : 979-10-95546-34-4. URL : https://aclanthology.org/2020.lrec-1.704.

KRISTOF, Cindy (oct. 1997). "Accuracy of Reference Citations in Five Entomology Journals". In : *American Entomologist* 43.4, p. 246-251. ISSN : 1046-2821. DOI : 10.1093/ae/43.4.246.

KULKARNI, Ajay, Deri CHONG et Feras A BATARSEH (2020). "Foundations of data imbalance and solutions for a data democracy". In : *Data democracy*. Elsevier, p. 83-106.

LIU, Haixia (2017). "Sentiment Analysis of Citations Using Word2vec". In : *CoRR* abs/1704.00177. arXiv : 1704.00177. URL : http://arxiv.org/abs/1704.00177.

MIKOLOV, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In : *arXiv preprint arXiv:1301.3781*.

PAYTON, Erica et al. (2017). "Parents' expectations of high schools in firearm violence prevention". In : *Journal of community health* 42, p. 1118-1126.

PEINELT, Nicole, Dong NGUYEN et Maria LIAKATA (juill. 2020). *tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection*. Online : Association for Computational Linguistics, p. 7047-7055. DOI : 10.18653/v1/2020.acl-main.630. URL : https://aclanthology.org/2020.acl-main.630.

PENNINGTON, Jeffrey, Richard SOCHER et Christopher D MANNING (2014). *Glove: Global vectors for word representation*, p. 1532-1543.

PORTET, François et al. (2013). "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects". In : *Personal and Ubiquitous Computing* 17, p. 127-144.

RADFORD, Alec et al. (2019). "Language models are unsupervised multitask learners". In : *OpenAI blog* 1.8, p. 9.

RAHUTOMO, Faisal, Teruaki KITASUKA et Masayoshi ARITSUGI (2012). *Semantic cosine similarity*. T. 4. 1, p. 1.

RAMOS, Juan (jan. 2003). "Using TF-IDF to determine word relevance in document queries". In.

TE, Sonita et al. (oct. 2022). *Citation Context Classification: Critical vs Non-critical*. Gyeongju, Republic of Korea : Association for Computational Linguistics, p. 49-53. URL : https://aclanthology.org/2022.sdp-1.6.

TIWARI, Ashish (2022). "Chapter 2 - Supervised learning: From theory to applications". In : *Artificial Intelligence and Machine Learning for EDGE Computing*. Sous la dir. de Rajiv PANDEY et al. Academic Press, p. 23-32. ISBN : 978-0-12-824054-0. DOI : https://doi.org/10.1016/B978-0-12-824054-0.00026-5. URL : https://www.sciencedirect.com/science/article/pii/B9780128240540000265.

TURKI, Turki et Sanjiban Sekhar ROY (2022). "Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer". In : *Applied Sciences* 12.13. ISSN : 2076-3417. DOI : 10.3390/app12136611. URL : https://www.mdpi.com/2076-3417/12/13/6611.

VASWANI, Ashish et al. (2017). "Attention is all you need". In : *Advances in neural information processing systems* 30.

VICKERS, Neil J. (2017). "Animal Communication: When I'm Calling You, Will You Answer Too?" In : *Current Biology* 27.14, R713-R715. DOI : https://doi.org/10.1016/j.cub.2017.05.064.

WALLACE, Eric et al. (2019). "Do NLP models know numbers? probing numeracy in embeddings". In : *arXiv preprint arXiv:1909.07940*.

8 Table des illustrations

Table des figures

Liste des tableaux

Annexes

Appendices

A Annexe

A.1 Jeu de données Good-in-bad-out

Objectif :

Ce jeu de données est construit pour nous permettre de choisir un format suffisant pour continuer les évaluations avec plus de données dans-le-domaines.

Quantité :

Il y a au total 10 citations, dont 6 sont bonne et dans-le-domaine, 4 sont mauvaises et hors-domaine.

Total	Bonne et dans-le-domaine	Mauvaise et hors-domaine
10	6	4

Exemples :

Bonne ou Mauvaise	Contexte de citation	Contexte référé
Bonne	The purpose of the study was to assess parental thoughts on what high school administrators should be doing to reduce the risk of firearm violence in schools	The purpose of this study was to examine what parents thought schools should be doing to reduce the risk of firearm violence in schools.
Mauvaise	Vacuum assisted closure (VAC) (Kinche, Concepts, Inc, San Antonio, TX, USA) treatment provides a good environment that allows for both open and closed treatment, better wound healing procedures under moist, hygienic, sterile conditions.7	0 (Pas de contexte référé trouvé, l'article cité est sur violence par arme à feu à l'école aux États-Unis)

Comment récupéré :

Les contextes et les abstracts de citations ont été trouvées manuellement dans les articles qui citent les papiers suivants : [Payton et al 2017]Payton et al., 2017, [Karthik et al 2014]Karthik et al., 2014. Les contextes et les abstracts sont extraits manuellement dans ces deux papiers référés.

Mais pour les citations hor-domaine, il n'est pas possible de retrouver le contexte

référé dans le papier cité, donc on a mis un '0' en tanq que contexte pour les mauvaises citations.

Abs-abs

A.2 Jeu de données Abs-abs

Objectif :

Ce jeu de données est construit pour observer la similarité cosinus sur les abstracts des papiers citants et papiers référés.

Quantité :

Il y a au total 726 citations, Nombre de différence de tokens entre l'abstract du papier citant et l'abstract référé varie de 0 tokens à 473 tokens, la cosinus similarité varie de 0.07 à 0.29.

Total	Différence de nombre de tokens	Similarité cosinus
726	0 - 473	0.07 - 0.29

Exemples :

abstract	abstract référé	Difference nombre tokens	Similarité cosinus
Oligosaccharides play a central role in plants and are involved in basal metabolism, and many other functions at cellular levels. ... will offer deeper insights into the spatio-temporal study of oligosaccharides in plants.	Male moths compete to arrive first at a female releasing pheromone. A new study reveals that additional pheromone cues released only by younger females may prompt males to avoid them in favor of older but more fecund females.	202	0.12542664
... Recently, school firearm violence and school shootings have received increasing attention from school personnel, policymakers, and in the mass media. ... Hardening of schools seems to be a questionable endeavor for most schools, given the dearth of evidence regarding effectiveness.	Firearm violence remains a significant problem in the US (with 2787 adolescents killed in 2015). ... Parents seem to have a limited grasp of potentially effective interventions to reduce firearm violence.	22	0.2372755

Comment récupéré :

Les abstracts de citations ont été trouvés dans les articles qui citent les papiers suivants : [Payton et al 2017]Payton et al., 2017, [Vickers et al 2017]Vickers, 2017, [Karthik et al 2014]Karthik et al., 2014, [Peinelt et al, 2020]Peinelt, Nguyen

et Liakata, 2020. Les abstracts référés sont extraits directement dans ces papiers référés.

A.3 Tokenization de GPT et Fasttext

D'un autre côté, GPT n'utilise pas un mécanisme de division en sous-mots. Son tokenizer se fonde sur le Codage de Paires d'Octets (Byte-Pair-Encoding) pour réaliser la tokenisation. La limite maximale de tokens est de 1024 pour le tokenizer de GPT, mais il ne peut pas effectuer automatiquement la tokenisation d'une paire de phrases.

En ce qui concerne FastText, les représentations vectorielles des mots sont élaborées à partir des vecteurs des sous-chaînes de caractères dont ils sont composés. Cette approche permet de générer des vecteurs, même pour les mots mal orthographiés ou absents du vocabulaire du modèle.

A.4 Autres formats

A.4.1 Contexte citant et le contexte référé

Dans ce scénario, nous comparons la similarité entre les contextes de citations dans les papiers citants et les contextes de référence. Pour chaque paire de citations de notre jeu de données, nous utilisons les modèles de langage pour obtenir d'abord les représentations vectorielles des deux phrases séparément. Ensuite, nous appliquons la métrique de similarité cosinus sur ces deux vecteurs, enregistrant ainsi le score de similarité dans un fichier.

A.4.2 Abstract citant et abstract référé

Ici, nous évaluons la similarité entre l'abstract du papier citant et l'abstract du papier référencé. Dans notre jeu de données, la taille des abstracts varie d'un document à l'autre.

Dans certains articles, comme celui de Vickers (2017)Vickers, 2017, l'abstract est très court, ne contenant que 38 mots. En revanche, d'autres articles, tels que [Andrade C 2011]Andrade, 2011, indiquent que la plupart des revues exigent un abstract de 200 à 250 mots. En calculant la similarité entre des abstracts de tailles différentes, nous avons constaté que la métrique de similarité cosinus est fortement influencée par la différence de taille entre les phrases. Dans la figure 15Plus cette différence est importante, plus le score de similarité est bas.

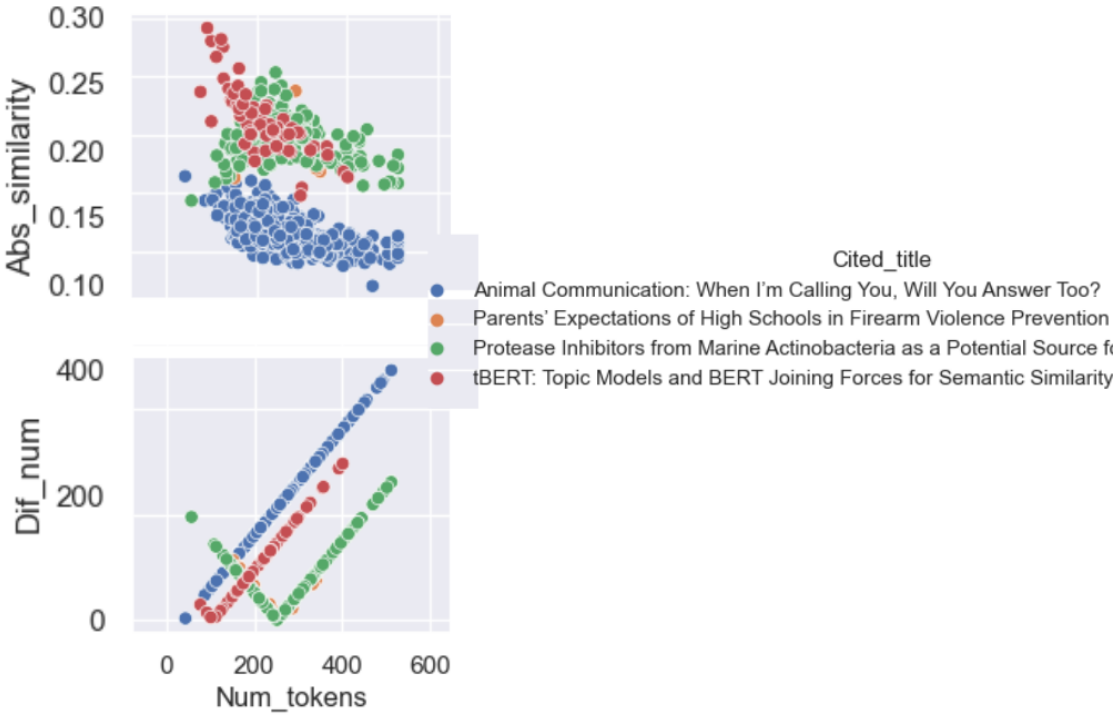

FIGURE 15 – Nombre de tokens et la similarité cosinus

Chaque couleur représente un papier référencé, et les points de cette couleur représentent l'abstract du papier qui a cité ce référencé. Par exemple, un point bleu représente l'abstract d'un papier ayant cité le document "Animal Communication : when I'm calling you, will you answer too ?" L'axe "Dif-num" représente la différence de nombres de tokens entre l'abstract du papier citant et l'abstract du papier référencé.

Il est clairement visible dans le graphique que lorsque la différence entre le nombre de tokens diminue, la similarité cosinus a tendance à augmenter. Par exemple, pour les points verts, à mesure que la différence de nombre de tokens se rapproche de zéro, la similarité cosinus atteint un pic, puis diminue à mesure que la différence de

nombre de tokens augmente. Les points d'autres couleurs montrent une tendance similaire. En résumé, il est évident que la différence de taille entre les textes a un impact significatif sur la similarité cosinus.

A.5 Résultat sur jeu de données Good-in-Bad-out

Ici, l'axe y représente le nombre de citations, et l'axe x représente la note de similarité cosinus entre le contexte de citation et le contexte dans le papier référé. Il y a en total 10 citations, dont 4 sont mauvaises et hors-domaine, 6 sont bonnes et dans-le-domaine.

A.5.1 Format contexte-contexte

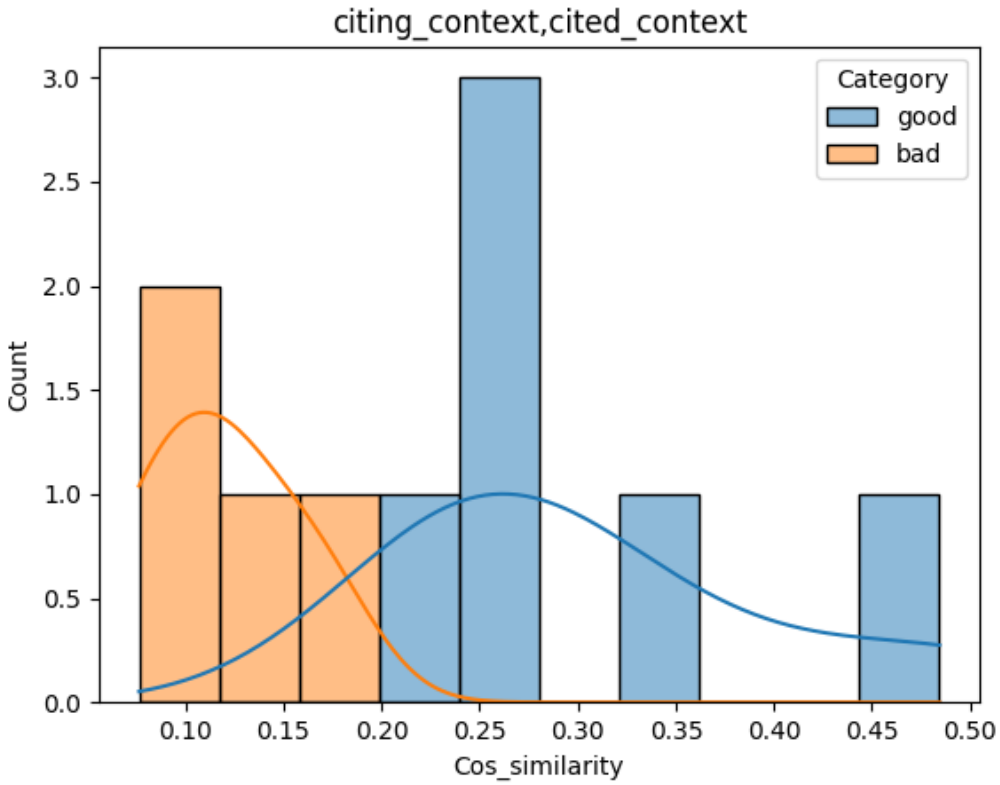

FIGURE 16 – contexte de citation avec contexte référé

A.5.2 Format abstract abstract

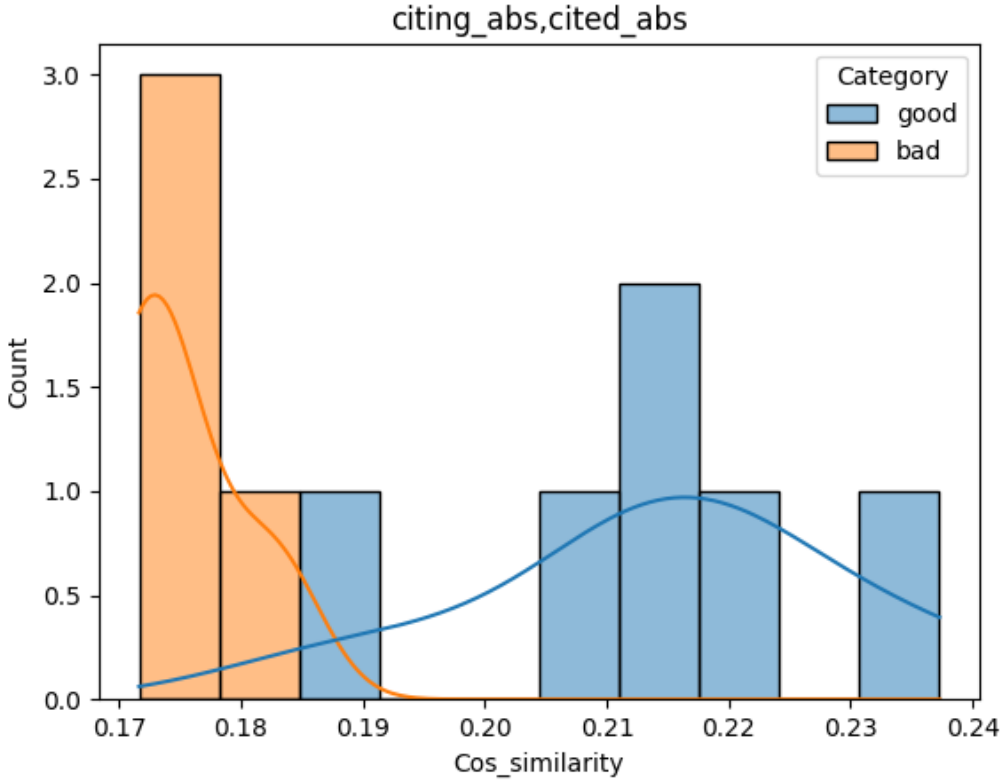

FIGURE 17 – l'abstract du papier citant avec l'abstract du papier référé

Table des matières

MOT-CLES : la pertinence de citations, similarité textuelle, articles scientifiques

Résumé

Les citations jouent un rôle important dans la recherche scientifique. Cependant, des études récentes ont révélé un problème préoccupant : de nombreuses citations inexactes sont présentes dans les articles scientifiques. Ces références erronées peuvent entraîner des interprétations erronées, une déformation des intentions de l'auteur original et potentiellement même des conséquences plus graves. Pour faire face à cette préoccupation, notre article présente deux heuristiques conçues pour évaluer l'exactitude des citations en utilisant des techniques de traitement automatique du langage (TAL). La première heuristique consiste à mesurer les similitudes textuelles à l'aide de représentations vectorielles. La deuxième heuristique implique l'utilisation d'une approche de classification de texte affinée pour trouver les citations fiables et les citations erronées. Ces deux méthodologies adoptent des modèles de langage. Selon nos résultats expérimentaux, il semble que la méthode de classification de texte affinée a démontré la meilleure performance.

KEYWORDS : citation accuracy, similarity of text, scientific paper

Abstract

Citations play an important role in scientific research. However, recent studies have unveiled a concerning issue : numerous inaccurate citations are present within scientific papers. These erroneous references can result in misinterpretations, distortion of the original author's intentions, and potentially even more serious consequences. To address this concern, our paper introduces two heuristics designed to assess citation accuracy using Natural Language Processing (NLP) techniques. The first heuristic involves measuring text similarities through vectorized text representations. The second heuristic entails employing a fine-tuned text classification approach to distinguish between reliable and flawed citations. Both of these methodologies adopt language models. Based on our experimental findings, it appears that the fine-tuned text classification method demonstrated the most optimal performance.