الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 8 ماي 1945 بسيدي بلعباس
École Supérieure en Informatique
8 Mai 1945 Sidi Bel Abbès

MÉMOIRE

Pour l'obtention du diplôme de **Master**
Filière : **Informatique**
Spécialité : **Système d'Information et Web (SIW)**

Thème

Mesures de similarité textuelle et modèles de type Transformer :

État de l'art

Réalisé par :
Lamisse Fatiha BOUABDELLI

Encadré par :
Pr. BENSLIMANE Sidi Mohammed (ESI-SBA)
Dr. BARON Mickael (LIAS)
Dr. CHARDIN Brice (LIAS)
Dr. JEAN Stéphane (LIAS)

Soutenu le 04 juillet 2024, devant le jury composé de :

Pr. Mimoun MALKI (ESI-SBA)	Professeur	Président
Dr. Badia KLOUCHE (ESI-SBA)	Docteur	Examinatrice
Pr. BENSLIMANE Sidi Mohammed (ESI-SBA)	Professeur	Encadrant
Dr. BARON Mickael (LIAS)	Ingénieur	Co-encadrant

Année Universitaire : 2023-2024

RÉSUMÉ

Ce mémoire explore les diverses mesures de similarité textuelle, telles que la similarité cosinus et la distance de Levenshtein, ainsi que les modèles de type Transformer, tels que RoBERTa, BERT et CamemBERT, dans le cadre de la comparaison de textes. L'objectif principal est de fournir une revue de littérature exhaustive sur ces concepts, en mettant en évidence les défis et les solutions existantes. Nous analysons ces techniques de comparaison textuelle pour comprendre leur efficacité dans divers contextes, en particulier ceux nécessitant une interprétation sémantique fine.

Nous abordons également les spécificités des modèles Transformer dans le cadre du traitement du langage naturel (NLP) et de l'apprentissage automatique (ML). Ce travail met en lumière l'importance de choisir des approches adaptées pour améliorer la qualité des analyses textuelles dans des domaines spécialisés.

Mots-clés : Similarité textuelle, Modèles Transformer, BERT, CammeBERT, RoBERTa, , Traitement du langage naturel, Apprentissage automatique,similarité cosinus, distance de Levenshtein

ABSTRACT

This thesis explores various text similarity measures, such as cosine similarity and Levenshtein distance, as well as Transformer-based models, such as RoBERTa, BERT, and CamemBERT, in the context of text comparison. The primary objective is to provide a comprehensive literature review on these concepts, highlighting existing challenges and solutions. We analyze these text comparison techniques to understand their effectiveness in different contexts, particularly those requiring fine semantic interpretation.

We also address the specifics of Transformer models within the scope of Natural Language Processing (NLP) and Machine Learning (ML). This work emphasizes the importance of selecting appropriate approaches to improve the quality of textual analyses in specialized domains.

Keywords : Text similarity, Transformer models, BERT, CamemBERT, RoBERTa, Natural Language Processing, Machine Learning, Cosine similarity, Levenshtein distance.

الملخص

يستكشف هذا البحث مقاييس التشابه النصي المتنوعة، مثل التشابه الكوني ومسافة ليفنشتاين، بالإضافة إلى نماذج من نوع المحولات مثل في سياق مقارنة النصوص. الهدف الرئيسي هو تقديم مراجعة شاملة للأدبيات حول هذه المفاهيم، مع إبراز التحديات والحلول الموجودة. نحن نحلل هذه التقنيات لمقارنة النصوص لفهم فعاليتها في سياقات مختلفة، وخاصة تلك التي تتطلب تفسيرًا دقيقًا للمعاني.

كما نتناول أيضًا خصائص نماذج المحولات في سياق معالجة اللغة الطبيعية والتعلم الآلي . يسلط هذا العمل الضوء على أهمية اختيار الأساليب المناسبة لتحسين جودة التحليلات النصية في المجالات المتخصصة.

الكلمات المفتاحية: التشابه النصي، نماذج المحولات، معالجة اللغة الطبيعية، التعلم الآلي، التشابه الكوني، مسافة ليفنشتاين.