جامعة صنعاء

الدراسات العليا والبحث العلمي

كلية الحاسوب وتكنلوجيا المعلومات

قسم علوم الحاسوب

برنامج الدكتوراه (الذكاء الاصطناعي)

تطوير نموذج جديد لمطابقة السلاسل العربية باستخدام التعلم الآلي

Developing a Novel model for Arabic string matching using machine learning

أطروحة مقدمة إلى قسم علوم الحاسوب، كلية الحاسوب وتكنلوجيا المعلومات، جامعة صنعاء

كجزء من متطلبات الحصول على درجة دكتوراه الفلسفة في علوم الحاسوب "الذكاء الاصطناعي"

الباحث

صلاح عبده محمد الحجري

المشرف الرئيسي

أ.د/ غالب حمود الجعفري

أستاذ علوم الحاسوب "الذكاء الاصطناعي"

جامعة صنعاء

2024

1445

الملخص:

في مجال الذكاء الاصطناعي(AI) ، حدثت طفرة ملحوظة مع ظهور النماذج اللغوية الكبيرة (LLMs)التي تم تحسينها لتتبع تعليمات البشر. أحد هذه النماذج ChatGPT (Chat Generative Pre-trained Transformer) من OpenAI ، والذي أثبت أنه أداة قادرة للغاية للعديد من المهام بما في ذلك الإجابة عن الأسئلة وتصحيح الأكواد وإنشاء الحوارات. ومع ذلك، على الرغم من أن هذه النماذج تشتهر بإتقانها للعديد من اللغات، إلا أن قدرتها على تحليل المشاعر بدقة، وخاصة في اللغة العربية، لم يتم التحقيق فيها بشكل واسع. ومن هذا المنطلق، نسعى لمعالجة هذا القصور من خلال إجراء تقييم شامل لقدرات ChatGPT في تحليل المشاعر العربية على وجه التحديد. تهدف هذه الأطروحة إلى اقتراح نموذج مبتكر لتحليل المشاعر في اللغة العربية باستخدام تقنيات التعلم الآلي. تتناول الأطروحة تحديات حرجة في تصنيف البيانات (DL)، ومطابقة السلاسل(SM)، وتوسيع البيانات(DA)، والنماذج اللغوية الكبيرة (LLM) وجمع البيانات لتحليل المشاعر في اللغة العربية. يعد تحليل المشاعر أمرًا بالغ الأهمية لفهم الآراء والمشاعر المعبر عنها في النصوص. ومع ذلك، فإن تطبيق تحليل المشاعر على النصوص العربية يقدم تحديات فريدة بسبب تركيبها اللغوي المعقد ونقص الموارد المتاحة بسهولة. يعتبر تحليل المشاعر العربي الدقيق أمرًا أساسيًا في تطبيقات متنوعة مثل مراقبة وسائل التواصل الاجتماعي، وتحليل تعليقات العملاء، والترجمة الآلية. تهدف هذه الدراسة إلى تحسين تقنيات تحليل المشاعر العربي لتسهيل فهم أفضل للمشاعر في النصوص العربية. الهدف الرئيسي من هذه الأطروحة هو تطوير نموذج مبتكر لتحليل المشاعر العربية يتفوق في الدقة على الأساليب الاخرى. ويتم تحقيق ذلك من خلال معالجة التحديات في أربعة مجالات رئيسية: أولاً، يتم تقديم نهج جديد يجمع بين ChatGPT والمُعلقين البشريين، مما يحقق دقة فائقة مقارنة بكل طريقة على حدة. يتم التحقق من صحة هذه الطريقة على مجموعة بيانات جديدة من تعليقات عملاء البنوك وشركات الصرافة في اليمن لتطبيقات أندرويد.

ثانيًا، يستكشف البحث دمج خوارزمية Levenshtein Distance مع خوارزميات K-Nearest Neighborsو Naive Bayesلتحسين أداء تحليل المشاعر العربي. يتضمن هذا النهج تقطيع الكلمات وإزالة الكلمات الشائعة وتقليل حجم متجه الكلمات، مما يؤدي إلى تصنيف أكثر دقة للمشاعر. ثالثًا، يتم التحقيق من فعالية دمج ChatGPT و Gemini Googleلتوسيع البيانات. تُظهر هذه الطريقة، إلى جانب التصنيف البشري لتحليل المشاعر، أداءً أفضل من الطرق الاخرى. رابعًا، يتم تقييم ChatGPT 4 و ChatGPT 3.5لأول مرة في مهام تحليل المشاعر العربية. يبحث هذا البحث في التعلم بدون أمثلة باستخدام نماذج لغوية كبيرة مختلفة ويقدم تقييماً شاملاً لتحليل المشاعر العربية على مجموعة بيانات جديدة من تعليقات يوتيوب من كأس العالم في قطر وكذلك مجموعة بيانات من الدراما اليمنية. تعترف الأطروحة بالتحديات المتأصلة في تحليل المشاعر العربية، بما في ذلك الطبيعة المعقدة للنصوص العربية، وقلة توفر بيانات التدريب، والحاجة إلى خوارزميات قوية لالتقاط المشاعر بدقة. يقدم هذا البحث حلولاً مبتكرة تعالج هذه التحديات المذكورة. يظهر النموذج المقترح تحسينًا كبيرًا في دقة تحليل المشاعر العربي مقارنة بالأساليب الحالية. بالإضافة إلى ذلك، يمثل تقييم ChatGPT 4 و ChatGPT 3.5 لمهام تحليل المشاعر العربي جهدًا رائدًا في هذا المجال. تسهم النتائج التي توصلت إليها هذه الأطروحة بشكل كبير في مجال تحليل المشاعر العربية. يقدم النموذج المقترح حلاً قويًا وفعالًا لتحسين دقة تحليل المشاعر العربية، مما يمهد الطريق لتطبيقات أكثر تقدمًا في مجالات متنوعة.

Abstract

In the field of Artificial Intelligence (AI), a remarkable surge has occurred with the advent of Large Language Models (LLMs) that have been fine-tuned to follow human instructions. One such model is ChatGPT (Chat Generative Pre-trained Transformer) from OpenAI, which has proven to be an extremely capable tool for various tasks, including answering questions, correcting codes, and generating dialogues. However, despite the renowned proficiency of these models in mastering numerous languages, their ability to accurately perform sentiment analysis (SA), particularly in Arabic and has not been extensively investigated. From this perspective, we aim to address this gap by conducting a comprehensive evaluation of ChatGPT's capabilities in Arabic Sentiment Analysis (ASA) specifically. This dissertation proposes a novel model for ASA using Machine Learning (ML) techniques. It addresses critical challenges in data labeling (DL), string matching (SM), data augmentation (DA), and the application of LLMs and data scraping for ASA. SA is crucial for understanding opinions and emotions expressed in texts. However, applying SA to Arabic texts presents unique challenges due to its complex morphology and the lack of readily available resources. Accurate ASA is essential in various applications, such as social media monitoring, customer feedback analysis, and machine translation. This study aims to improve ASA techniques to facilitate a better understanding of sentiments in Arabic texts. The main objective of this dissertation is to develop an innovative model for ASA that surpasses other methods in accuracy. This is achieved by addressing challenges in four key areas: **First**, a novel approach is introduced that combines ChatGPT with human annotators, achieving superior accuracy compared to each method individually. This method is validated on a new

dataset of customer comments from banks and exchange companies in Yemen for Android applications. **Second**; the research explores integrating the Levenshtein Distance (LD) algorithm with K-Nearest Neighbors (K-NN) and Naive Bayes (NB) for improved ASA performance. This approach involves stemming words, removing stop words, and reducing word vector size, leading to more accurate sentiment classification. **Thirdly**, the effectiveness of integrating ChatGPT and Gemini Google for DA is investigated. This method, along with human classification for SA, shows better performance compared to other techniques. **Fourth**, ChatGPT 4 and ChatGPT 3.5 are evaluated for the first time in ASA tasks. This research investigates zero-shot learning using various LLMs and provides a comprehensive evaluation for ASA on a new dataset of YouTube comments from the Qatar World Cup, as well as a dataset from Yemeni drama series. The dissertation acknowledges the inherent challenges in ASA, including the complex nature of Arabic texts, limited availability of training data, and the need for robust algorithms to capture sentiment accurately. This research introduces innovative solutions that address these mentioned challenges. The proposed model demonstrates significant improvement in ASA accuracy compared to existing methods. Additionally, the evaluation of ChatGPT 4 and 3.5 for ASA tasks represents a pioneering effort in this field. The findings from this dissertation contribute significantly to the field of SA for Arabic texts. The proposed model offers a robust and effective solution for improving ASA accuracy, paving the way for more advanced applications in various domains.