وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

نظام ذكي لتحويل الكلمات العامية الانكليزية إلى كلمات رسمية

رسالة مقدمة
إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
جزءاً من متطلبات نيل شهادة دبلوم عالٍ في
علوم الحاسوب

من الطالب

أحمد عبدالستار إبراهيم اليونس

بإشراف

م.د. بان شريف مصطفى

١٤٤٥ هـ ٢٠٢٣ م

الملخص

يواجه العديد من المستخدمين تحديا يتمثل في استخدام الكلمات العامية غير المعترف بها في القواميس القياسية، وتكون هذه الكلمات مفهومة ضمن النطاق الضيق للمجتمع الذي ينتمون إليه، وقد تكون محدودة للمجتمع المحلي، ومع ذلك، يواجه المستخدمون الذين ينتمون إلى خارج هذا المجتمع صعوبة في فهم النصوص التي تحتوي على تلك الكلمات العامية، وتعد هذه المشكلة تحديا كبيرًا أمام وظائف معالجة اللغة الطبيعية المتاحة حاليا مثل الترجمة وتلخيص النصوص والإجابة على الأسئلة، في بعض الأحيان تتعرف نماذج معالجة اللغة الطبيعية على بعض المفردات العامية، ولكنها تواجه صعوبة في التعامل مع مفردات اللهجات المحلية وفهمها، ومن هنا يصبح وجود نماذج معالجة اللغة الطبيعية التي تتعامل مع اللهجات المحلية ذات أهمية بالغة في تطوير مجال البحوث المستقبلية.

قد اشتهر العديد من الباحثين بتطوير نماذج تحويل الكلمات العامية إلى الكلمات الفصحى وقد حققت تلك النماذج بعض النجاح في فهم الكلمات العامية وتحويلها إلى لغة قياسية. ومع ذلك لا تزال هناك صعوبات في إيجاد حلول شاملة لهذه المشكلة، إذ تختلف العاميات واللهجات من منطقة إلى أخرى ومن مجتمع إلى آخر، فضلاً عن ذلك يمكن أن تكون الكلمات العامية متعددة الدلالات وتعتمد على السياق لفهم معناها الصحيح من أجل التغلب على هذه التحديات، ويمكن تطوير نماذج معالجة اللغة الطبيعية التي تأخذ في الاعتبار العاميات واللهجات المحلية وتسعى لفهمها وترجمتها بشكل أفضل، يمكن أيضًا تطوير قواعد بيانات ضخمة تحتوي على العديد من الكلمات العامية والعبارات المستخدمة في اللهجات المحلية واستخدامها في تدريب النماذج لتحسين أدائها في التعامل مع هذه الكلمات.

تم في هذه الرسالة اختيار مجموعة بيانات أولية من موقع (GitHub) تم تغيرها بما يخدم المجتمع وللحصول على الأداء الأفضل، بلغ عدد الكلمات (233) كلمة عامية مختارة من اللغة الإنجليزية.

تم استخدم لغة بايثون ومنصة (Colab) لتنفيذ العديد من النماذج المتاحة في منصة (Hugging Face)، ولا سيما المكتبة الأساسية (Transformer). وبعد إجراء عملية التدريب على 80% من مجموعة البيانات وعملية الاختبار على 20% من مجموعة البيانات، توصلنا إلى أن النموذج (facebook/bart-base) أظهر نتائج الأفضل في الأداء بقيمة دالة الخسارة (0,05299).

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Intelligent System to Transform English Slang Words into Formal Words

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Higher Diploma
in
Computer Science**

By

Ahmed Abdel Sattar Ibrahim Al-Younes

Supervised by

Lecturer.Dr. Ban Sharif Mustafa

2023 A.D. **1445 A.H.**

Abstract

Many users face the challenge of using colloquial words that are not recognized in standard dictionaries. These words are understood within the narrow scope of the community they belong to and may be limited to the local community. However, users from outside that community struggle to understand texts containing these colloquial words. This poses a significant challenge for current natural language processing tasks such as translation, text summarization, and question answering.

Sometimes, natural language processing models can recognize some colloquial vocabulary, but they struggle with local dialects and their understanding. Hence, the existence of natural language processing models that handle local dialects is of great importance for the development of future research .

Several researchers have become renowned for developing models that transform colloquial words into standard language. These models have achieved some success in understanding and converting colloquial words into the standard language. However, comprehensive solutions to this problem are still challenging to find. Colloquial expressions and dialects vary from one region to another and from one community to another. Furthermore, colloquial words can have multiple meanings that rely on context to understand their correct interpretation. To overcome these challenges, natural language processing models can be developed to better understand and translate colloquial expressions and local dialects. Additionally, large databases containing numerous colloquial words and phrases used in local dialects can be developed and utilized to train models and improve their performance in handling these words.

In this study, an initial dataset was selected from the GitHub platform and modified to be compatible with the current community and achieve the best performance. The dataset consisted of (233) colloquial words in English. Python language and the Colab platform were used to implement various models available on the Hugging Face platform, especially the Transformer library. After training on 80% of the dataset and testing on the remaining 20%, it was found that the "facebook/bart-base" model showed the best performance with a loss function value of (0.05299).