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

قياس نسبة التضليل وموثوقية المحتوى الرقمي في منصة X

رسالة مقدمة
إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
وهي جزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب

من قبل
صفا محمد علي شيت حسين

بإشراف
أ. م. د باسم محمد محمود علي

١٤٤٧ هـ
٢٠٢٥ م

الملخص

يعد مجال كشف المعلومات المضللة في الشبكات الاجتماعية مجالا ناشئا ويحتاج إلى تقنيات حديثة وطرائق جديدة ومتطورة بشكل مستمر، ومشكلة بحثية بالغة الأهمية في مجال الشبكات الاجتماعية. نظرا لتزايد كمية المحتوى غير المرغوب فيه والمضلل بكميات كبيرة بسبب تطور تقنيات الذكاء الاصطناعي التوليدي والنماذج اللغوية الكبيرة وخاصة منصة X مع تزايد الحسابات الوهمية وانتشارها وروبوتات الدردشة الآلية، لذلك توصي الرسالة بطريقة جديدة لنظام مقترح لكشف المعلومات المضللة والمحتوى غير المرغوب فيه (المضلل) . إذ تعتمد الطريقة المقترحة على تحليل المحتوى وتصنيف النصوص لتسمح للمستفيدين والمستخدمين وإدارات منصات التواصل الاجتماعي والمؤسسات في اتخاذ قرارات بتشخيص المحتوى المضلل الذي زاد انتشاره باستخدام طريقة جديدة وذلك بدمج طرائق عدة مع بعضها البعض للحصول على مقدار النسبة المئوية لموثوقية المحتوى المنشور، إذ تم تطوير نظام مقترح رئيس يتكون من 7 أنظمة هجينة مقترحة (3 أنظمة رئيسة و4 أنظمة فرعية) والنظام الرئيس المقترح الأول المطور لفحص الحسابات الوهمية يتكون من نموذج هجين مطور (RF و DT وBotometer) مع نموذج الاساس (Generative Pre-Trained Transformer-2:GPT-2)، وتم الحصول على (Score الأول) يمثل النسبة المئوية لفحص حسابات المستخدمين الحقيقية والآلية و النسبة المئوية لسمعة المستخدم كإسهام أول، ودمجه مع النظام الرئيس المقترح الثاني المطور الهجين المكون من(GPT2 وDistlBERT وBERT) لفحص موثوقية التغريدات كإسهام ثاني لهذا العمل و استخدام التعلم العميق كنموذج هجين مطور لموديل الذكاء الاصطناعي التوليديGPT2 لتحليل النصوص المنشورة وتم الحصول على (Score الثاني) يمثل النسبة المئوية لموثوقية التغريدات ودمجه بعد ذلك مع النظام الرئيس المقترح الثالث لتحليل تشابه التغريدات ويتكون من 4 أنظمة مقترحة فرعية لتحليل التشابه (الدلالي والتركيبي والحرفي والمشاعر) ودمجت معاً لحساب ال (Score الثالث) يمثل النسبة المئوية لتشابه التغريدات المنشورة باستخدام نموذج هجين مطور(Cosine Similarity وMean وMax) مع GPT2 على مجموعة البيانات في منصة X كإسهام ثالث في نموذج هجين مطور. وللحصول على النسبة المئوية النهائية Score، تم إدخال نتيجة كل من Scoreالأول لفحص الحسابات والسمعة مع نتيجةScore الثاني لنموذج الذكاء الاصطناعي التوليدي المدرب مسبقا على مجموعة البيانات مع Score الثالث الناتج من

حساب نسبة التشابه بين التغريدات لإخراج Score نهائي لنسبة التضليل للمحتوى المنشور في التغريدات. وقد وصلت الدقة في النموذج الهجين الأول المطور(GPT2 و RF و DT وBotometer) 92.3% ونسبة الدقة في نموذج فحص الموثوقية للمحتوى في النموذج الهجين الثاني المطور(GPT2 وDistlBERT وBERT) 79.37%، أما نسبة الدقة في نموذج تحليل التشابه في النموذج الهجين الثالث المطور(GPT2 و Cosine Similarity وMean وMax) فوصلت إلى80.00%، وهذا النظام المقترح اثبت انه الأفضل من حيث الدقة في الفحص ليساعد المستخدم على اتخاذ قرار مناسب وتمييز المحتوى المشبوه الذي يتم نشره في صفحات التواصل الاجتماعي. ويسهم في تحقيق بعض أهداف التنمية المستدامة منها (التعليم الجيد والعمل اللائق ونمو الاقتصاد والصناعة والابتكار والهياكل الاساس ومدن ومجتمعات محلية مستدامة وتحقيق السلام والعدل).

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

Quantifying Misinformation Level And Digital Content Credibility On X Platform

A Thesis Submitted to the Council of the College of Computer Science and Mathematics University of Mosul is a Partial Fulfillment of Requirements for the Degree of Master of Science in Computer Science

By

Safa Mohammad Ali Sheet Hussain

Supervised by

Associate Prof. Dr. Basim Mohammed Mahmood Ali

2025 A.D. **1447 A.H.**

Abstract

The field of disinformation detection in social networks is an emerging domain that requires advanced and continuously evolving techniques and methods. It is a critical research challenge within the realm of social media platforms, especially with the increasing amount of unwanted and misleading content due to the development of generative AI techniques and large language models, particularly on Platform X. The rise of fake accounts and the proliferation of chatbots have compounded this problem. Therefore, this thesis proposes a novel approach for detecting disinformation and unwanted content. The proposed method is based on content analysis and text classification, enabling users, social media platform administrators, and institutions to make decisions regarding the diagnosis of misleading content and the removal of fake pages that have proliferated. This is achieved by integrating several methods to calculate the reliability percentage of the published content.

A main system has been developed, consisting of seven proposed hybrid systems (three main systems and four sub-systems). The first proposed main system is developed for detecting fake accounts, and the first score represents the percentage of real and automated user accounts examined using a hybrid model developed with GPT-2, along with the user's reputation percentage as the first contribution. This is integrated with the second proposed main system developed to examine the reliability of tweets as the second contribution to this work. Deep learning is used as a hybrid model for the generative AI model GPT-2 to analyze the posted texts, and the second score represents the reliability percentage of the tweets. This is further integrated with the third proposed main system for analyzing tweet similarity, which consists of four proposed sub-systems (for semantic, syntactic, lexical, and sentiment similarity). These were combined to calculate the third score, representing the percentage of similarity between

the posted tweets using a developed model with GPT-2 on the dataset from Platform X as the third contribution to the hybrid model.

To obtain the final percentage score, the results from the first score for account and reputation verification, the second score from the pre-trained generative AI model on the dataset, and the third score from calculating the similarity percentage between tweets were inputted to produce the final score for the level of misinformation in the published content on the tweets.

The accuracy of the first developed hybrid model (GPT-2 & RF & DT & Botometer) reached 92.3%, and the accuracy of the second model for content reliability verification (GPT-2 & DistilBERT & BERT) reached 79.37%. The accuracy of the third model for similarity analysis (GPT-2 & Cosine Similarity & Mean & Max) was 80.00%. This proposed approach has proven to be the best in terms of speed and accuracy in helping users make appropriate decisions and distinguish between suspicious content being published on social media platforms. It also contributes to achieving some of the Sustainable Development Goals, including quality education, decent work, economic growth, industry, innovation, infrastructure, sustainable cities and communities, and achieving peace and justice.