Offensive Language Detection in Arabic Social Media Using Machine Learning and the Five Phases Optimization Algorithm for Feature Selection

Prepared By

Saleem Nahed Abu Lehyeh

Supervisor By

Dr. Mahmoud Ahmad Omari

Abstract

Social communication has become very important in our lives, as it allows people to communicate with each other and share their feelings. However, it is misused by some individuals for the purposes of cyberbullying or using abusive language. Detecting offensive language in Arabic social media faces significant challenges due to the complex nature of the Arabic text, including the diversity of dialects. This creates difficulties in accurately identifying offensive content, necessitating the development of robust automated detection systems based on machine learning to mitigate the negative effects of digital toxins and promote safer online dialogue.

This study investigates the application of machine learning techniques combined with Five Phases Algorithm (FPA) for feature selection in detecting offensive Arabic content on social media. The main

aim of this study is to answer important question related to the formulation of a cost-sensitive function to achieve a balance between false positives and false negatives, and their impact on classifier performance and feature reduction. The impact of feature selection based on the FPA was evaluated by experimenting on different expressions of an optimal cost-sensitive function and testing classifiers such as Support Vector Randomization (SVM), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). The dataset “ArCybC” consisting of 4,505 tweets was used in experiments to evaluate the performance of machine learning models.

Experimental results showed that utilizing FPA improves the performance of machine learning models, especially in recall rate and F1-score. Specificaly, the SVM and RF classifiers showed improved accuracy and recall when using FPA, while DT showed improved recall despite a compromise in accuracy. The study highlights the important role of the FPA in refining feature selection that enhances the ability of classifiers to accurately detect offensive content.

Keywords: offensive language, cyberbullying, ArCyC dataset

اكتشاف اللغة الهجومية في وسائل التواصل الاجتماعي العربية باستخدام تعلم الآلة وخوارزمية الخمسة مراحل التحسينية لاختيار الميزات

اعداد

سليم ناهض سليم أبو لحية

اشراف

الدكتور محمود احمد العمري

الملخص

التواصل الاجتماعي أصبح أمرًا مهمًا بشكل كبير في حياتنا، حيث يتيح للأشخاص التواصل مع بعضهم البعض ومشاركة مشاعرهم. ومع ذلك، يساء استخدامه من قبل بعض الأفراد لأغراض التنمر الإلكتروني أو استخدام لغة مسيئة. يواجه اكتشاف اللغة المسيئة في وسائط التواصل الاجتماعي العربية تحديات كبيرة نظرًا لطبيعة النص العربي المعقدة، بما في ذلك تنوع اللهجات. وهذا يخلق صعوبات في التعرف بدقة على المحتوى المسيء، مما يستدعي تطوير أنظمة قوية للكشف الآلي بناءً على التعلم الآلي للتخفيف من الآثار السلبية للسموم الرقمية وتعزيز حوار أكثر أمانًا عبر الإنترنت.

تستكشف هذه الدراسة تطبيق تقنيات التعلم الآلي بالاشتراك مع "خوارزمية خمس مراحل" لاختيار السمات في اكتشاف المحتوى العربي المسيء على وسائل التواصل الاجتماعي. تهدف الدراسة إلى الإجابة على أسئلة مهمة تتعلق بصياغة دالة حساسة للتكلفة لتحقيق توازن بين الايجابيات الزائفة والسلبيات الزائفة، وتأثيرها على أداء المصنف وتقليل السمات. تم تقييم فعالية اختيار السمات بناءً على خوارزمية خمس مراحل من خلال التجربة على تعبيرات مختلفة لدالة حساسة للتكلفة مثلى واختبار المصنفات مثل Support Vector Machine (SVM)، Random Forest (RF)،

Decision Tree (DT)، Logistic Regression (LR). تم استخدام مجموعة البيانات المسماة بـ "ArCybC" والتي تتألف من 4505 تغريدة في التجارب لتقييم أداء نماذج التعلم الآلي . أظهرت نتائج التجارب أن دمج خوارزمية خمس مراحل يحسن أداء نماذج التعلم الآلي، خاصة في معدل الاستحضار (recall) و (F1-score). وبشكل خاص، أظهرت مصنفات SVM وRF تحسينًا في الدقة والاستحضار عند استخدام خمس مراحل، بينما أظهرت DT استحضارًا محسنًا على الرغم من التنازل في الدقة (accuracy). ويسلط البحث الضوء على الدور المهم لخوارزمية خمس مراحل في تنقية اختيار السمات، مما يعزز قدرة المصنفات على اكتشاف المحتوى المسيء بدقة.

الكلمات المفتاحية: اللغة الهجومية، التنمر الالكتروني، مجموعة بيانات التنمر الالكتروني