الكشف عن الجماعات الخبيثة في شبكات التواصل الاجتماعي بناء على تحليل السلوك الجماعي

ريم مسفر عبد الجبار الحارثي

اشراف
د. كوثر موريا و د. أريج الهذلي

الملخص

يتضمن السخام الإلكتروني العديد من الأنشطة الضارة التي تهدد مستخدمي الشبكات الاجتماعية من خلال مشاركة محتوى غير مرغوب فيه أو مضلل أو مسيء إما بإرسال إعلانات غير مرغوب فيها أو التلاعب بالرأي العام أو نشر برامج ضارة. وقد أدى ذلك إلى تطوير أبحاث ونهج كثيرة لمكافحة هذه الممارسات في الشبكة الاجتماعية. ومع ذلك، معظم الأعمال تم تطويرها للكشف عن مرسلي هذه المحتويات باللغة الإنجليزية وقليل من الجهد اعطي لمستخدمين العرب. وبالتالي ، تقدم هذه الأطروحة أولاً تحليلًا شاملاً لخصائص الحملات او المجموعات العربية الضارة على توتير. تضمنت الدراسة تحليل المجموعات من عدة جهات، بما في ذلك محتواها، والرسم البياني للتفاعلات الاجتماعية، وعمرها المتوقع، والأنشطة اليومية لهذه الحسابات. إلى جانب الكشف عن استراتيجيات هذه الحسابات، فقد وجد انها أكثر نجاحًا في تجنب الكشف من قبل توتير من الحسابات المذكورة مسبقًا في الاعمال السابقة. استنادًا إلى نتائج التحليل، تم اعتماد خوارزميات شبه خاضعة للإشراف، واستخدمت مجموعة من ١٦ ميزة للكشف عن حسابات الحملات الضارة تلقائيًا. تم اختبار أداء النموذج المقترح على مجموعة بيانات من ١٦٨٥ حسابًا، وأظهرت النتائج أن النموذج حقق دقة تساوي ٠,٩١. نظرًا لوجود حاجة كبيرة إلى تقنيات الكشف في الوقت الفعلي لتصفية التغريدات الضارة أو منخفضة الجودة ، فقد طورنا أيضًا أسلوبًا للتعلم العميق في الوقت الفعلي يستخدم البيانات النصية للتغريدة للكشف عن محتويات الضارة او السخام الالكتروني. تم تقييم النهج المقترح باستخدام مجموعة بيانات حقيقة تتضمن مجموعة واسعة من التغريدات منخفضة الجودة، ويظهر نموذجنا اداء متفوقا مقارنة بالأنظمة الموجودة في كل من الدقة وقياس F1 (٠,٩٨). لقد اقترحنا أيضًا طريقة بسيطة باستخدام التعلم العميق لتصنيف حسابات توتير في الوقت الفعلي الى حساب حقيقي او

حساب ضار بناء على التغريدات الأخيرة للحساب. الطريقة المقترحة حققت دقة تنصيف وقياس F1 عالية (٠,٩٨). في أقل من خمسة ميلي ثانية ، تمكن نظامنا في الوقت الفعلي من تصنيف تغريدة و تصنيف حساب في أقل من ثانية.

Detecting Malicious Campaigns in Social Networks based on A Collective Behavior Analysis

Reem Mesfer Alharthy

Supervised By
Dr. Kawthar Moria and Dr. Areej Alhothali

ABSTRACT

Social spam involves several malicious activities threatening social network users by sharing unwanted, misleading, or spiteful content to either send unwanted ads, manipulate public opinion, or spread harmful malware. This has led to the development of considerable researches and approaches to combat such practices in the social network. Most works, however, were developed to address the detection of English language spammers with little efforts for the Arabic language spammers. This thesis, therefore, first provides a comprehensive analysis of the characteristics of malicious Arab campaigns on Twitter. The study involved analyzing these campaigns from several perspectives, including their content, social interactions graph, lifespan, and day-to-day activities. Besides exposing their spamming tactics, these accounts were found to be more successful in avoiding Twitter suspension than previously reported spammers in the literature. Based on the outcomes of the analysis, two semi-supervised algorithms were adopted, and a set of 16 features were used to identify the campaigns' accounts automatically. The proposed model performance was tested on a dataset of 1685 accounts, and the results show that the model achieved a 0.91 accuracy. Since there is a considerable need for real-time detection techniques to filter malicious or low-quality tweets, we have also developed a real-time

deep-learning approach that utilizes tweet textual data to detect spam content and spammers profiles. The proposed approach was evaluated in a real-world dataset that includes a wide range of low-quality tweets, and our model demonstrates superior performance compared to existing solutions in both accuracy and F1 measure (0.98). We also proposed a lightweight deep learning approach to classify Twitter accounts as spam or genuine accounts based solely on the account recent tweets. The lightweight method yielded high accuracy scores and an F1 measure (0.98). In less than five milliseconds, our real-time approach was able to classify a tweet and an account in less than a second.