جامعة الفيوم
كلية الحاسبات والمعلومات
قسم نظم المعلومات

البحث رقم (٥)

Automated Ham-Spam Lexicon Generation Based on Semantic Relations Extraction

اسم البحث	**إنشاء معجم آلي للرسائل الصحيحة والغير مرغوب فيها على أساس استخراج العلاقات الدلالية**
ملخص المشكلة	يعد البريد الإلكتروني (البريد الإلكتروني) إحدى الطرق الأساسية الحالية للاتصال. تعتبر حاليًا الطريقة الرسمية للأنشطة التجارية المختلفة مثل عقد الاتفاقيات وإعداد الاجتماعات الرسمية والتعاون الجماعي. هذا الاهتمام المستمر برسائل البريد الإلكتروني كقناة اتصال لفت الانتباه إلى الحاجة إلى القضاء على البريد الغير مرغوب فيه والذي له تأثير حيوي على موارد الشبكة وأنشطة الأعمال. لذلك، يسلط البحث الضوء على الحاجة الحيوية لبناء مستكشف فعال لرسائل البريد الإلكتروني الغير مرغوب فيها من خلال اكتشاف علاقة متعددة الاتجاهات من النص الشخصي.
سياق البحث	ينصب تركيز البحث على إنشاء معجم لرسائل البريد الإلكتروني الشخصية والذي يعتمد على تقنيات تحليل النص. المفتاح الرئيسي للبحث المقترح هو الكشف عن الكلمات المفتاحية الموزونة التي هي ركيزة مهمة التصنيف الناجحة. أحد المفاتيح الناجحة في النهج المقترح أنه يركز على عنصرين من مكونات البريد الإلكتروني الرئيسية وهما الموضوع وأجزاء الرسالة بدلاً من التركيز على جزء الرسالة فقط،يوفر هذا النطاق اكتشافًا أكثر دقة للكلمات الرئيسية بالإضافة إلى تصنيف أكثر دقة النتائج.

جامعة الفيوم
كلية الحاسبات والمعلومات
قسم نظم المعلومات

أسلوب البحث	يقدم البحث إطار عمل للكشف عن علاقة متعددة الاتجاهات من نص شخصي. يهدف البحث إلى الكشف عن نوع رسائل البريد الإلكتروني الشخصية سواء أكانت بريدًا إلكترونيًا حقيقيًا أم بريدًا غير مرغوب فيه. يركز الإطار المقترح على عنصرين من مكونات البريد الإلكتروني "الموضوع والرسالة"، بينما يتم إجراء تحليل النص على المكونين مع تحديد العلاقة بين المصطلحات المميزة ومكونات البريد الإلكتروني. تتمثل الفكرة الرئيسية للإطار المقترح في تحديد المصطلحات الأساسية الشخصية عن طريق تحليل النص واستخراج أعضاء العلاقة المتضمنين، ثم تطبيق تقنية التصنيف لهدف اكتشاف البريد الغير مرغوب فيه. تضمن تحليل النص استخراج الكلمات الرئيسية، وإنشاء مصطلحات N-Gram ، وتحديد وزن لكل مصطلح باستخدام تقنية TF-IDF ، وأخيراً بناء العلاقة المضمنة التي توفر سمة إضافية للمصطلحات التي تم إنشاؤها. في المرحلة التجريبية من البحث، تم استخدام مجموعة بيانات من ست مجموعات بريد إلكتروني شخصية بإجمالي ٣٣٧١٦بريدًا إلكترونيًا مقسمة على ١٦,٥٤٥بريدًا إلكترونيًا هامًا و١٧١٧١ بريدًا إلكترونيًا غير مرغوب فيه تم جمعها في الفترة من مايو ٢٠٠١ إلى سبتمبر ٢٠٠٥. عدد الرموز المميزة المستخرجة بعد مرحلة ما قبل المعالجة كان ما مجموعه ٨٠٠١٧٦ توكن. تم إنشاء مصطلحات N-gram التي نتج عنها ١,٧٦٠,٣٨٦مصطلحًا بمدى من ١ إلى ٣. ثم تم تطبيق تقنية الترجيح TF-IDF لتوفير ترجيح للمصطلحات التي تم إنشاؤها، ثم تم إنشاء معجم شخصي للرسائل العشوائية غير المرغوب فيها. كخطوة أخيرة، تم تطبيق خمسة من خوارزميات التصنيف للكشف عن نوع رسائل البريد الإلكتروني. تم إجراء تقييم نتائج خوارزميات التصنيف من خلال نهج التحقق وبلغت نسبة النجاح ٨٩,٨٥٪ لخوارزمية الشبكة العصبية.
النتائج المستخلصة	يتبع الإطار المقترح نهج التعاون بين تقنيات التنقيب عن النصوص وتقنيات الترجيح بالإضافة إلى تقنيات التصنيف للكشف الفعال عن الرسائل الغير مرغوب فيها. يعد البحث خطوة في نظرة أوسع لإطار عمل تعاون وتواصل ذكي عام. يخدم النهج المقترح في مجال توجيه اكتشاف رسائل البريد الإلكتروني الغير مرغوب فيها وهو أمر مفيد للغاية لمزودي خدمة الإنترنت بما في ذلك خدمات Hotmail وYahoo وGmail وOutlook. إن دعم خدمات البريد الإلكتروني بتقنيات ذكية للكشف عن الرسائل غير المرغوب فيها من شأنه أن يقلل من مستوى الضعف في حسابات البريد الإلكتروني من التعرض لتهديدات مختلفة بما في ذلك انتشار الفيروسات والبيانات الحساسة للتصيد الاحتيالي، فضلاً عن تهديد خطير للنظام بأكمله، وبالتالي، سيكون فاعليا لإعاقة تهديد حسابات البريد الإلكتروني الشخصية بدءًا من صانعي القرار مروراً بجميع المستويات الإدارية إلى الموظفين العاملين في نظام المؤسسة.