جامعة البعث

كلية الهندسة المعلوماتية

قسم هندسة البرمجيات ونظم المعلومات

التنقيب عن المعلومات في محتوى الشبكات الاجتماعية

دراسة أعدت لنيل درجة الماجستير في الهندسة المعلوماتية باختصاص هندسة البرمجيات ونظم المعلومات

إعداد

م. هزار هيسم هرموش

إشراف

د. محسن حسين

أستاذ مساعد في قسم هندسة البرمجيات ونظم المعلومات

٢٠١٥ م-١٤٣٥هـ

ملخص

فتحت وسائل الإعلام الاجتماعية، وعلى وجه الخصوص منصات التدوين المصغر، إمكانيات جديدة للناس لاستهلاك وإنتاج البيانات. أصبح تويتر واحداً من أكثر منصات التدوين المصغر شعبية ويمتلك الملايين من المستخدمين الذين ينشرون يومياً كماً هائلاً من التغريدات التي تعبر عن التجارب الشخصية للأشخاص وتلهم المستخدمين المرتبطين معهم عبر تويتر. أتاح توافر هذا الغنى و الحجم الهائل من البيانات ومشاعرهم – خلافاً للأساليب التقليدية المحدودة و المعتمدة بشكل كبير المستخدمين على الويب حول آراء على استطلاعات الرأي – فرصة ذهبية لدراسة الرأي العام وحث الباحثين على تطوير تقنيات التنقيب في البيانات لتلائم طبيعة وسائل الإعلام الاجتماعية الضخمة و الموزعة و غير المهيكلة. ويشار ب "تحليل المشاعر و التنقيب في الآراء" إلى المجال البحثي ضمن حقل التنقيب في بيانات وسائل الإعلام الاجتماعية المختص بدراسة الرأي العام.

تتناول هذه الدراسة فعالية طرق تجميع المصنفات في تحسين أداء أنظمة تحليل المشاعر و التنقيب في الآراء ضمن تويتر. و بشكل أكثر دقة فعالية طرق تجميع المصنفات في تحسن الأداء في حال استخدمت عند انجاز المهمتين الأساسيتين: تصنيف الذاتية (تحديد فيما إذا كانت التغريدة تحتوي مشاعر أم لا) و تصنيف قطبية المشاعر (إيجابية أم سلبية). كان من الممكن تحويل المشكلة إلى تصنيف ثلاثي للتغريدات إلى(إيجابي، سلبي، و عادي)، لكن تم تفصيل الدراسة على المهمتين لنقدم الإجابة المناسبة لمن يريد استخدام تجميع المصنفات لبناء إحدى المهمتين لأغراض أخرى.

تتلخص الإضافة العلمية التي تقدمها هذه الدراسة في نقاط عدة. الأولى: عرض دقيق لتقنيات التنقيب في بيانات وسائل التواصل الاجتماعية. الثانية: مسح للجهود البحثية التي تناولت تحليل المشاعر : ضمن بيانات تويتر الثالثة: مقارنة لأربع أدوات مجانية ومفتوحة المصدر للتنقيب في البيانات. الرابعة : دراسة فعالية طرق التجميع في دراسة فعالية طرق التجميع في تحسين أداء تصنيف الذاتية. الخامسة : تحديد آلية وزن المميزات الأفضل عند استخدام المميزات تحسين تصنيف قطبية المشاعر. السادسة التقليدية. وأخيراً، توضيح تأثير تقنية المعالجة المسبقة المستعملة (استخراج المميزات و اختيار المميزات).

Abstract

Social media, in particular micro-blogging platforms, opened new possibilities for people to data consumption and production. Twitter has become one of the most popular micro-blogging platforms and has millions of users, who publish daily a huge amount of tweets that reflect their personal experiences and inspire their followers on Twitter. This richness and volume of data about user's opinions and feelings is available on the Web- as opposed to the traditional methods, limited and depends mainly on opinion polls- provided a golden opportunity for the study of public opinion and urged researchers to develop drilling techniques in the data to suit the nature of distributed and non-structured social media data. "Sentiment analysis and opinion mining" referred to the area of research within the field of social media mining specialist to study public opinion.

This study examines the effectiveness of the ensemble methods in improves the performance of Twitter sentiment analysis and opinion mining. More precisely; the effectiveness of the ensemble methods in improves performance when used to perform two main tasks of sentiment analysis: subjective classification (to determine whether Tweet contains feelings or not) and polarity classification (to determine whether Tweet contains positive or negative opinion). It was possible to transform the problem into a multi-class classification of Tweets (positive, negative, and neutral), but we provide a detailed study for each task to provide the appropriate answer for anyone who wants to use the ensemble methods in the construction of one of those two tasks for other purposes.

The scientific addendum provided by this thesis can be summarized by several points: first: exploration of social media mining techniques. Second: survey state of art research efforts in Twitter sentiment analysis.Third: Evaluating four of the most popular Open Source and Free Data Mining Tools. Fourth: Studding the effectiveness of the ensemble methods to outperform the state-of-art algorithms used to implement twitter subjectivity classification. Fifth: Studding the effectiveness of the ensemble methods to outperform the state-of-art algorithms used to implement twitter polarity classification. Sixth: identification of the best weighting scheme of features when using traditional BOW representation. Finally, discuss the effect of the use preprocessing method: feature extraction and feature selection.