نهج مزدوج لتحليل الآراء والمشاعر في النصوص العربية باستخدام تصنيف تعلم الآلة والتصنيف القائم على المعجم

مشرفة الرسالة / منيرة طيلب

اسم الطالبة / أمل ضيف الله الكبكبي

المستخلص

مع النمو الملحوظ لمنصات التواصل الاجتماعية، تنتشر كمية كبيرة من البيانات عبر الشبكات، وتحتوي هذه البيانات على معلومات قيمة يمكن أن تكون ذات فائدة كبيرة في العديد من المجالات. استخراج معلومات مفيدة من هذه البيانات أصبح تحديًا، وذلك بسبب أنها كبيرة في الحجم وذات أنواع مختلفة وسريعة النمو. يعد تحليل الآراء أحد الأساليب التي يمكن أن تساعد في استخراج المعلومات من كمية كبيرة من البيانات، وهو يعد من أحد مجالات البحث في التنقيب عن النصوص. تصنيف الآراء هي مهمة استخراج الرأي أو تصنيف قطبية جملة أو نص معين وتحديد ما إذا كان يحمل مشاعر إيجابية أو سلبية أو محايدة.

لقد تمت العديد من الدراسات في هذا المجال للغة الإنجليزية كما انه في السنوات الأخيرة هناك دراسات وأبحاث للغة العربية أيضًا، إلا أن اللغة العربية تعتبر أكثر صعوبة نظرًا لطبيعة اللغة العربية وقواعدها ووجود لهجات متعددة في الدول العربية. هناك مشكلة صعبة أخرى في تصنيف الآراء وهي الحاجة إلى البيانات التي تحمل علامات (مثل إيجابي أو سلبي) للتدريب وحقيقة أن عملية وضع العلامات يتم تنفيذها يدويًا بواسطة البشر وبالتالي فهي تستغرق وقتًا طويلاً. هناك مشكلة أخرى تتمثل في تحديد خوارزمية التعلم الآلي الأكثر ملاءمة لتصنيف البيانات بدقة.

في هذه الأطروحة، تم اقتراح نموذج هجين جديد لتحليل المشاعر والآراء، بهدف تحقيق تصنيف فعال لنصوص عربية غير معلمة بدقة عالية. يستخدم النموذج المقترح كلاً من النهج القائم على المعجم ونهج التعلم الآلي باستخدام تقنية التعلم في المجموعات: التصويت بالأغلبية. تمت معالجة مشكلة الحاجة إلى بيانات معلمة من خلال الاستفادة من النهج القائم على المعجم لتصنيف وتعليم النص. ويستخدم التعلم الجماعي القائم على التصويت بالأغلبية لتحسين أداء التصنيف، حيث يتم استخدام مصنفات متعددة بدلاً من مصنف واحد ويتم دمج نتائجها. تم النظر في عدة مجموعات من المصنفات، وأظهرت النتائج التجريبية أن أداء النموذج المقترح يتفوق على جميع النماذج باستخدام مصنف واحد.

Hybrid Approach for Arabic Sentiment Analysis Using Machine Learning and Lexicon-based classification

Supervised by\ **Mounira Taileb**
Student name\ **Amal Dhaifallah Alkabkabi**

Abstract

With the remarkable growth of the social media platforms, a large scale of data is scattered throughout the networks, those data contain valuable information that can be of a great help in many areas. It became a challenge to extract useful information from a big amount of data, since they are large in volume, variety and velocity. Sentiment analysis (SA) or Opinion Mining (OM) is one of the techniques that can help to extract information from a large amount of data, it is a research field in text mining. SA is the task of extracting the opinion or classifying the polarity of a given sentence or text and determine whether it carries positive, negative or neutral sentiments.

SA has been well studied for the English language, however, Arabic SA has been considered more challenging due to the natural of the Arabic language, its rules and the fact that multiple dialects exist in the Arab countries. Another challenging issue in SA classification is the need for labeled data for training and the fact that the labeling process is usually carried out manually by humans and thus it is time consuming. Another challenging issue is identifying the most suitable machine learning algorithm for classifying the data accurately.

In this thesis, a new hybrid sentiment analysis model is proposed for Modern Standard Arabic (MSA) text, with the goal of achieving an efficient classification of a given unlabeled Arabic text with optimal accuracy. The proposed model uses both lexicon-based approach and machine learning approach by using the ensemble learning technique: majority voting. The issues of data labeling are addressed by taking advantage of the lexicon-based approach for text labeling. The majority voting-based ensemble learning is used to improve the classification performance, where multiple classifiers are used instead of a single classifier and their results are combined.

Several sets of classifiers and test datasets have been considered. The experimental results show that the proposed model with the set of classifiers: Naive Bayes, Logistic Regression and Stochastic Gradient Descent outperforms all the models with a single classifier in terms of accuracy, recall, precision and F-score.