استخدام تقنيات التعلم الآلي لتحسين كفاءة نموذج تحليل المشاعر باللغه العربية

إعداد

سالي نواف عليان علاونه

إشراف

الدكتور محمد نصار

الملخص

في الفترة الاخيرة، تم انتشار العديد من التطبيقات والبرامج التي تعبرعن آراء و مشاعر المستخدمين حول وجهات نظر معينه ومتوفرة من خلال الشبكه العنكبوتيه ومثل هذه التطبيقات تويتر، فيس بوك و انستجرام. حيث هذه التطبيقات اتاحت الفرصه للباحثين لتطوير الكثير من الأبحاث في مجال تحليل المشاعر وبالذات في اللغه العربيه بالاعتماد على تقنيات تصنيف البيانات المتوافرة. لتساعد المستخدمين في اتخاذ القرار لشراء منتج معين وتساعد شركات الانتاج لتحسين منتجاتهم بالاعتماد على هذه الآراء .

واجه الباحثين الكتير من التحديات والعقبات مع استخدام اللغه العربيه في مجال تحليل المشاعر. من هذه التحديات طبيعة اللغه العربيه المعقده في مصطلحاتها ونحوها ومن ناحية اخرى قلة المصادر المتوفرة والادوات التي من الممكن استخدامها في هذا المجال.

هذه الدراسة تقدم نظرة شامله ومكثفه عن المراحل التي يجب اتباعها وتطبيقها على مجموعات البيانات التجريبية المتوفرة لتحسين ورفع كفاءتها في تحليل المشاعر خصوصا في مجال اللغه العربيه، سيتم ذلك عن طريق تقديم نموذج عام يحتوي على المراحل الاساسيه التي يجب اتباعها في مجال تحليل المشاعر. تبدأ هذه المراحل بتجهيز البيانات, يتضمن تجهيز البيانات ثلاثة خطوات: أولا: تقسيم قواعد البيانات بالاعتماد على قطبية النص سالبا ام موجبا ،ثانيا :تجهيزها من خلال التخلص من الكلمات احرف الجر وادوات الترقيم واستخلاص الجذر،ثالثا استخلاص خصائص الكلمه والممثله بوزنها ونسبة تكرارها وذلك باستخدام ثلاث طرق وهي:

1. استخلاص وزن الكلمه ونسبة تكرارها في الملف كامل (TF).
2. استخلاص وزن الكلمه ونسبة تكرارها في نص معين وعدم تكرارها في نص اخر(TF-IDF) .
3. استخلاص وزن الكلمه ونسبة تكرارها في المجلد المصنف الى قطبيه سالبه ام موجبه (TF-IDF-CF) .

هذه الدراسه تهدف الى اقتراح استخدام خوارزميات تصنيف البيانات المناسبه والتي تعطي نتاءج وكفاءه عاليه في مجال تحليل المشاعر في اللغه العربيه ومن ثم فحص هذه الخوارزميات وعمل مقارنه بينها ، وقد تمت هذه المقارنه على عدة اوجه وهي كالتالي :

1. تم المقارنه بين الخوارزميات التاليه KNN, SVM, DT,NB بتطبيقها على الطرق الثلاثه لاستخلاص الخصائص وهي : (TF,TF-IDF,TF-IDF-CF) ، وكانت النتيجه الافضل باستخدام خوارزمية SVM مع TF-IDF-CF .
2. تم المقارنه بين الخوارزميات التاليه KNN, SVM, DT,NB ولكن بعمل تعديلات على معاملات الخوارزميات السابقه وتطبيقها ايضا على الطرق الثلاثه التاليه : (TF,TF-IDF,TF-IDF-CF) ، وكانت النتيجه الافضل بتعديل معامل البذره (seed) لخوارزمية DT مع TF-IDF-CF .
3. تم تطبيق الخوارزميات التاليه Stacking و Baggingو Boosting، وكانت النتيجه الافضل استخدام خوارزمية ال bagging مع TF-IDF-CF .
4. تم تطبيق المقارنه عن طريق دمج خوارزميتين مع بعضهم البعض ، وكانت النتيجه الافضل دمج خوارزمية ال SVM مع KNN او DT باستخدام TF-IDF-CF

Using of Machine Learning Techniques for Improving Efficient Arabic Sentiment Analysis Model.

Prepared by:

Sally Nawwaf Alawneh

Supervised by:

Dr.Mohammad Othman Nassar

Abstract

Recently many popular web applications have been developed such as Facebook, Twitter and Instagram which express the feelings of users and their views on a particular topic available on the World Wide Web. Because of that researchers have developed much in the field of Opinion Mining specifically in Arabic; based on Machine Learning Techniques. This can help the customers make purchase decision and help businesses to improve the quality of products depend on users views.

For the Arabic language; there are some challenges facing researchers in this area, which are divided into two types: the first type is related to the

complex nature of the Arabic language, and the other is related to the resources and tools used and the difficulty of providing them.

This thesis provides comprehensive overview about the stages that must be applied on the available datasets to improve the effectiveness of Arabic sentiment analysis process by providing general framework that summarizes the main stages needed to be followed in Arabic sentiment analysis process. The main stages needed to be followed starts by preparing the dataset; preprocessing and feature extraction by using Vector Space Model to improve the effectiveness of Arabic sentiment analysis process, the following feature extraction methods: TF, TFIDF and TFIDFCF was used. Finally; testing and evaluating the model using machine learning techniques.

This thesis aims to compare Machine Learning methods to improve the effectiveness of Arabic sentiment analysis process. We divide the work into four main stages:

In the first stage we compared between the following Machine Learning methods: K-Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM).The comparison was based on using default parameters for each Machine Learning method; this

comparison was also conducted over three levels using the following proposed feature extraction methods: TF, TFIDF and TFIDFCF. We find that the best algorithm is Support Vector Machine (SVM) with TFIDFCF model.

In the second stage we manipulated the parameter's that can be used for each learning technique; we changed the default parameters to get better results; we get the best results when the seed parameter in decision tree (DT) classifier is modified with TFIDFCF model.

In the third stage; we used the available Ensemble learning models available in weka that combine more than one Machine Learning Technique such as Bagging, Boosting and Stacking; our results shows that those methods can enhance performance; the best one of them was the Bagging method which gives the highest results.

In the fourth stage; we suggested new combinations for the Machine Learning method to get enhanced results by using our own proposed Hybrid models; our results shows that we can enhance performance; we find the best technique was to join SVM with NB or KNN classifiers based on TFIDF model.