جامعة الموصل
كلية علوم الحاسوب والرياضيات

تحليل المشاعر لتغريدات تويتر العربية باستخدام آلة متجه الدعم

رنا زهير عبد الغني العبيدي

رسالة ماجستير
علوم الحاسوب

بإشراف
د. غيداء عبد العزيز الطالب
أستاذ مساعد

١٤٤١هـ ٢٠١٩ م

الملخص

إن منصات التواصل الاجتماعي من المساحات المفتوحة التي تسمح لمستعمليها بالتعبير عن آرائهم بكل حرية، مما جعلها من أكثر مواقع الانترنت شعبية واستعمالاً، ومنها موقع تويتر الذي يعد من بين أكثر مواقع التواصل الاجتماعي زيارة، إذ يزداد عدد مستعمليه يوماً بعد يوم، ونظراً لكمية المعلومات، الآراء ووجهات النظر التي تحتويها هذه المواقع، ظهرت أهمية تحليل واستخلاص هذه الآراء والاستفادة منها في مجالات شتى، لتسمح للمستفيدين من هذه المعلومات في اتخاذ القرارات المناسبة وفقاً لنتيجة تحليل النصوص المكتوبة فيها وتصنيفها على وفق تصنيفات معينة. إذ لاقى حقل التنقيب في الآراء وتحليل المشاعر اهتماماً كبيراً من الباحثين لكن معظم الدراسات ركزت على النصوص الإنكليزية. لذلك تم في هذا البحث دراسة النصوص العربية في هذا المجال خصوصاً بعد زيادة الطلب على أدوات تحليل المشاعر للنصوص العربية والمكتوبة بالفصحى والعامية. واعتمد البحث على تقانة تعلم الآلة وباستعمال خوارزمية آلة متجه الدعم Support Vector Machine لتصنيف التغريدات إلى تغريدات ذات بصمات شعورية (عاطفية) إيجابية، أو سلبية، أو محايدة، وذلك لكونها من الخوارزميات الجيدة في تصنيف النصوص بشكل عام. واستنبطت قواعد انتولوجي مكونة من ثلاثة أقسام هي قواعد انتولوجي لتحديد الاسم الذي يدل على وصف، والاسم الذي لا يدل على وصف، والأفعال، وتكوين معجم للكلمات العربية الذي يتضمن خليطاً من كلمات فصيحة وعامية، ليكون قاعدة بيانات في مرحلة المعالجة الأولية لكلمات نص التغريدة، وتحديث ملف الكلمات العربية المستبعدة Arabic Stop Words، إضافة كلمات عربية مستبعدة جديدة وإزالة كلمات أُخر منه، لأنها تؤثر في عملية تحليل المشاعر. إن عملية استخلاص الحروف العربية من الكلمات في نص التغريدة، هي أفضل طريقة في إزالة الحروف غير العربية، الأرقام، العلامات الصوتية، التنقيط والتشكيل المرافقة للنص. وتمت معالجة الكلمة أوليا بإزالة البادئات واللواحق بطريقة دقيقة لتجنب عملية الحذف العشوائي للحروف التي قد تكون من أصل الكلمة وليست مجرد حروف بادئة أو لاحقة، وتضمنت عملية معالجة البادئات واللواحق العامية التي استحدثت في نصوص التغريدات، وإرجاع حروف عربية قد تم حذفها أو تبديلها بحروف أُخر عندما تمت إضافة اللواحق إليها. علاوة على ذلك، تمت معالجة عدد من الأخطاء المطبعية في نص التغريدات، فضلاً عن عملية فصل بعض الكلمات المدمجة التي ظهرت في المعالجة ككلمة واحدة. إن نصوص التغريدات تضمنت بعضاً من اللهجات التي تستعمل في مثل هذه المواقع والتي تمت معالجتها في البحث. فضلاً عن معالجة جزء من سياق الكلام Context الذي يغير البصمة

الشعورية للكلمة، العبارة، أو الجملة، منها معالجة خاصة لعدد من الكلمات التي تسبق بكلمات، ومعالجة المضاف والمضاف إليه في التغريدات ومعالجة العبارة الاسمية في التغريدة، ومعالجة النفي الذي يعد ذا أهمية كبيرة في تغيير البصمة الشعورية، ليتم دمغ الكلمة أو العبارة أو الجملة المنفية بالبصمة الشعورية وفق المعالجات السابقة، ومعالجة علامات # (الهاشتاغات) والوجوه التعبيرية ورموز تعبيرات الوجه لإعطاء البصمة الشعورية للتغريدة بشكل آلي. إن السمات المستعملة هي أربعة أنواع تكون مدخلات إلى خوارزمية آلة متجه الدعم (الأفعال، والأسماء، والأفعال والأسماء والوحدة القواعدية)، مضافاً إليها أدوات النفي التي عُدّت من ضمن السمات إذ تم إدخالها مع الأنواع جميعها السابقة من السمات وتتضمن أدوات النفي الفصحى والعامية. وتم إدخال كلمات التغريدات إلى متجه السمات لخوارزمية آلة دعم المتجه لإنتاج أنموذج تدريب، أُدخلت لاحقاً إلى المصنف مع متجه سمات الاختبار (الذي تم إنتاجه من اقتطاع جزء من التغريدات)، وعبر هذه العملية تم تصنيف التغريدات إلى تغريدات إيجابية، أو سلبية، أو محايدة، وتم مقارنة التصنيفات الناتجة عن الأنواع الأربعة من السمات وتحديد الأفضل منها وفق مقاييس تقييم الأداء الآتية: الصحة، الدقة، الاستدعاء، مقياس–ف، معدل الايجابي الصحيح، ومعدل السلبي الصحيح، إذ أظهرت نتائج المقارنة أن سمات الأفعال هي الأفضل، إذ بلغ معدل الايجابي الصحيح ٦٧.٥٪ ومعدل السلبي الصحيح ٨٧.٥٪ .

University of Mosul
College of Computer Sciences
And Mathematics

Sentiment analysis for Arabic tweets using support vector machine

Rana Zuhair Abdul Ghany Al- Obaidy

M.Sc./Thesis
Computer Sciences

Supervised By

Dr. Ghayda Abdul Aziz Al-Talib
Assistant Professor

2019 A.D **1441 A.H**

Abstract

The platforms of social networking is open area that allow users to express their opinions, reviews and emotions about topics, events, news and products freely. Therefore, these platforms have become the most popular among users such as Twitter that numbers of users are increasing day by day. So, opinion mining and sentiment analysis appears to extract and analyze sentiments of users from their written tweets, blogs and other texts to help them for decisions making. This field has taken attention of the large researchers, but most of the researches focus on text of English language. Therefore, this thesis focus on opinion mining and sentiment analysis in Arabic language because the demand on its has increased. Tweets are usually written using Modern standard Arabic (MSA) and dialectal Arabic (DA) which have a lot of mistakes such as typos that causes many challenges ,in addition to challenges of opinion mining and sentiments analysis. Therefore, this thesis try to solve number of challenges in context and typos. The goal of this thesis is building a system that classify tweets into three classes, positive, negative or neutral, by using SVM algorithm and four types of features. The features are verbs, nouns, verbs plus nouns and unigram that tag with them for each word in tweet by POS tagging. The existing POS tagging only for English words. So, this research focus on building POS tagger for Arabic words which is divided into two steps, first step consists of rules of ontology that were consulted to three sets , first set uses to identify noun that refers to description, second set uses to identify noun that does not refer to description and the third uses to identify the verbs. while second step are using lexicon which had built manually. Accuracy of designed Arabic POS tagger highly depends on accuracy of natural language processing. For example, the preprocessing of words like removal of prefixes and

suffixes, as well as, the way to handle words after the removal. where in some cases, the word needs to return its letters which were part of the original word because each change in the word may change polarity of it and of all of tweet. The file of Arabic stop words have been used to remove stop word from the text but this file had modified by adding new words to it, and had removed some of words from the file because they are important for sentiment analysis. The process of extracting of Arabic letters from word is better way than removing Non Arabic words, numbers, punctuation, symbols, diacritics and formalization because fencing of them is very difficult. Also, this research mange number of typos, as well as, merged words that appears as one word , in addition to the treatment of some words that was written in dialectical Arabic. This research also handling the context of word, phrase, sentence, and particular words, as well as, a special treatment have been done with negation. The polarities of hashtags, emojis had considered as word of tweet. Process of labelling each tweet has done automatically. Finally, training model is generated by training all features vector for one of the features types , then this model have been used with testing data to classify tweets into three classes. A comparison between results of four types of features have been done using the following metrics of Accuracy, precision, recall, F-measure, TP rate and TN rate. Those accuracy measurements shows that choosing features for verbs are the better.