تحليل المشاعر للأحداث الترفيهية

وجدان باجابر

بحث مقدم لنيل درجة الماجستير في العلوم
(نظم المعلومات)

د. فهد صالح العتيبي

كلية الحاسبات وتقنية المعلومات
جامعة الملك عبد العزيز
جدة – المملكة العربية السعودية
محرم1440هـ-سبتمبر2018 م

المستخلص

الهدف الرئيسي من هذه الرسالة هو معالجة مهمة تحليل المشاعر للمدونات العربية. من خلال استخدام نهج التعلم الآلي الخاضع للإشراف. رصد وتحليل المعلومات المنتشرة على تويتر يمكن الحصول على تقدير دقيق لموقف المستخدم ومشاعره تجاه هذه الأحداث. تمت تغطية هذه المهمة على نطاق واسع بلغات متعددة، معظمها باللغة الإنجليزية، في حين أن العمل باللغة العربية محدود. أسباب ذلك هي التحديات النحوية والتركيبية التي تطرحها اللغة العربية وندرة البيانات العربية الموسومة والمتاحة للجمهور. وبالتالي، فإن تطبيق حلول لمعالجة هذه المسالة يتضمن العديد من التحديات، بعضها خلق موارد البيانات النصية المصنفة، وفحص تقنيات التعلم الآلي المختلفة، والتحري عن واستخراج مجموعة من الميزات التي تعمل بشكل أفضل في مجال تحليل المشاعر.

تحقيقا لهذه الغاية، ستتم مراجعة الأسئلة البحثية الأربعة التي تم تناولها في هذه الرسالة:

س١) كيف يمكن إنشاء ذخيرة بيانات نصية تحتوي على آراء عامة تجاه الأحداث الترفيهية في المملكة العربية السعودية؟

س٢) ما هي طريقة التعلم الآلي الأكثر فعالية في تنفيذ تحليل المشاعر تجاه الأحداث الترفيهية في المملكة العربية السعودية؟

س٣) ما هي مجموعة الميزات الدلالية المستخرجة في مجال تحليل المشاعر؟

س٤) كيف تمثل الميزات الدلالية لتحسين أداء النموذج؟

- **نتائج الرسالة الرئيسية**

- **تطوير الموارد:** لقد أنشأنا مجموعة من النصوص العربية التي تم الحصول عليها من تويتر كواحدة من أشهر منصات التدوين المصغر. بسبب بنية البيانات والطبيعة في تويتر، كانت مهمة تحدي تنظيف وتجهيز وحفظ كمية معقولة من البيانات النصية مهمة صعبة. كانت مهمة توسيم او تحشية البيانات واحدة من التحديات الرئيسية. عملية التوسيم مكلفة ومهمة تستغرق وقتًا طويلاً، حيث يعمل مستخدمان عربيان لتوسيم المجموعة. تم إنشاء ٢٧٦٩ تغريده موسمه الى (إيجابية، سلبية، محايدة).

- **استخراج الميزات والخصائص:** قمنا بتنفيذ نموذج خط الأساس الخاص بنا باستخدام خوارزمية MNB. لقد استخدمنا مجموعة من الميزات القائمة على العد، ونماذج

n-gram (unigram, bigram, trigram) and TF-IDF كانت التجربة على مستوى التصنيف (ثنائي ومتعدد). حقق أداء النموذج نسبة ٨٨٪ في التصنيف الثنائي و ٨٠٪ في التصنيف متعدد الطبقات باستخدام نموذج unigram . إحدى نتائج هذه التجربة المبدئية هي أن عدم التوازن الطبقي يؤثر على أداء الخوارزمية حيث أن خوارزميات التعلم تحتاج إلى مجموعة متوازنة من البيانات لكل قطبية لإنتاج نتائج ثابتة. في التجارب المتقدمة، تم أخذ المزيد من استخلاص الميزة مع مراعاة تمثيل الكلمات إلى جانب حقيبة كلمات n-gram models وTF-IDF؛ كان ذلك عبارة عن نمذجة العبارات. توضح التجربة أن نمذجة العبارات جنبًا إلى جنب مع n-gram تؤثر على أداء النموذج حيث تشير تلك الكلمات التي تظهر في كثير من الأحيان معًا إلى "المعنى" الدلالي المختلف الذي يؤثر على قدرة النموذج للتنبؤ بدقة أكثر. قمنا بالبحث في استخدام خصائص تضمين الكلمات لاستخراج الخصائص النحوية والدلالية من الكلمات. قمنا بتطوير نموذجين لتمثيل ناقلات، CBOW و SG وتم تدريب كلا النموذجين على أساس مجموعة مختلفة المعايير. قدم نموذج skip-gram توقعات أفضل للكلمات. ومع ذلك، فإن تضمين الكلمة كخاصية دلالية لا يعزز دقة المصنّفين نظرًا لوجود مجموعة تدريبية صغيرة في متناول اليد. للتغلب على هذه المشكلة، قمنا بزيادة حجم البيانات، ومع ذلك يبقى أداء الخوارزمية ثابتًا. وبالتالي، كان التضمين المتولد على (العبارات الجذعية والمتمثلة) إجراء آخر تم اتخاذه لتعزيز اكتشاف الأنماط على مجموعة بيانات التدريب. تحسين الحل المقترح أداء الخوارزمية على كلا حزم البيانات ومع ذلك، فإن النتائج على مجموعة البيانات الكبيرة أسفرت عن أداء أعلى. أجابت نتائج هذه التجارب على سؤال البحث الثالث.

- **خوارزميات التعلم الآلي:** تم فحص ثلاثة خوارزميات مختلفة وتمت مقارنة نتائجها مع بعضها البعض بناءً على طريقتين أساسيتين للتقييم. الأول هو التحقق من الصحة المتقاطع الذي ينطبق لضمان تدريب النموذج على جميع أجزاء البيانات وحساب متوسط النتائج. والثاني هو اختبار النماذج على مجموعة من البيانات للتأكد من دقة البيانات القائمة على البيانات غير المرئية لضمان قدرة النماذج على التعميم. أداء MNB يتفوق على SVM خوارزمية LRبمتوسط دقة يبلغ ٩١٪ على مجموعة البيانات المحجوبة بينما حقق MNB و SVMنتائج مماثلة على التحقق المتقاطع بدقة ٨٧٪. ومع ذلك، فإن نتائج نماذج الخوارزميات باستخدام السمات الدلالية على مجموعة صغيرة، تعلن

أن SVM يسجل نتائج مماثلة بنسبة ٨٩٪ مثل LR على مجموعة التثبيت. ومن المثير للاهتمام أن أداء LR يظل ثابتًا كما كان من قبل من خلال إضافة خصائص تضمين الكلمات بدقة مماثلة على نتائج التحقق المتقاطع، حيث كانت الدقة ٨٦٪. والمثير للدهشة، أن SVM حقق دقة أعلى قبل استخدام الخصائص الدلالية للكلمة. انخفض أداء النموذج عند تطبيقه على مجموعة كبيرة في التحقق المتقاطع، حيث سجل 85 LR٪ وسجل 84 SVM٪ .ومن المثير للاهتمام، أن أداء الخوارزميتين قد زاد على مستوى الوقف حيث حقق ٩٠٪ من الدقة. في الرد على السؤال الثاني من البحث، كان أداء مصفوفة التعلم الآلي من MNB جيداً بالمقارنة مع الخوارزميات الأخرى التي تم فحصها لمهمة تصنيف المشاعر. بالنسبة لسؤال البحث الرابع، يمكن لتضمين الكلمات التقاط الخصائص الدلالية للكلمات في السياق، ولكنه يتطلب استخدام مجموعة تدريب كبيرة.

- **المساهمة الرئيسية للرسالة:**

- ساهمت الدراسة في تعزيز موارد المشاعر النصية العربية.
- وضعت الدراسة تصنيف المشاعر العربية من خلال تعلم خوارزميات مختلفة لإجراء دراسة مقارنة بشأن الدقة في أداء مهمة التصنيف.
- عرضت الدراسة مجموعة من السمات الإعلامية على النموذج المطور. حيث تم تطبيق خوارزميات مصمّمة مختلفة ومراقبتها وفقًا لمجموعات الخصائص المختلفة.

بحثت الدراسة في تمثيلات الكلمات المختلفة التي يمكن التقاط العلاقة بين الكلمات وكان فحص طرق دمج الكلمات كميزة دلالية في أداء النموذج المتقدم أحد الإسهامات الرئيسية لهذا العمل

Sentiment Analysis for Entertainment Events

By
Wejdan M. Bajaber

A thesis submitted for the requirements of the degree of Master of Science in Information Systems

Supervised By
Dr. Fahd S. Alotaibi

Faculty of Computing and Information Technology
KING ABDULAZIZ UNIVERSITY
JEDDAH-SAUDI ARABIA
Moharam 1440 H – September 2018 G

Conclusion and Feature Work

The main aim of this thesis is to address the task of sentiment analysis of Arabic microblogs. Precisely, analyze the public's opinions regards the entertainment events hosted by the Saudi general entertainment authority through utilizing a supervised Machine Learning (ML) approach to perform Sentiment Analysis (SA). Monitoring and analyzing the information scattered on twitter can obtain an accurate estimation of the user's attitude and feelings toward these events. This task has been covered widely in multiple language, mostly in English, while the work in Arabic language is limited. The reasons for that is the morphological challenges Arabic language poses and the scarcity of publically available annotated Arabic corpora. Consequently, addressing this task posture many challenges, some are the creation of Arabic sentiment resources, examining different machine learning techniques, investigating and extracting a set of features that work the best in the domain of sentiment analysis.

To this end, the four research questions addressed in this thesis will be revisited:

RQ1: How to create a sentimental-based textual corpus containing public opinions toward entertainment events in Saudi Arabia?

RQ2: Which machine learning method is the most efficient in implementing sentiment analysis towards entertainment events in Saudi Arabia?

RQ3: What are the set of extracted informative features in the domain of sentiment analysis?

RQ4: How to represent semantic features to improve the model performance?

In this final chapter, we sum up our main conclusions, findings, and contributions concerning our work toward answering the above questions.

1.1 Main Thesis Results

This thesis addressed the task of Arabic sentiment analysis from different aspects, with goal to contribute to different area.

1. **Resource Development:** Utilizing machine learning algorithms to develop Arabic SA, requires the creation of textual resource (i.e. annotated corpus). Chapter V answered the first research question (RQ1) in which we introduced our methods of constructing a domain-specific Arabic corpus for sentiment analysis. We created an Arabic corpus obtained from Twitter as one of the most popular microblogging platforms targeting the trend hashtags of Saudi entertainment events. Due to Twitter data structure and nature, it was a challenging task to clean, prepare and keep a reasonable amount of textual data. The annotation task was one of the main challenges for this chapter. The annotation process is expensive and time consuming task, where two Arab users are employed to label the corpus. At the end of the chapter, a corpus contained 2769 tweets labeled to three categories (positive, negative, and neutral) was constructed.
2. **Features Extraction:** Chapter VI represents the baseline model where it served as a benchmark for the upcoming machine learning classifiers experiments. We implemented our baseline model using Multinomial NB algorithm. We utilized a set of count-based feature, n-gram models (unigram, bigram, and trigram) and TF-IDF. The experiment was on tow level of classification (binary and multiclass). The model performance achieved 88% on binary classification and 80% on multi-class classification using the unigram

model. One of the findings of this chapter is that class imbalance affects the classifier performance as the learning algorithms need a balanced corpus to produce steady results. In Chapter VII, further feature extraction considering words representation was taken beside the bag of words n-gram models and TF-IDF; that was phrase modeling. The experiment demonstrates that Phrase modeling combined with n-gram affects the model performance as those words who frequently appears together indicate different semantic "meaning" which influence the model ability for more accurate prediction. In Chapter VIII, we investigated the use of word embedding features to extract syntactic and semantic properties out of the words. We developed two models for vector representations, CBOW and SG and both models were trained based on a different set of parameters. The skip-gram model provided better word predictions. However, word embedding as a semantic feature did not enhance the classifier's accuracy, due to the small training corpus in hand. To overcome this problem, we increased the data size and, yet the classifier's performance remains constant. Thus, generated embedding on (stemmed and tokenized phrases) was another action taken to **strength the discovery of patterns over the training dataset**. The proposed solution improved the performance of the classifiers on both corpora. However, the results on the large corpus yielded to higher performance. The findings of these three chapters answered the research question (RQ3).

3. **Machine Learning Classifiers:** Three different ML classifiers were examined and their results were compared with each other based on two main evaluation methods. The first is cross-validation that applied to ensure training the model on all portions of data and calculates the average results. The second was testing the models on a holdout set of data

to validate the accuracy based on unseen data to assure the models' ability for generalization. The experiments of Chapter VII and Chapter VII outcomes declare that MNB outperforms SVM and logistic regression classifiers with average accuracy of 91% on the holdout dataset while MNB and SVM achieved similar results on the cross validation with 87% accuracy. However, the results of classification models using the semantic features on the small corpus, declares that SVM score similar results with 89% as the LR on the holdout set. Interestingly, the LR performance remains constant as before adding word embedding features and similarly dose on the cross-validation results, where the accuracy was 86%. Surprisingly, SVM achieved higher accuracy before utilizing the word2ved semantic features. The model performance decreased when applied on the large corpus in the cross validation, where LR scored 85% and SVM scored 84%. Interestingly, the classifiers performance increased on the holdout set achieving 90% accuracy. In respond to (RQ2), MNB machine learning classifier performed well in comparison with the other examined classifiers for the sentiment classification task. Regarding to (RQ4), word embedding can capture semantic properties of the words in context, yet it requires the use of large training corpus.

1.2 Main Contribution

In summary, for the above discussed thesis findings, the present work has made the following contribution:

1. The study contributed toward enhancing the Arabic language sentiment resources. We Built an annotated corpus from domain-specific microblogging (i.e., Arabic tweets) related to Saudi Arabia entertainment events in which we will make publically available.

2. The study developed Arabic sentiment classification by learning three different classifiers (SVC, MNB, and LR), in which to perform a comparative study concerning the accuracy in performing classification task.
3. The study presented a set of informative features on the developed model. Involving the BOW, n-grams, phrase modeling and TF-IDF, in which different supervised classifiers algorithms were applied and compared according to different feature sets.
4. The study investigated different word representations that can capture the relation between words (syntactically and semantically). An examination of word embedding methods as a semantic feature on the developed model's performance was one of the main contributions of this work.

1.3 Limitations

Upon completing this study, different limitations of the conducted work will be point out in this section.

1. Corpus size: The small dataset was one of the limitations of this work. Constructing a large Arabic sentiment corpus requires intensive time for all crawling, cleaning, and annotation process. Due to the time constraints to finish this work and meet deadlines, we were able to collect data over 4 months crawling period. Indeed, during the time period over 80K tweets were collected. However, the context of those data was full of noise and irrelevant data. Thus, the constructed dataset contained only 2769 tweets which represent the sentimental corpus of this work.
2. Computational resource: The size of the corpus was a major factor concerning the word embeddings features extraction. As it requires a huge amount of data for the model to

deduce semantic relations between the words. However, using a pertained corpus might assess to overcome this issue. Nevertheless, the lack of powerful computational resource was one of the factor that prevent us of using pertained corpus.

3. Features extraction: The work did not investigate the syntactical features of the language such as parts of speech (POS) and base phrase chunk (BPC).
4. Negation and sarcasm: in the scope of our work we did not conceder handling negation. As we mentioned in the literature review the effect of negation on the overall semantics of the text. Also, dealing with sarcasm where the context of the text conveys opposite meaning to the words presented which accordingly affect the overall sentiment of a given text.

1.4 Future Work

There are many different directions to carry on based on this work. Arabic sentiment analysis is still an emerging topic in comparison with the conducted work in other languages. Researchers could do some of the following path toward evolving the research area in Arabic SA.

1. Developing more Arabic annotated corpora to overcome the limitation of the publically available corpus. The existence of Arabic sentiment recourses, will encourage the researchers to continue in advancing the research area.
2. As for extracting semantics features, using a pertained corpus as for the model to deduce better relations between the words.
3. As for a high inflectional and morphologically complex language as Arabic; Language dependent features such as "morphological-based, syntactic-based, and dependency-

based features" caught less attention than "word gram and microblogging features" and needs to be investigated and evaluated.

4. Handling negation is one of the major challenges in Arabic sentiment analysis, though few works have tried to address this issue. Negation in Arabic needs to be investigated more, especially with the use of machine learning algorithms; as most of the trials used opinionated words count and lexicon based methods.