استخدام تويتر لتحليل مبادرات الشركات الناشئة: دراسة حالة تمكين المراءة للعمل في شركة كريم.

بشاير علي محمد العتيبي

بحث مقدم لنيل درجة الماجستير في العلوم
(نظم المعلومات الحاسوبية)

الدكتور محمد احتشام أسلم

المستخلص

في هذا العصر، تمثل منصة تويتر مصدراً هامًا للآراء العامة التي يمكن الوصول إليها بسهولة. في سياق مجال الأعمال، سيمكّن الاستخدام الصحيح لتحليل بيانات تويتر الشركات الناشئة من تتبع تغييرات الأعمال والتحسينات بسهولة للحفاظ على وجودها. علاوة على ذلك، سيدعم هذا التحليل الشركات الناشئة في تحليل البيئة التنافسية، تجنب أسباب الفشل، وتحسين مشاركة العملاء، ودعم عمليات صنع القرار، وقياس استجابة الجمهور لتلبية احتياجات المستهلكين بشكل أفضل.
يعاني الوضع البحثي الحالي الخاص بتحليل نشاطات الشركات الناشئة باستخدام بيانات تويتر من عدة قيود. بشكل عام، هناك نقص في وجود إطار تحليلي والذي يستخدم مجموعة بيانات تويتر بشكل أفضل من خلال الجمع بين طرق التحليل المختلفة. من حيث الشركات الناشئة، هناك نقص في البحوث التي تدرس دور تويتر لقياس أداء نشاط معين (على سبيل المثال المبادرات والحملات التسويقية)، وكانت جميع الأعمال السابقة على نطاق الشركات الناشئة الأوروبية والأمريكية (أي لغة البيانات هي اللغة الإنجليزية أو أوروبا). كان الهدف من هذه الأطروحة هو سد الثغرات البحثية السابقة من خلال اقتراح إطار عمل قائم على تحليل بيانات تويتر (SIRA)، لدعم الشركات الناشئة في قياس استجابة الجمهور فيما يتعلق بمبادرتهم. تم تطوير إطار SIRA استنادًا إلى ثلاث تقنيات؛ تصنيف النص (text classification) وتحليل المشاعر (sentiment analysis) والتحليل الإحصائي. (statistical analysis)
سيكون هذا الإطار قادرًا على تحديد استجابة الجمهور فيما يتعلق بأنشطة تويتر ورضا العملاء والانتشار الزمني. تم التحقق من صحة الإطار المقترح من خلال دراسة حالة شركة كريم باللغة العربية لإثراء الفجوة البحثية في التنقيب عن النص العربي.
تم إجراء التجربة بناءً على مجموعة بيانات مختلطة باللغتين العربية والإنجليزية، تتألف من ٣،٠٧٤ تغريدة تم تصنيفها يدويًا. مجموعة البيانات العربية تتكون من لهجات عربيه وبعض الأشكال القياسية الحديثة. لذلك، كانت هناك حاجة للتغلب على هذا التحدي وتحسين أداء تصنيف اللغة العربية من خلال تحليل مقارن لمجموعة مختلفة من تقنيات المعالجة المسبقة. أسفرت التجربة عن النتائج التالية: في كلا من نموذجي التصنيف لمجموعة البيانات العربية، حقق مصنف CNB مقياس F1 أعلى. بينما يبدو أن (text cleaning and normalization) يزيد بشكل كبير من أداء التصنيف الثنائي، في حين أظهرت نتائج نموذج تصنيف المشاعر مقياس F1 أعلى عندما تم استخدام (text cleaning and (normalization with ISRI stemmer.
من ناحية أخرى، في كلا من نموذجي تصنيف مجموعة البيانات الإنجليزية، حقق المصنف NN مقياس F1 أعلى. بشكل عام، مجموعة البيانات صغيرة نسبيًا، وهذا ما يفسر لماذا قيم قياس الأداء ليست عالية جدًا في نموذجي التصنيف لكل لغة. كذلك، يبدو أن أداء المصنف يعتمد على مجموعة البيانات ويعتمد على الغرض من تصنيف النص. استنادا إلى نتيجة مجموعة البيانات المصنفة تم إجراء العديد من التحليلات الإحصائية وتقديمها.
في الواقع، عادة ما يعبر الناس عن آرائهم من خلال تويتر فيما يتعلق بموضوعات وقضايا مختلفة، لا تقتصر على المنتجات أو الخدمات. وبالتالي، فإن الإطار SIRA قابل للتطبيق في أي مجال وغرض كإطار تحليلي قائم على بيانات تويتر.

Using Twitter to Analyze Initiatives of a Startup Company: The Case of Empowering Women in CAREEM.

By

Bashayer Ali Mohammad Al-otaibi

Supervised by

Dr. Muhammad Ahtisham Aslam

Abstract

Public opinions are significant to almost any organization, companies, and governments, in which such entities are aware of the significance of utilizing the unstructured data in the social networks, and it has become a growing research area lately. Twitter, as a microblogging platform, represents a significant source of public opinions that is easily accessible. In the business domain, the previous research efforts in analyzing startups activities through Twitter analysis are generally limited, especially for the Arabic language. In the Twitter analysis filed, there is a lack of a twitter-based analytics framework that combines different analysis methods to utilize the Twitter dataset better.

This thesis study aims to fill the literature research gaps by proposing a Twitter analytics-based framework called Startup Initiatives Response Analysis (SIRA). SIRA assesses the performance of an initiative taken by startup using text classification, sentiment analysis, and statistical analysis techniques. The proposed framework is validated empirically through a case study of Arabic startup (i.e., Careem), regarding the initiative of empowering women to work in the Careem. The study experiment was carried out based on using supervised machine learning (SML) in building the subject and sentiment classification models. As well, the classification models are evaluated through a comparative analysis in terms of examining a variety of machine learning (ML) classifiers, and various levels of preprocessing techniques to improve the performance of Arabic text mining.

The study experiment yielded the following results: for both two Arabic classification models, Complement Naïve Bayes (CNB) achieved a higher F1 measure with applying text cleaning and normalization as text preprocessing techniques. While the Neural Network (NN) classifier achieved the highest F1 measure for the binary classification model of the English dataset, and the Random Forest (RF) classifier outperformed other classifiers for the sentiment classification model of the English dataset.

In contrast, based on the classified dataset, several statistical analyses were conducted and presented (e.g., Tweets Reply frequency, the temporal distribution of Tweets). The experiment results analysis confirms the effectiveness of such a framework in delivering valuable insights regarding the public responsiveness, based on a comprehensive qualitative and quantitative analysis of the Twitter dataset.

The proposed framework (SIRA), is applicable as a Twitter-based analytics framework in any domain and for any purpose. One of the future work recommendations is validating the proposed framework with other experiments of different datasets.