وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

إيجاد الموضوع الرائج في تغريدات تويتر في الزمن الحقيقي

رسالة مقدمة
إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب

من قبل

رقيه خليل إبراهيم كشمولة

بإشراف
أ.م. د. غيداء عبد العزيز الطالب

١٤٤٤ هـ
٢٠٢٣ م

الخلاصة

تنتشر دائما العديد من القضايا والمواضيع في وسائل التواصل الاجتماعي بشكل كبير، مما يثير الرأي العام وتؤثر نتائجها بشكل مباشر على المجتمع سواء كان التأثير كبير أَو ضئيل، هذا ما يُسمَّى بالموضوع الرائج. قد يستمر انتشار ورواج موضوع أكثر من غيره لعدة أَيام أَو يستمر فقط لعدة ساعات. الكثير من التجار وأَصحاب المعامل والتسويقيين وحتى السياسيين يقومون باستغلال هذه المواضيع مما يخدم أَعمالهم؛ إذ إنَّها تؤثر في تفكير المجتمعات بشكل كبير؛ إذ قد تزيد نسب مبيعات شركات بالمليارات أَو قد تخسر الشركة مبالغ طائلة بسبب موضوع رائج سيء تجاه المنتج أَو الشركة أَو أَي شيء يلامسها. لذلك وجد أَنَّ من المهم اكتشاف ما هي المواضيع الأَكثر رواجًا لما لها من أهمية سياسيًا واقتصاديًا وإعلاميًا وغير ذلك.

تم تحديد النتائج لمنطقة جغرافية معينة وهي منطقة الوطن العربي وشمال أَفريقيا الذي تم الحصول عليه من خرائط Google؛ إذ زود بها النظام المقترح بمرشح الموقع، أَيضًا تم تحديد اللغات التي سيتم معالجتها ضمن النظام المقترح وهي اللغتين العربية والإنكليزيَّة بالمرشح الخاص بها لكيلا تتشتت النتائج عندما تكون بلغات ومناطق جغرافية عشوائية. تم اختيار تويتر مصدرًا للبيانات كونه منصة مفتوحة المصدر ويمثل آراء المجتمع بإيجاز ويُعَدُّ من أكثر المواقع تداولًا في العالم. تم سحب التغريدات التي تحوي ضمن كلماتها رمز التصنيف "#", كما تم سحب اسم ناشر التغريدة ووقت وتاريخ النشر من الناس المشتركين بشكل عشوائي في تويتر بالوقت الحقيقي (Real time) عن طريق الحصول على حساب تويتر أولا ثم الحصول على حساب مطوري تويتر، عن طريق بوابة أَو منصة واجهة برمجة تطبيقات تويتر Application Programming Interface (API) التي تتيح سحب البيانات من تويتر بتفاصيل معينة يمكن تحديدها، وبصلاحيات يحددها تويتر حسب حاجة المطور وبحسب وجود المعلومة وكميتها. تم تنظيف البيانات (التغريدات) أولًا، ثم معالجتها بطرائق معالجة اللغات الطبيعية، لغرض تنظيم البيانات وتوحيد صيغتها من أَجل إدخالها للنظام المقترح، لكي تعطي دقة أَعلى وسرعة في التنفيذ بتقليل العبء على النظام المقترح. تم تقطيع النص إلى كلمات أَو مقاطع، ثم إزالة الزوائد من النص مثل الرموز والوجوه التعبيرية وعناوين URL وغيرها، ثم إرجاع الكلمات إلى جذورها بالخوارزميات والفهارس الخاصة باللغات العربية والإنكليزيَّة كل منهما على حدا. عند الوصول إلى صيغة واحدة للبيانات تم تطبيق إحدى تقنيات معالجة اللغات الطبيعية وهي تقنية تردد المصطلح _ تردد المستند العكسي(Term Frequency _ Inverse Document Frequency (TF_IDF)) لإعطاء قيم عددية تمثل اهمية كل كلمة أَو مقطع ضمن التغريدات في ملف البيانات ككل، وأَخيرًا تم استخدام

التوجيه (Vectorization) لنتائج خوارزمية TF_ IDF، لترتيب المواضيع من الأَكثر أَهمية (الأَكثر رواجًا) إِلى الأَقل أَهميَّة.

بلغت دقة النتائج من النظام المقترح في الأَيَّام العادية 60% للتغريدات باللغتين الإِنكليزيَّة والعربية وبعدد 1000 تغريدة، بينما وصلت 100% في حالة وجود حدث كبير في ذلك اليوم مثل (كأس العالم) لنفس العدد من التغريدات، بالمقارنة مع نتائج تويتر حول الموضوع الرائج في المُدَّة الزمنية نفسها.

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and Mathematics
Department of Computer Science

Real-time Trends Finding in Twitter Tweets

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Computer Science**

By

Roqaya Khalil Ibrahim Kashmola

Supervised by

Assistant Professor

Dr. Ghayda Abdul Aziz Al-Talib

2023 A.D. **1444 A.H.**

Abstract:

There are always many issues and topics on social media, which raise public opinion and directly affect the results of society, whether the impact is large or small, this is what is called the trending topic. The spread and popularity of a topic may continue more than others for several days or only for several hours. Many merchants, factory owners, marketers, and even politicians exploit these topics, which serves their businesses. As it greatly affects the thinking of societies; As the percentage of sales of companies may increase by billions, or the company may lose huge amounts of money due to a trending topic that is bad about the product, the company, or anything that touches it. Therefore, found important to discover what are the trending topics because of their political, economic, media, and other importance.

The results were determined for a specific geographical region, which is the region of the Arab world and North Africa, which was obtained from Google Maps; As the proposed system was provided with a site filter, the languages that will be processed within the proposed system, which are Arabic and English, were also specified with their own filter so that the results would not be dispersed when they are in random languages and geographical areas. Twitter was chosen as the source of the data because it is an open-source platform, it succinctly represents the opinions of the community, and it is one of the most popular websites in the world. Tweets containing the hashtag "#" were withdrawn, and the name of the publisher of the tweet and the time and date of publication were withdrawn from random people participating in Twitter in real time by obtaining a Twitter account first and then obtaining a Twitter developer account, by Through a portal or platform of the Twitter Application Programming Interface (API) that allows data to be pulled from Twitter with specific identifiable details, and with permissions determined by Twitter according to the need of the developer and according to the availability and quantity of information. The data (tweets) were cleaned first, then processed by natural language processing techniques, in order to organize the data and unify its format in order to enter it into the proposed system, in order to give higher accuracy and speed of implementation by reducing the burden on the proposed system. The text was cut into terms or phrases, then the excesses were removed from the text such as symbols, emoticons, URLs, etc., then the terms were returned to their roots using algorithms and indexes for the Arabic and English languages separately. When reaching a unified format for the data, one of the natural language processing techniques, Term Frequency _ Inverse Document Frequency (TF_IDF), was applied to give numerical values that represent the importance of each term or phrase within the tweets in the data file as a whole. Finally, Vectorization was used on the results of the TF_IDF algorithm, to arrange topics from most important (trending topics) to least important.