Using Big Data Tools to Analyze Tweets Related to Hajj Sentimentally

Gassan Bati

Umm Al-Qura University

Abstract

Millions of Muslims gather annually at The Sacred Places in Makkah to perform Hajj. Many of them use Twitter to talk about their experience as well as communicate with their relatives and friends. Sentiment analysis helps to measure people's opinion about a matter which facilitates taking decisions. This paper uses Spring XD, Hadoop, Hive, and Microsoft Excel to collect, refine, and visualize tweets. There is no doubt that Hajj involves many data, so this paper presents the importance of using big data tools effectively in Hajj to enhance the services provided by different organizations.

Introduction

We live currently in the era of tremendous amount of available data. Big data is a term that is widely spread and used these days. Also, it has many definitions. One definition that is relevant to the proposed use of big data in this paper is "a new attitude by businesses, non-profits, government agencies, and individuals that combining data from multiple sources could lead to better decisions" [1] [2]. According to The Saudi Arabian Central Department of Statistics and Information, 2,085,238 Muslims coming from various countries all over the world have performed Hajj this year [3].

Twitter is one of the most popular social networks worldwide [4]. It is also widespread in Saudi Arabia as Saudi has the highest number of active Twitter users in the Arab World (2.4 million active users) [5] [6]. This paper uses Spring XD, Hadoop, Hive, and Microsoft Excel as a proof of concept to collect, refine, and visualize tweets. Furthermore, there is no doubt that Hajj involves many forms and types of big data, as the previous numbers show, so this paper presents the importance of using big data tools effectively in Hajj to enhance the services provided by businesses, non-profits, government agencies, and individuals.

Related Work

Big data tools and techniques are used in many fields. For instance, the White House declared a nationwide initiative related to big data which includes six federal departments and agencies pledging $200 million to research projects in big data arenas [7]. Also, similar initiatives have been established in other countries like UK [8]. In Saudi Arabia, the Saudi mobile carrier Mobily has recently established its own infrastructure for big data with the help of Teradata and Hortonworks. It aims to provide enhanced services that are designed to target individuals [9]. The research community has some good studies for Twitter sentiment analysis and big data like [10], [11], and [12].

To IBM, big data means big return on investment. Furthermore, it claims that 20٪ decrease in patient mortality could be achieved by analyzing streaming patient data, 92٪ decrease in processing time could be reached by analyzing networking and call data, and 99٪ enhanced accuracy in terms of placing new resources of power generation could be accomplished by analyzing 2.8 petabytes of untapped data. It provides other examples and use cases for big data in the fields of automotive, banking, consumer products, energy and utilities, government, healthcare, insurance, oil and gas, retail, telecommunications, and travel and transportation [13]. The UN published in its Global Pulse that big data is a key change for development in 21st century if it is utilized [14].

Implementation and Results

For majority of the subsections of this part of the paper, Hortonworks sandbox 2.1 is used. Hortonworks sandbox is portable Hadoop environment which is equipped with many Hadoop tutorials and runs in virtual machines. Also, it contains many packages as shown in figure one [15]. Similarly, a Cloudera quick start virtual machine could be used as an alternative [16].

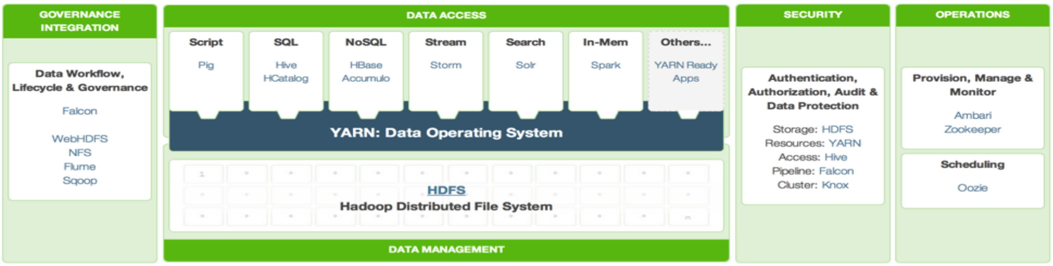


Figure One: Apache Hadoop Environment [17].

Collecting Tweets

There are many ways to collect tweets (data) from Twitter. One good way is done using Spring XD [18]. Spring XD (extreme data) is a "unified, distributed, and extensible system for data ingestion, real time analytics, batch processing, and data export". The goal of Spring XD is to facilitate developing big data applications [19]. In this paper, the set of data (tweets) that are used were collected by Hortonworks for the lunch of the movie Iron Man 3. The tweets were collected using Flume and are provided freely online in the website of Hortonworks [20].

Refining Tweets

In order for the tweets to be refined, they have to be uploaded to the sandbox. Then, a hive script is used to refine the tweets. This hive script is responsible for converting the raw tweets into a tabular format, scoring the sentiment of each tweet by comparing the number of positive and negative words, assigning a neutral, positive, or negative sentiment rate to each tweet, and finally creating a new table which contains the sentiment rate for each Tweet [20].

Visualizing Tweets

In this part, Microsoft Excel Professional Plus 2013 is used to access the refined sentiment data (tweets) generated in the previous step to visualize them depending on their location in the world. Figure two shows a map that presents the refined tweets using orange for positive, blue for negative, and red for neutral sentiment [20].

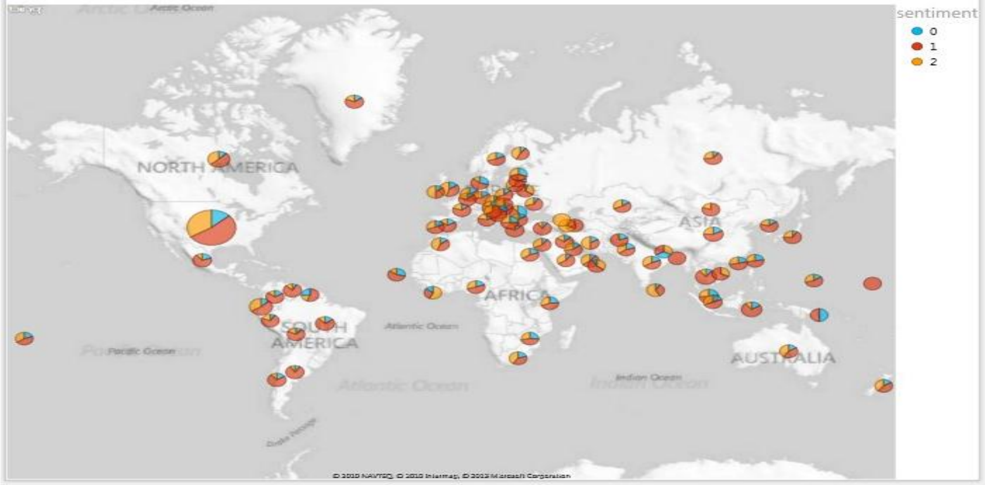
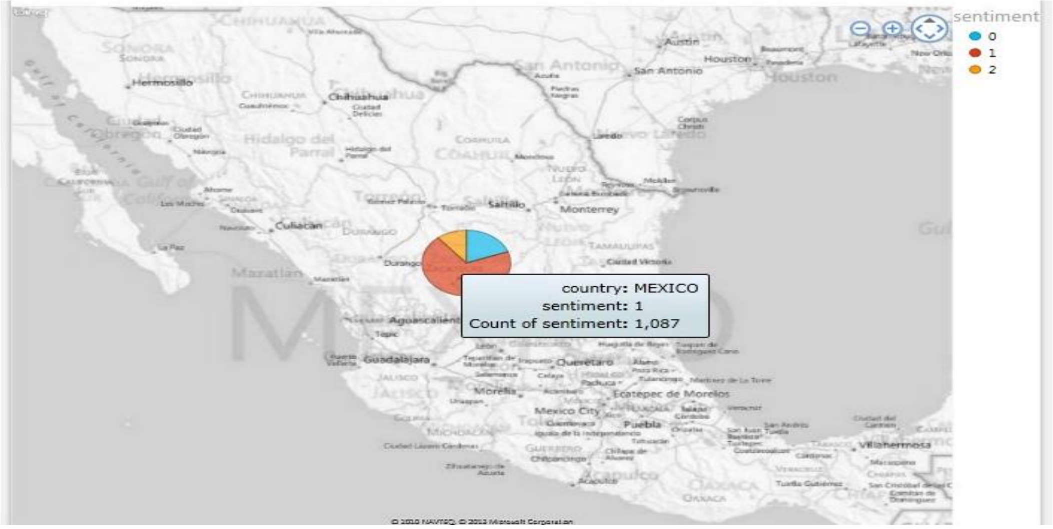


Figure Two: Presenting refined tweets worldwide [20].

Zooming in to any particular country will give more details that help in the analysis. For example, figure three presents the sentiment analysis for Mexico [20].



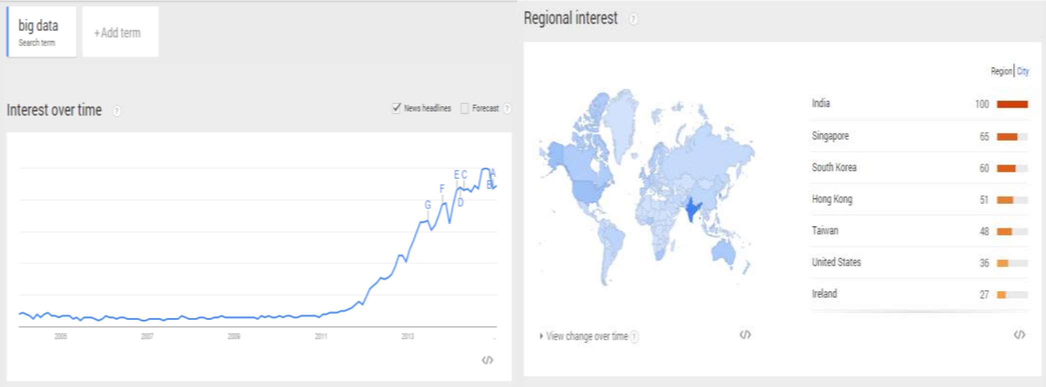
country: MEXICO
sentiment: 1
Count of sentiment: 1,087

Figure Three: Sentiment analysis for Mexico [20].

Suggested Uses Related to Hajj

Now and after presenting this example of analyzing tweets sentimentally, the paper is going to suggest some useful uses of big data tools and techniques that may be utilized in Hajj to improve services. For example, Al Mashaaer Al Mugaddassah Metro (Arabic: قطار المشاعر المقدسة) could generate a Twitter hashtag that the officials of the metro monitor and analyze as one factor out of many they have already to maintain the quality of the provided services. This hashtag could be publicized quickly and easily using text messages, billboards in airports and streets, in online and TV ads, the metro tickets and cards, and many more. Then, pilgrims would use it to tweet about the metro. Other agencies could replicate the same explained idea. The Saudi Project for Utilization of Hajj Meat could benefit from big data as well by analyzing the website's server log data to increase the revenue and the security of the website [21]. Other organizations might find the same idea very beneficial. It is clear after taking a look at the Descriptive and Cumulative Index of the Studies, Reports, and Researches done by the Custodian of the Two Holy Mosques Institute of Hajj and Umrah Research that many researchers are interested in using sensors in Hajj [22]. Big data tools could be used effectively to analyze sensors' data as described in [23].

The author would love to collaborate with any innovative and useful ideas that use big data to facilitate Hajj since big data is gaining popularity in many countries all over the world as could be seen in figure four, so are we ready to benefit from it to improve the provided services in Hajj?



India	100
Singapore	65
South Korea	60
Hong Kong	51
Taiwan	48
United States	36
Ireland	27

Figure Four: Trends of the keyword "Big data" from Google [24].

Future Work

Although this paper talks about sentiment analysis for tweets, this choice does not mean that it is the only social networks to analyze. Analyzing many social networks reaches more people who do not necessarily use Twitter and provides more data. The book in [25] covers how to analyze many social networks and email clients.

Arabic sentiment analysis could be added to this work easily as a future work. The author is not going to reinvent the wheel, but rather he may use an Arabic Twitter corpus for subjectivity and sentiment analysis that has been developed by Eshrag Refaee and Verena Rieser [26].

Conclusion

This paper presents a proof of concept for the importance of using big data tools and techniques to analyze tweets related to Hajj sentimentally to improve the provided services. The proposed concept is suitable to implement by various agencies, organizations, and individuals. Also, it could be expanded in many ways and directions.

References

[1] G. Press, "12 Big Data Definitions: What's Yours?," Forbes, 03 09 2013. [Online]. Available: http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/. [Accessed 09 01 2015].

[2] J. Dutcher, "What Is Big Data?," datascience@berkeley, 03 09 2014. [Online]. Available: http://datascience.berkeley.edu/what-is-big-data/. [Accessed 09 01 2015].

[3] Centeral Department of Statistics and Information, "Hajj_1435," 2014. [Online]. Available: http://www.cdsi.gov.sa/pdf/Hajj_1435.pdf. [Accessed 09 01 2015].

[4] Statista, "Global social networks ranked by number of users 2014," Statista, 12 2014. [Online]. Available: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed 09 01 2015].

[5] K. T. AbduRabb, "Saudi Arabia has highest number of active Twitter users in the Arab world," Arab News, 27 06 2014. [Online]. Available: http://www.arabnews.com/news/592901. [Accessed 09 01 2015].

[6] R. Mourtada, F. Salem and S. Alshaer, "Citizen Engagement and Public Services in the Arab World: The Potential of Social Media," MBRSG's Governance and Innovation Program, 06 2014. [Online]. Available: http://www.mbrsg.ae/getattachment/e9ea2ac8-13dd-4cd7-9104-b8f1f405cab3/Citizen-Engagement-and-Public-Services-in-the-Arab.aspx. [Accessed 09 01 2015].

[7] Office of Science and Technology Policy, "OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS," 29 03 2012. [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf. [Accessed 11 01 2015].

[8] BBC, "Alan Turing Institute to be set up to research big data," BBC, 19 03 2014. [Online]. Available: http://www.bbc.com/news/technology-26651179. [Accessed 11 01 2015].

[9] Al Madina Newspaper, "موبايلي تنجح في تنفيذ البنية التحتية لمشروع البيانات الضخمة," Al Madina Newspaper, 01 01 2015. [Online]. Available: http://www.al-madina.com/node/579263. [Accessed 11 01 2015].

[10] A. H. A. Rahnama, "Distributed real-time sentiment analysis for big data social streams," in *International Conference on Control, Decision and Information Technologies (CoDIT)*, Metz, France, 2014.

[11] A. Minanovic, H. Gabelica and Z. Krstic, "Big data and sentiment analysis using KNIME: Online

reviews vs. social media," in *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2014.

[12] A. Lima and L. de Castro, "Automatic sentiment analysis of Twitter messages," in *Fourth International Conference on Computational Aspects of Social Networks (CASoN)*, Sao Carlos, 2012.

[13] IBM, "Big data in action," IBM, [Online]. Available: http://www-01.ibm.com/software/data/bigdata/industry.html. [Accessed 11 01 2015].

[14] UN Global Pulse, "Big Data for Development - UN Global Pulse June 2012," UN, 06 2012. [Online]. Available: http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf. [Accessed 11 01 2015].

[15] Hortonworks, "Hortonworks Sandbox," Hortonworks, [Online]. Available: http://hortonworks.com/products/hortonworks-sandbox/. [Accessed 12 01 2015].

[16] Cloudera, "QuickStart VMs for CDH 5.3.x," Cloudera, [Online]. Available: http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms/cdh-5-3-x.html. [Accessed 12 01 2015].

[17] Hortonworks, "Introducing Apache Hadoop to Developers," Hortonworks, [Online]. Available: http://hortonworks.com/hadoop-tutorial/introducing-apache-hadoop-developers/. [Accessed 12 01 2015].

[18] mehzer, "Hadoop Tutorial: Using Spring XD to stream Tweets to Hadoop for Sentiment Analysis," Hortonworks, [Online]. Available: http://hortonworks.com/hadoop-tutorial/using-spring-xd-to-stream-tweets-to-hadoop-for-sentiment-analysis/. [Accessed 11 01 2015].

[19] Pivotal Software, Inc, "Spring XD," Pivotal Software, Inc, [Online]. Available: http://projects.spring.io/spring-xd/. [Accessed 11 01 2015].

[20] Hortonworks, "Hadoop Tutorial: How To Refine and Visualize Sentiment Data," Hortonworks, [Online]. Available: http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-sentiment-data/. [Accessed 11 01 2015].

[21] Hortonworks, "How to Refine and Visualize Server Log Data," Hortonworks, [Online]. Available: http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-server-log-data/. [Accessed 12 01 2015].

[22] M. S. O. Nojoum, "http://hajj.edu.sa/Indexing/Indexing.pdf," 06 2008. [Online]. Available: http://hajj.edu.sa/Indexing/Indexing.pdf. [Accessed 12 01 2015].

[23] Hortonworks, "How To Analyze Machine and Sensor Data," Hortonworks, [Online]. Available: http://hortonworks.com/hadoop-tutorial/how-to-analyze-machine-and-sensor-data/. [Accessed 12 01 2015].

[24] Google, "Google Trends - Web Search intrest: big data - Worldwide, 2004 - present," Google, 11 01 2015. [Online]. Available: http://www.google.com/trends/explore#q=big%20data. [Accessed 11 01 2015].

[25] M. A. Russell, Mining the Social Web, Sebastopol, CA: O'Reilly Media, Inc., 2014.

[26] E. Refaee and V. Rieser, "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis," in *The International Conference on Language Resources and Evaluation*, Reykjavik (Iceland), 2014.