Fabricated information on social media: A Case Study of Twitter

Ali Ebrahimi[1], Maizatul Akmar Ismail[1], Salinah Jaafar[2], Suad Awab[3], Rohana Mahmud[1], Abdullah Ghani[1], Ishak Suliaman[4], Amirrudin Kamsin[1], Norisma Idris[1]

[1]Faculty of Computer Science and Information Technology;
[2]Academy Of Malay studies
[3]Faculty of Language and Linguistics
[4]Academy of Islamic Studies
University of Malaya,
50603 Kuala Lumpur, Malaysia.
maizatul@um.edu.my

Abstract: By increasing the accessibility to social media, such as Twitter, Facebook and Google+, from various devices it is a challenge to recognize fabricated or false information. This misinformation might be created by the wrong interpretation or translation of a subject, which might affect its core significance. Recent research shows that most people are not aware of the distribution of fabricated and false information on social media, a problem that is also evident in respect of information pertaining to Islam. Consequently, there are various people who are sharing false interpretations and wrong translations of Islamic hadiths on social media. In this work, we shall be developing a system that will be able to extract recent tweets about Islamic hadiths from Twitter and investigate their authentication by comparing them with a reliable database. Therefore, we will develop a database system that will be able to store all the authenticated Islamic hadiths with their Malay language translation. This system will also be able to extract recent tweets about Islamic hadiths from Twitter via the Twitter API functions. Additionally, from a comparison of the extracted tweets and al-hadith stored on our database, it will be possible to analyze the percentage of fabricated tweets about al-hadiths in the region of Malaysia. The development of this system will be completed when all the data are stored appropriately. The extraction of relevant tweets from Twitter will be achieved, and, simultaneously, the authenticity of tweets will be investigated.

***Keywords*:** Data Extraction, Data Analysis, Online Social Media, Twitter, Fabricated Information

1. Introduction

In the past few years, online social media (OSM), which encourages people to generate a particular network to share information, opinions, and thoughts in the form of text, photos and videos, has emerged as one the most popular phenomenon on the Web. This scenario is an advantage of Web 2.0, which shifted users from being passive consumers of information to data producers. Thus, OSM platforms provide novel and unique research opportunities; for instance, the analysis of a large scale of user interaction provides a chance to answer questions concerning how humans indulge in friendships and how these friendships grow over time [1]. In addition, human activity and shared opinion in OSM may also benefit scientists and enable them to understand human behavior and identify groups of users that share information on OSM without necessary reviewing a certain topic. Therefore, in this study, we are planning to use social media to investigate an OSM in order to determine the number of users who share fabricated information.

Web data extraction systems have become powerful technologies that are used extensively for applications in the analysis process of web documents in full text format for a wide range of usage in business intelligence [2-5]. Such uses include medical [6, 7] and company management, to understand human behavior [8-10], bio-informatics [11], tourism and travel [12], affectivity of media on politics [13] and crawling of social web platforms [14, 15]. However, there is insufficient research about fabricated data and the information that is shared on OSM platforms.

Several OSMs empower their users to communicate with their network via a variety of functions. A brief study on popular social media – Facebook[16], Twitter[17] and LinkedIn [18] – shows that the most common function among them is the Share function which is called “Tweet” on Twitter, “Post” on Facebook, and “Share” on LinkedIn. This function allows users to share their opinion or information via different formats like texts, images, videos and sounds with additional abilities like tagging a contact or attaching location and emoticons to their posts. This function brings a huge amount of public data that has the potential to deliver valuable information to the scientific community, marketing agencies, NGOs and other organizations interested in people’s behavior and their points of view [19].

As the number of registered users on OSMs is rapidly growing, and, at the same time, the size of content shared is rising, there is a great opportunity for scientists to extract valuable information about human behavior and social interactions. Twitter is an OSM that can be utilized as a tool for research. Currently it has over 284 million monthly active users that generate 500 million tweets per day [20]. The advantage of Twitter in comparison to other social media is that Twitter is an open platform, which allows researchers to access tweets by using the Twitter application programming interface (API) [21]. This presents three APIs including SEARCH API, STREAM API and REST API. These functions offer an opportunity for researchers to access a large quantity of data [22].

In this project, our main objective is to develop a system that will be able to extract recent tweets from the Twitter website regarding Islamic hadiths. This information will then be compared to a database that stores all Islamic hadiths in the

Malaysian language [23]. Moreover, these compared tweets will be analyzed to study the percentage of fabricated information that people are tweeting on the Twitter website.

The Islamic world is growing at a faster rate than ever before, which has led to various Islamic sects and perspectives. Simultaneously, the rapid rise of using social media for sharing information on the World Wide Web has caused an increase in the amount of fabricated data concerning the translation and interpretation of Islamic hadiths. As discussed earlier, some OSMs allow developers to gather public information. For instance, Twitter presents API functions that enable developers to extract recent tweets for different aspects.

Nowadays, identifying the accuracy of Islamic hadiths and the interpretation and authenticity thereof is a problem that Muslims face, and Muslim scholars are consistently working hard to impede the spread of false, inappropriate transliterations and interpretations. However, it is difficult for them to prevent the spread of these problems in cyberspace, especially on social media. Therefore, developing an application that might be able to gather information from social media can provide a solution for the above-mentioned problems. Such an application will create a base on the Twitter API enabling us to extract recent tweets from the Twitter website.

The crux of this project is to develop a system to extract the recent tweets (specified tweets) from Twitter and store them in the form of structured data on a database to prepare them in order to make a comparison between the tweet and an Islamic hadith database that was previously developed by Bimba et al. [23].

The core objectives of this project are to review the available research, methods and architectures of data extraction from social media that have been carried out by other scholars and researchers in this field. In addition, this study intends to design and develop a system architecture that will be able to extract recent tweets about Islamic hadiths in the region of Malaysia from the Twitter website and store them on an appropriate database. Moreover, extracted tweets will be analyzed to determine the amount of fabricated information about Islamic hadiths that is shared on Twitter.

This paper is organized as follows: Section 2 presents a brief review of the use of Twitter in other areas of research; Section 3 presents the tools and toolsets that will be used in this project; in Section 4, the challenges of OSM data extraction and the advantages of OSM data extraction are discussed.

2. Related work

The theme of Twitter's Streaming API has been referenced by a number of reviews. Some researchers used Twitter's Streaming API to generate systems that enable scientists to analyze the behavior of the users of Twitter while they are using social media and online networks. In addition, [19] developed an architecture for a Twitter data collection. There are several studies that show that Twitter has been used to collect data for content modeling [24, 25] and statistical analysis of content [26].

Nevertheless, because of the popular usage of Twitter's Streaming API in several scientific fields, it is necessary that we understand how having a collection of information or data affects our final results. For instance, many researchers have used Twitter to detect real time social and physical events. [27] investigated the usage of Twitter as a device to detect frequent and diverse social and physical events in real-time; they examined the 2010-2011 US National Football League (NFL) games. [28,29] also studied a real time sport event. [30, 31] examined the detection of earthquakes, and [32] studied events pertaining to natural hazards (grassfires and floods). [33] examined recent news topics and bunched the corresponding tweets into breaking news. In terms of political viewpoints, researchers used Twitter to determine people's reactions to the media [13].

One of the advantages of real time detection is that the collected data enables scientists to analyze the behavior of people at the time that they tweet their opinion on social media. As mentioned previously, the aim of this research is to identify fabricated information on Twitter, which requires the extraction of data in real time. Thus, our work mimics cases that detect real time events. The other aspect that we have to consider is the accuracy of the collected data the same as other researchers with similar objectives.

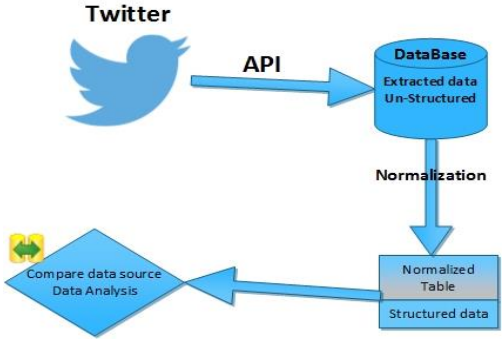

Figure 1. Architecture of Twitter extraction

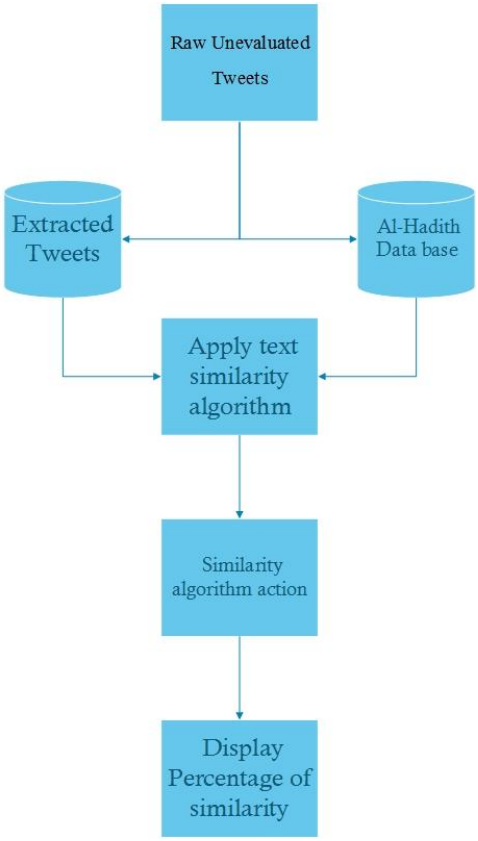

Figure 2. Evaluation of extracted tweets

3. System development

Twitter is a microblogging system that enables users to post and receive short messages of up to 140 characters, which are called "tweets". The actual number of the characters is 160 based on the size defined for the Short Message Service. However, this limit is split into 140 characters for the post-box and another 20 characters for the username. On the other hand, this limitation motivates users to produce creative posts. Twitter allows users to share URLs that may contain Videos, articles and any other information that the tweeter would like to share with his followers by Twitter's short messaging service, which is called twl.tl URL. This poses an additional difficulty for scientists if the URL is a part of the analysis system. Twitter offers considerable functionality to users for posting a simple message including the ability to send a message or reply to a friend. The message can be addressed to a specific user either on their follower list or by starting the message with "@Username". The other functionality of Twitter is the 'hashtag', which introduced users to tagging their post with a word, number or phrase in order to make it easier for people to find and follow discussions about a specific subject.

Twitter offers three application programming interfaces (APIs): Rest API, Search API and Streaming API. The Rest API allows developers to read and write tweets by programmatic access. This means that it allows developers to access Twitter data and store it in a database; the Search API returns recent or popular tweets that match with a specific query, and the Streaming API provides real-time access to Twitter data. Therefore, these APIs allow developers to achieve their target when real time data analysis is used for scientific reasons.

Fig. 1 shows a simplified architecture proposed for tweet collection. The process starts by choosing the right API and applying a code to the system and then storing tweets on a database as unstructured data. The initial step for gathering tweets will be done with the help of the PHP programing language, which will run as a continuous background process. When a user sends a tweet in real time, the Twitter streaming API will employ the PHP query to insert it into the MySQL table. Afterwards the unstructured data may be normalized and stored in a database in order to be used for future reference.

After collecting data from Twitter and storing it as structured data, it will be evaluated. In this process, the database of collected tweets will be compared with al-hadith data by applying a text similarity algorithm that is responsible for determining the similarity between texts. This means that the algorithm will be applied to identify the most comparable al-hadith in the database and compare it to the extracted tweets. This will display the percentage of similarity of the tweets based on the al-hadith database. Figure 2 illustrates the overall evaluation process. Raw unevaluated tweets will be sieved by the extracted tweets database and al-hadith database. The text similarity algorithm will be applied to both these databases and will eventually provide true percentages of tweets when compared to the authenticated al-hadith database.

In this project, to actualize the above-mentioned tweet collection architecture, Twitter APIs are going to be developed in conjunction with the PHP programming language and MySQL. More specifically, the first step of data collection will be achieved using the PHP programming language, which is a continuously running program that works to extract new related tweets about the assigned subjects. The extracted tweets will then be inserted into the MySQL table. After collecting data, a text similar algorithm will be applied in order to do the evaluation process.

4. Discussion

4.1. Challenges of Social Media Extraction:

Although information extraction from OSM and data analysis has been widely used by many organizations that analyze data for different fields, research on web data extraction shows there are still many critical challenges that may affect the quality of web data extraction. Typically, the problems in extracting data from the Web are challenging because they require several tools and techniques that are not easily accessible. In this section, the system development, velocity of extraction, availability of data and data management system are elaborated upon.

- System development:

To extract raw data from an OSM like Twitter, developers often create various system that are able to extract the requested data. This means that the best way to create a system for extraction is to divide the data in the extraction program into many modules, and, subsequently, command them to extract a part of the data. This data is then combined together as a dataset. Such a process requires an expert programmer who is able to face this challenge.

- Velocity of extraction:

One of the main targets of data extraction from OSMs is to prepare a set of information either for analysis in exclusive areas, such as businesses, or create a collection of vital information for organizations like the government. For instance, in the field of business intelligence, analyzing data quickly and efficiently can provide a competitive advantage. It is a strength for a company to have the first set of data in order to plan their strategy at the appropriate time. Therefore, the tools and techniques that are used for data extraction must be able to mine a large amount of data in the shortest possible time.

Status updates on social media also raise many practical performance challenges in that data extraction needs to be performed in the run time. A huge amount of information must be extracted within an appropriate time, which exacerbates the difficulty of the technical aspects of data extraction.

- Availability of data:

Information on the Web is heterogeneous. Many users of OSM may provide the same or similar information using different words and formats. This problem raises the challenge of choosing the right source for the data extraction process, which requires the extraction source that presents the most accurate information.

- Data management system:

To be more efficient and valuable, the data extracted from OSM should be stored on a database. The rapid growth of OSM data may cause an increase in the volume of the extracted data, which may be called Big Data. Big Data refers to "massive data sets" with a large and complex structure that requires a high capacity storage database and an architecture to manage it. For these kinds of system, applications must have a strong management system in order to categorize the extracted data and create integration between other applications for future actions.

4.2. Advantages of Twitter for Data Extraction

For OSM data extraction, Twitter is a great tool. The key to its popularity and broad appeal to people lies in its innate openness for community consumption. It also provides high performance tooling for developers and researches because it offers thorough and well-organized data and information. Twitter data is especially remarkable because it is available to its users as soon as it is uploaded. Hence, this multifaceted application represents the true cross-section of the community it serves. As mentioned previously, Twitter "following" interactions can range from small dialogues to interest groups that cater to the information for which people have an affinity. Another important advantage of Twitter data is that each tweet has its own author. To perform sentiment analysis, we need to label data before applying suitable learning algorithms. For example, [34, 35] found that for sentiment analysis, the training data needs to be stored and Twitter has the advantage that it possesses tweeter-provided sentiment indicators.

5. Conclusion

In conclusion, this case study will demonstrate that Twitter is a viable tool for data extraction and analysis for mass consumption and research in Malaysia. This study will attempt to pool data from Twitter and evaluate its authenticity since people can freely state what they had in mind on OSMs without reviewing their comments. It is imperative to a have type of 'information police' so that only true and authentic information is allowed to appear on OSM. Further studies could incorporate Twitter as a device for data collection, extraction and analysis from OSM because of Twitter's broad appeal, user compatibility and API tools for developers.

Acknowledgement

We gratefully acknowledge the University of Malaya for supporting this research through UMRG Grant (RP003C-14HNE).

References

[1] J. Kleinberg, *The convergence of social and technological networks.* Communications of the ACM, 2008. **51**(11): p. 66.

[2] A. Mikroyannidis, B. Theodoulidis, and A. Persidis, *PARMENIDES: towards business intelligence discovery from web data*. in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*. 2006. IEEE.

[3] R. Baumgartner, et al. *Web data extraction for business intelligence: the lixto approach*. 2005: na.

[4] W. F. Cody, et al. *The integration of business intelligence and knowledge management.* IBM systems journal, 2002. **41**(4): p. 697-713.

[5] J. Srivastava, R. Cooley, *Web business intelligence: Mining the web for actionable knowledge.* INFORMS Journal on Computing, 2003. **15**(2): p. 191-207.

[6] K. Denecke, W. Nejdl, *How valuable is medical social media data? Content analysis of the medical web.* Information Sciences, 2009. **179**(12): p. 1870-1880.

[7] G. Eysenbach, et al. *Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review.* Jama, 2002. **287**(20): p. 2691-2700.

[8] J. Kleinberg, *The small-world phenomenon: An algorithmic perspective*. in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. 2000. ACM.

[9] L. Backstrom, et al. *Four degrees of separation*. in *Proceedings of the 4th Annual ACM Web Science Conference*. 2012. ACM.

[10] M. E. Newman, *The structure and function of complex networks.* SIAM review, 2003. **45**(2): p. 167-256.

[11] C. Plake, et al. *AliBaba: PubMed as a graph.* Bioinformatics, 2006. **22**(19): p. 2444-5.

[12] Z. Xiang, U. Gretzel, *Role of social media in online travel information search.* Tourism Management, 2010. **31**(2): p. 179-188.

[13] D. A. Shamma, L. Kennedy, and E. F. Churchill. *Tweet the debates: understanding community annotation of uncollected sources*. in *Proceedings of the first SIGMM workshop on Social media*. 2009. ACM.

[14] M. Gjoka, et al. *Walking in Facebook: A case study of unbiased sampling of OSNs*. in *INFOCOM, 2010 Proceedings IEEE*. 2010. IEEE.

[15] S. A. Catanese, et al. *Crawling facebook for social network analysis purposes*. in *Proceedings of the international conference on web intelligence, mining and semantics*. 2011. ACM.

[16] *Facebook*. Available from: http://www.facebook.com.

[17] *Twitter*. Available from: http://www.twitter.com/.

[18] *Linkedin*. Available from: https://www.linkedin.com/.

[19] M. Oussalah, et al. *A software architecture for Twitter collection, search and geolocation services.* Knowledge-Based Systems, 2013. **37**: p. 105-120.

[20] *Online source*. Available from: http://www.about.twitter.com/company.

[21] *Online documentation*. Available from: https://dev.twitter.com/overview/documentation.

[22] B. K. Chae, *Insights from Hashtag# SupplyChain and Twitter Analytics: Considering Twitter and Twitter Data for Supply Chain Practice and Research.* International Journal of Production Economics, 2015.

[23] A. Bimba, et al. *Towards Enhancing the Compilation of Al-Hadith Text in Malay*. in *International Conference on World Islamic Studies* 2015. Seoul, South Korea.

[24] L. Hong , B. D. Davison, *Empirical study of topic modeling in twitter*. in *Proceedings of the First Workshop on Social Media Analytics*. 2010. ACM.

[25] A. Pozdnoukhov, C. Kaiser, *Space-time dynamics of topics in streaming text*. in *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks*. 2011. ACM.

[26] M. Mathioudakis, N. Koudas. *Twittermonitor: trend detection over the twitter stream*. in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 2010. ACM.

[27] S. Zhao, et al. *Human as real-time sensors of social and physical events: A case study of twitter and sports games.* arXiv preprint arXiv:1106.4300, 2011.

[28] D. Chakrabarti, K. Punera, *Event Summarization Using Tweets.* ICWSM, 2011. **11**: p. 66-73.

[29] J. Hannon, et al. *Personalized and automatic social summarization of events in video.* 2011: p. 335.

[30] T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, in *Proceedings of the 19th international conference on World wide web*. 2010, ACM: Raleigh, North Carolina, USA. p. 851-860.

[31] Y. Qu, et al. *Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake*. in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 2011. ACM.

[32] S. Vieweg, et al. *Microblogging during two natural hazards events: what twitter may contribute to situational awareness*. in *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010. ACM.

[33] J. Sankaranarayanan, et al. *Twitterstand: news in tweets*. in *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*. 2009. ACM.

[34] A. Bifet, E. Frank. *Sentiment knowledge discovery in twitter streaming data*. in *Discovery Science*. 2010. Springer.

[35] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. 2013: " O'Reilly Media, Inc.".