Initiating an Online Reputation Monitoring System with Open Source Analytics Tools

Mohd Ilias M Shuhud, Najwa Hayaati Md Alwi and Azni Haslizan Abd Halim
Faculty of Science and Technology, University Sains Islam Malaysia, Nilai, Malaysia
ilias@usim.edu.my

Abstract. Online reputation is an invaluable asset for modern organizations as it can help in business performance especially in sales and profit. However, if we are not aware of our reputation, it is difficult to maintain it. Thus, social media analytics is a new tool that can provide online reputation monitoring in various ways such as sentiment analysis. As a result, numerous large-scale organizations have implemented Online Reputation Monitoring (ORM) systems. However, this solution is not supposed to be exclusively for high-income organizations, as many organizations regardless sizes and types are now online. This research attempts to propose an affordable and reliable ORM system using combination of open source analytics tools for both novice practitioners and academicians. We also evaluate its prediction accuracy and we discovered that the system provides acceptable predictions (sixty percent accuracy) and demonstrate a tally prediction of major polarity by human annotation. The proposed system can help in supporting business decisions with flexible monitoring strategies especially for organization that want to initiate and administrate ORM themselves at low cost.

1. Introduction
In early October, 2017, the government of a Middle-Eastern country banned four Malaysian universities for administrative issues [1]. The incident raised various reactions, especially by local and international netizens, which indirectly or directly affected the reputation of the involved universities. Furthermore, there were rumours made on social media which claimed that the incident was primarily due to education quality. However, in another report, the vice chancellor of one of the universities took a quick action to refute the rumour with the fact that the reason involved administrative matters [2]. It is expected that the management of each of the affected universities will strategize for rectification of the damage caused to its reputation and public image. However, they might have interested to the current status of their online reputation.

The above scenario is an example of an online reputation attack, which is a growing concern of many organizations today that is probably more harmful than computer viruses. The rapid development of social media has challenged various aspects of a business, including online reputation. While positive reputation can influence purchases, investment and loyalty towards products or services, a negative reputation may have an opposing effect, e.g., customer attrition, sales reduction, and distrust. Therefore, organizations should be aware of their reputation by monitoring what people say about our brands or products. “If a company is unaware of its own reputation, it cannot manage it successfully” [3]. One way to manage reputation is by implementing online reputation management to support business decisions [4], [5]. Unfortunately, many organizations do not implement ORM due to cost, complexity and reliability [6]. For example, the cost of negative review removal in search engine can be from RM50,000 (USD 12,800) according to an ORM company.
This research attempts to propose an affordable and reliable ORM system using open source data analytics tools and then evaluate its performance. The research contributes in enabling an online reputation monitoring implementation for any type and size of organizations at almost no cost. It also allows academic research to explore social media data for various purposes. In this research we combine, different open source tools such as data science software platforms, business intelligence platforms, text analytics Application Programming Interfaces (API) and Microsoft Excel Online. We began with a background study of the existing literature, followed by a context study which involved an interview. Later, we conducted experiments to evaluate the developed ORM system environment. We discovered that the system produced acceptable polarity prediction accuracy (over 50 percent of human annotation) and produced similar prediction of major polarity type with human annotation in sentiment analysis of a news.

This paper is organized as follows. The next section reviews the background of important keywords in the research and theoretically support the study. Next, we present the research methodology in Section 3, and then present our experiment results and evaluate selected analytics tools in Section 4. In Section 5, we provide some insights from the experiments, including advantages and limitations. Finally, we conclude in Section 6.

2. Background
This section provides a background on the key terms in the research from the relevant literature.

2.1. Social media analytics
In defining social media analytics, [7] describes it as “the practice of gathering data from social media websites and analysing the data using social media analytics tools to make business decisions”. The definition highlights making business decisions as the ultimate goal, regardless of the type of organization whether commercial, government and non-profit organizations. The definition by [7] is in-lined in a more comprehensive definition by [8] in his book, where "social media analytics is the art and science of extracting valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making. It is a science, as it involves systematically identifying, extracting, and analysing social media data (e.g., tweets, shares, likes and hyperlinks) using sophisticated tools and techniques. It is also an art, interpreting and aligning the insights gained with business goals and objectives. To obtain value from analytics, one should master both its art and science”[8].

Social media analytics falls under interdisciplinary fields in computer sciences and social sciences. The practice of social media analytics falls under computer science, however, the analysis produced is useful in a wide spectrum of disciplines in social science such as political science, sociology, geology, business and marketing and journalism [9]. Among the various purposes of social media analytics in commercial organizations is to support marketing and customer service processes by mining customers’ sentiment from comments, review, and tweets. In social media analytics, the most critical step is to determine how the collected and analysed social media data can benefit specific business goals [8].

2.2. Online reputation monitoring
Online reputation monitoring or management (ORM) is a new element of public image for entities, especially organizations that establish their presence online. Online reputation can be defined by search results on a search engine, social media comments, a ranking of consumer reviews, articles and news about the organization or the related attributes such as products and services [10]. What do we see when we search for our name and company name on Google or Facebook? If the first page is all good about us, then this is a sign of good online reputation. But if it is bad information, we might be in trouble, as people normally do not check the next pages. Large organizations such as eBay.com and Amazon.com have embedded an ORM in their e-commerce platform so that customers can research how people think about and rate certain products. This is because 90 percent of online purchases are driven by product reviews [11]. This shows that online reputation is an invaluable intangible asset that should be monitored by organization of all sizes, types, and popularities. Some companies or
government departments outsource the ORM to a third-party company for analytics services, while others employ data analysts to work internally [12]. Among of the useful tools in ORM is sentiment analysis, which provides an overall binary polarity value for an entity’s online reputation. There are many works in the literature that have discussed ORM in various industries such as healthcare, politics, journalism, hotels, tourism and automotive. For example, in [13], researchers suggest that online reputation with sentiment analysis can enhance competitiveness in the automotive industry.

2.3. Sentiment analysis
Sentiment analysis has been established as among the hot research topics across domains in the past decade [14]. It is one of many content analysis methods in social media analytics, and is also called opinion mining, subjectivity analysis, and appraisal extraction. Sentiment analysis is essentially a natural language processing task that involves mining attitudes, opinions and emotion from text or speech [15]. Sentiment analysis is a classification problem where opinionated texts are categorized as “positive”, “negative” or “neutral” using either machine learning algorithms or predefined lexicons [16]. Businesses can use sentiment analysis to understand consumers’ emotional reactions towards their brand, company or signature products and services, and whether they are accepted positively or negatively on the Internet [17].

2.4. Related works on social media analytics using open source tools
There are currently many ORM companies offering their solutions. However, the service cost varies based on the needs and difficulty [6]. Thus, we attempt to explore other options in open source systems. In the work of [18], researchers attempt to develop a Java program to fetch real-time data from various social media platforms through various APIs for analysis purposes. Although they managed to extract some data, there are some restrictions set by social media platforms that limited the output. Meanwhile, in [19], an open source social analytics tool called Netlytic was evaluated. The researchers gathered contents from Facebook and YouTube for analytics and proved that social media analytics is not costly if we are able to explore alternative ways rather than to purchase predefined software. In the work of [20], the researchers discusses the ethical and legal issues on gathering social media data from APIs, which is another important angle that should not be overlooked.

Besides social media data gathering, [21] introduced an open source API for text analytics, crawling and interpretation called TACIT. The API features an intuitive tool and interface to utilise the technologies in text analysis and big data management for use by other researchers. Using multiple APIs in developing analytics environment can be a complex process. [22] provides a relevant method mapping across APIs called TMAP by using text mining on natural language API method descriptions. The method can assist software developers to migrate applications from a source API to the target API automatically.

3. Research methodology
In this research, we divided the research methodology into three phases, namely: (1) Foundation Phase (2), Experimentation Phase, and (3) Evaluation Phase.

3.1. Foundation phase
In this phase, we have conducted three sub-phases of groundwork, which are; background study, context study and tools selection.

3.1.1. Background study
We have conducted a review of relevant studies in the literature around important key words such as online reputation, online reputation management, social media analytics, user-generated contents (UGC) and sentiment analysis. The online databases used for searching for the literature are Google Scholar, Scopus, Mendeley and others. Part of the review is presented in Section 2.

3.1.2. Context study
To acquire a deeper understanding on the context of the study, especially online reputation monitoring practices, we have conducted an interview with a key person at a public relations unit of one of the involved universities in the banned incident mentioned above. The collaboration with the public relation unit of the university has been started a week before the banned issue was raised. The interview was transcribed and analysed for a better understanding on the practices by the organization.

3.1.3. Tools selection

Furthermore, we select open source data analytics tools which are appropriate for the experiments. The tools are required for social media data extraction, text processing, sentiment analysis, and report analysis. Three aspects are considered in selecting the appropriate tools: ease of use, accessibility, and reliability. Ease of use means that the tools are not too complex for a non-IT background user and a graphical user interface (GUI) is necessary. Accessible criteria mean that we have the access to the tools in terms of knowledge, e.g., training and acquisition of the tools. In terms of reliability, this refers to open-source text analytics APIs, which include sentiment analysis module in the package. For this, we need a tool that provides acceptable accuracy, or at least over fifty percent from human annotated polarity prediction. As mentioned earlier, we have selected a data science software platform called RapidMiner Studio for data processing which allows 10,000 rows per transaction to be processed in the free version. For sentiment analysis, we used a machine learning text analytics API which will be referred as ‘M API’ in this paper. M API is among of text analytics API available in RapidMiner Marketplace. In extracting comments from Facebook, we used Microsoft Power BI, an open source business intelligence and analytics program. We chose Facebook comment as it has more than 140 characters in a single post, unlike Twitter. Thus, more texts can be analysed. As for customization and filtration of the extracted data and results, we used Microsoft Excel Online a free version of the software included in Microsoft Office Online or Microsoft Office 365.

3.1.4. Experimentation phase

In this phase, we conducted two experiments to evaluate the accuracy of the selected API in predicting sentiment polarity of UGC and to evaluate the system capability to perform sentiment analysis towards bad or sensitive news related to an organization.

3.1.5. Evaluation phase
In the evaluation phase, we provide some insights on the performance of the open source analytics tools in the experimentation phase. In summary, our proposed online reputation monitoring system using opensource analytics tools involved three main processes which are data extraction, execution, and analysis as shown in Figure 1. Firstly, data extraction is the process of obtaining the social media contents in this case Facebook using an open source business analytics platform Microsoft Power BI. The second main process is the execution of sentiment analysis using selected text analytics API which is connected to a data analytics software platform, RapidMiner. Finally, the generated results can be exported into Microsoft Excel for analysis using various formulas. Visualization using Excel is also useful for better understanding and to support business decisions especially relation to online reputation.

4. Findings
In this section we present and discuss the findings and connect them with entire phases of the research.

4.1. Foundation phase outcome
Based on the conducted interview, we discovered that the organization does not have any preventive and curative strategy to use social media analytics for dealing with such online reputation attacks. However, they do invest in improving their organization’s visibility on social media, e.g., through the promotion of achievements of the university and organized programs.

The key person quoted “We have not implemented any social media analytics so far for any purposes including Online Reputation Management like you mentioned, but we agree on the importance of online reputation protection… but we think it is not urgent at the moment and I believe it is a complex and costly thing”. The quotation shows that the officer’s perception towards the connection of ORM and the role of social media analytics in the job is not convincing. They perceived that social media analytics is a complex process, and is costly for acquiring the tools and services, and hiring staff e.g. data analysts. However, the key person expressed an interest to know what people are saying about his organization, either from the inside or outside. The interviewee quoted “But, we are interested to know how students respond to our posts on social media, so we always check their comments … negative comments are extremely low on our page, but we don’t know if there are many on other pages”. The lack of negative comments at an organization’s social media account does not mean our online reputation is 100 percent positive. Many factors can contribute to the polarity such as internal social communication culture. In a university context, the students may be worried to comment negatively, but this cannot be generalized to all universities, as some are more open for criticism. Furthermore, online reputation as explained in [5] is defined as how people in the world are looking at us, not only from inside the organization. The issue on the banned universities is one example of how an organization’s online reputation can be at risk. Thus, ORM with social media analytics provide preventive measure before a certain issue raised and also monitor an ongoing issue in news report to preserve reputation.

4.2. Experiment phase outcome
There are two tests in this phase which is accuracy test and sentiment analysis test. First, we want to observe and evaluate the accuracy of the selected open source text analytics APIs on product posts from Facebook. This is to know the reliability of the opensource API to be used in reputation monitoring. Second, we attempt to observe the sentiment polarity on news posted on Facebook that is predicted by the text analytics API. Figure 1 above shows the entire process involved in the proposed system. As we have measured the accuracy of the M API in experiment one, we believe that the API is reliable for experiment two. In the following, we described the results of the experiment according to the processes involved in the experiment in the next section.

4.3. Data extraction
In this study, we have extracted Facebook comments from two official Facebook pages. We have two different tests for different datasets extracted from Facebook. Dataset one consists of 1,122 lines of comments from three posts of a Facebook page of an international car company operating in Malaysia
between May 1st to July 14th 2017. Meanwhile, dataset two consists of 89 comments from October 2nd to October 4th, 2017, from a local English online newspaper on the report of the banned universities.

The Facebook comments were extracted using Microsoft Power BI. We need to connect a specific Facebook account name to extract comments from it. There was a simple query in DAX (Data Analysis Expression) function language needed to be edited. Meanwhile, the speed of the extraction process depends on computer performance and the Internet connection. The extracted data were placed in an Excel worksheet for easy customization. For example, we can filter the extracted data based on our needs such as comments that contain certain words or comments posted at a certain date and time. The selected data were downloaded in CSV (comma separated values) format then exported in Microsoft Excel format (.xlsx) to be processed by RapidMiner Studio for sentiment polarity prediction.

Some comments needed to be removed for example those Malay, Chinese and India Language as well as comments that only consist of tagged Facebook ID and those in symbols or emoticons only. The API and RapidMiner generally support English and few other languages such as Spanish and German. We hired an advanced English user to annotate all 1,122 extracted comments with positive, ‘negative’, ‘neutral’ and ‘other’ for each comment. Other means that the comment needed to be removed. For dataset one, only 571 comments accepted, and 551 comments were removed. However, all 89 comments for dataset two were accepted.

4.4. Execution
In experiment one on the evaluation of the accuracy of selected API, we discovered that the results were not convincing but acceptable. The API predicted 305 comments or 0.60 of the total correct or similar with human annotated polarity. The matching process was done using Microsoft Excel’s “EXACT” formula for text: ‘=EXACT(text1, text2)’ where text1 = annotated text and text2=predicted text. If the text matched, e.g., text1 = negative, text2=negative, then it will print “TRUE”. The count of TRUE and FALSE comments was done using COUNTIF formula: COUNTIF (range, criteria).

We observed that, polarity prediction of Facebook comments by the API are acceptable. This is because both manual and computerized prediction show majority of negative polarity. Manual annotation, records 286 and API predicted 214 of negative comments. The second highest polarity type by manual and API prediction is neutral, which is 199 and 207 comments. Based on the extracted comments, we suspected that the high amount of neutral comments predicted was due to the mixture of languages, short form of words, as well as sarcastic words. Meanwhile, the very low positive comments mean that the car company should do something to reduce the negative comments which can risk their online reputation. We can conclude that the customers were strongly dissatisfied with products from the car company.

### Table 1. Results of Accuracy Test (Car Company) and Sentiment Analysis Test (Banned Universities Reports) by M Text Analytics API

|                  | Accuracy Rate          |
|------------------|------------------------|
| **Experiment 1** |                        |
| True             | 0.60 (331 comments)    |
| False            | 0.40 (220 comments)    |
| **Sentiment Prediction** |               |
| Positive         | 0.07 (6 comments)      |
| Negative         | 0.12 (11 comments)     |
| Neutral          | 0.81 (72 comments)     |

In experiment two of banned universities news dataset, the API predicts a high number of neutral comments, as shown in Table 1. However, based on the comments from the news, we discovered another reason for the high amount of neutral comments is due to the tagging of other Facebook IDs. There is no comment written, but readers wanted to alert their contacts about the news. Nevertheless,
negative comments are only slightly higher than positive comments which is to us not very impactful in our opinion.

The API also provides polarity confidence between 0 to 1.0, where a number close or equal to 1.0 means a higher polarity, e.g., very negative comment. We filtered the data in Excel Online to select only the comments which have a polarity confidence of more than 7.0. One of three comments classified as 1.0 by API 1 sounds harsh and emotional as “Shame Shame Where U put your s*****d face Now?”. We also filtered the lowest polarity confidence of negative classified comment, which has score of 0.475, and was quoted as “*** is a joke university”. The comment referred to one of the affected universities in a short form word. The comment looks cynical or sarcastic, and is not literally negative, but its meaning is negative. However, we supposed that collecting and combining more online news about certain issue with embedded Facebook comments can bring better results. Probably, the use web crawler is good option.

5. Implication for design

In the following, we conclude several advantages and limitations of the designed open source online reputation system.

5.1. Support in business decision making

Although the polarity prediction is not 100 percent accurate, it is acceptable for business decisions such as strategy development, marketing research and action plans. The prediction can be used to overview subjective opinions from customers toward certain issues, people, products, and services. This provides evidence-based reputation for organizations to allow them to manage it [3]. For instance, in of sentiment analysis on the car company above, proactive actions can be taken to reduce negative reputation. This is in line with [13], who mention that online reputation can improve competitiveness in the automotive industry. In case the organization participated in context study above, this approach can help it to analyse people’s reactions towards the banned universities issue for appropriate action planning and online reputation recovery. Furthermore, they can go directly to negative comments or rumours which show a high polarity confidence and respond directly to the comments for rebuttal or denial.

5.2. Flexible monitoring strategies

We observed that the approach can be used for periodical monitoring, e.g., daily, weekly, or monthly; event-based monitoring, e.g., rumours or sudden issues; and entity monitoring, e.g. products, services and employees [11]. Periodical monitoring is probably a good preventive action to always keep pace with online reputation status. In case of the interviewed organization, event-based monitoring can be carried out to keep track of people’s reactions towards the news.

5.3. Affordable for novice

In general, this approach can be applied to any organization or individual, e.g., celebrities or politicians. However, we suppose that large organizations prefer to outsource ORM services to analytics companies or acquire a sophisticated solution. The experiments above do not involve any cost, as all the tools used are open source with certain limits. We suggest that this approach is probably appropriate for small to medium sized profit-based or non-profit-based organizations or even individuals who are novice to ORM, but are interested to implement it [6]. It is also useful for academic researchers who are new to social media analytics and ORM.

6. Limitation

Despite the advantages of the approach, we discovered some limitations which are worthy to be shared. Among important limitations is the lack of Malay Language support by either the data science
software platform, RapidMiner, or the text analytics API. This is because contents in the local language has a greater value to be analysed.

7. Conclusion and future work
This research attempts to propose an affordable and reliable ORM system using open source tools for any type and size of organization. The task can be a new role for public relation unit and it can be performed by a staff with IT background. The research also evaluates the performance of the system in term of prediction accuracy. We discovered that the accuracy of predictions by M text analytics API is acceptable (over 50 percent accurate). However, different text analytics API may produce different accuracy rate. This shows that the system is reliable to support business decisions and offers flexible monitoring purposes as preventive action. Monitoring strategies can be product or brand based or issue based. In future work, we want to extend the research on the evaluation of topic and entity extraction from user-generated contents (UGCs) using open source tools. We want to know what topics are involved in the conversation based on the extracted social media data.

References
[1] Jay B (Update) Oman Bans Four Malaysian Universities Over Alleged ‘Academic, Administrative Abuse, Kuala Lumpur, 02-Oct-2017
[2] Bernama “USIM Says Administrative Issues to Blame for Oman Ban,” Seremban, pp. 6–11, 06-Oct-2017
[3] Aula P and Heinonen J 2016 Evidence-based reputation management Reputable Firm. Manag. Prof., 1, pp 163–199
[4] E. Commentary 2017 “Management Mitigating Risks through Online Reputation,” pp. 1–7
[5] Conner C 2014 “Top Online Reputation Management Tips for Brand Marketers,” Forbes
[6] Hartzer B 2017 “How Much Does Online Reputation Management Cost?” Billhartzer.com, [Online]. Available: https://www.billhartzer.com [Accessed: 29-Oct-2017]
[7] Rouse M 2017 “What is social media analytics?” searchbusinessanalytics.techtarget.com [Online]. Available: https://www.cio.com/article/3185725/security/9-biggest-information-security-threats-through-2019.html. [Accessed: 09-Sep-2017]
[8] Khan G 2015 Seven Layers of Social Media Analytics: Mining Business Insights from Social Media Text, Actions, Networks, Hyperlinks, Apps, Search Engine, and Location Data 1st ed. (CreateSpace Independent Publishing Platform) chapter 1 p 1
[9] Segerberg A and Bennett L 2011 Social media and the organization of collective action : using twitter to explore the ecologies of two climate change protests social media and the organization of collective action Commun. Rev. 14 3 pp 197–215
[10] Capatina A and Schin 2017 Perceptions on Online Reputation Management in Tourism Industry (The Case of a Romanian Hotel) J. Manag. Res. 9 2 p 71
[11] Guo S, Wang M and Leskovec J 2011 The Role of Social Networks in Online Shopping: Info. Passing, Price of Trust, and Consumer Choice Proc. ACM Conf. E-Commer pp 157–166
[12] Dorčák P and František P 2015 Analysis of the possibilities of improving an online reputation of public institutions J. Soc. Sci. 4 1 pp 275–282
[13] Dorčák P, Peter M, and Pollák F 2015 Multifactor analysis of online reputation as a tool for enhancing competitiveness of subjects from automotive industry Ekon. 65 2 pp 173–186
[14] Hussein D 2016 A survey on sentiment analysis challenges J. King Saud Univ. - Eng. Sci.
[15] Mejova Y 2012 Sentiment analysis within and across social media streams p. 190
[16] Kharde V and Sonawane 2016 Sentiment analysis of twitter data : A survey of techniques Int. J. Comput. Appl. 139 11 pp. 5–15
[17] Laeeq F and Tabrez N 2017 Sentimental Classification of Social Media using Data Mining Int. J. of Adv. Res. in Comp. Sci. (Udaipur) 8 5 pp 546-549
[18] Liao W and Chhoeut R 2017 Using Social Networking APIs to Collect Real-time Data for Analytics Social Proc. West Virginia Acad. Sci. vol. 89, no. 1
[19] Sathish R, Ambika M, Prabhu R and Krishnakumar 2017 R Real time data collection and
analytics of social media sites using netlytic Middle-East J. Sci. Res. 25 1 pp 174–180
[20] Lomborg S and Bechmann A 2014 Using APIs for Data Collection on Social Media Using APIs for Data Collection on Social Media Int. J. Inf. Soc. 30 4 256–265
[21] Dehghani M, Johnson K, Garten J, Boghrati R, Hoover J, Balasubramanian V, Singh A and Shankar Y 2016 TACIT: An open-source text analysis, crawling, and interpretation tool Beh. Res. Met. 49 2 pp 538-547
[22] Pandita R, Jetley R, Sudarsan S, Menzies T and L. Williams 2017 TMAP: Discovering Relevant API Methods Through Text Mining Of API Documentation Special Issue SCAM