TWITTER DATA ANALYTICS IN EDUCATION USING IBM INFOSPHERE BIGINSIGHTS

Miloš Popović,
Milan Milosavljević,
Pavle Dakić

Singidunum University,
32 Danijelova Street, Belgrade, Serbia

Abstract:

This paper explains the usage of Twitter data in a number of subjects on higher education institutions. It examines procedures for defining the interactions between students and professors through social networks, in order to develop more effective teaching. Describing in detail the way in which Big data technology such as "IBM InfoSphere BigInsights" allows processing of the data, which are primarily downloaded from social networks and later organized in Hadoop environment. To that end, we have designed an application for creating data file based on the search query. Evaluation through analysis of exam results confirms that the proposed solution produces better results in terms of student's academic achievements. Students are encouraged to achieve better results, which ensures better cooperation in communication of all participants in the educational process.

Key words:

Twitter, data analytics, education, big data, IBM InfoSphere BigInsights.

1. INTRODUCTION

In the last three years, human society has created and saved more information than at any time in history [1]. Twitter on daily occurrences has more than 200 million tweets. Students are widely using social networks in education, indicating that Twitter has a large amount of information that can be further analyzed and used in the existing educational systems. It is possible to find out who and what term they are talking about, and what they made the search query for.

The most important result of this research is the implementation of techniques for the analysis of Twitter social network and the application in terms of higher education in Serbia. Originality is reflected in defining the methodological procedures for analyzing Twitter data using the application "TDAedu" that with help of the Big data technology - IBM BigInsights, can be used for statistical analysis in creating better educational system.

The overall objective of this work is to help teachers on certain subjects in higher education institutions to modernize their teaching content, using data that has been collected from social networks. Current trend in the world is in using the links to the latest blogs, or comments of the most influential Twitter users, in connection with the most frequently searched term.

Correspondence:
Miloš Popović

e-mail:
popmilos7@yahoo.com

DOI: 10.15308/Sinteza-2016-74-80

2. RELATED WORK

Social networks such as Twitter and Facebook provide an opportunity for researchers in different fields to understand human behavior and social phenomena [2]. Three researchers jointly analyzed 3,000 tweets that were randomly selected from the tweets containing hashtag **#engineeringProblems**, and through analysis we can get interesting information that can be later used in various ways. This paper examines sentimental analysis which has recently become the focus of many researchers, since the analysis of online texts is useful and required in a variety of applications, aiming to inform users about the pros and cons in using various products [3]. It has been shown that, sentimental analysis of tweets, by using different algorithms, detecting sarcasm and summary of sentiment, is possible to obtain significant results. In addition, the paper [4] presents the methods of using Twitter data and sentimental analysis for prediction, such as the election predictions in Pakistan in 2013 and India in 2014. The analysis results were successful, since the studies have shown the exact result, even two months before the official results of the elections had been revealed.

The paper suggests a method for interaction between students and professors during the lecture, by using Twitter. The professor would ask surprise questions in the form of a quiz, after which students would submit their answers via smartphone [5]. Only a limited number of students who first answer correctly, for a certain amount of time, would be awarded with points.

This method would enable solving the problem of reduced concentration, resulting from frequent usage of smartphones for non-educational purposes. It forces students to concentrate on the lectures if they want to achieve a good result. The analysis of exam results and post-course surveys confirms that the suggested solution provides better results in terms of academic achievement of students, and their concentration during the lectures. This approach can be implemented without additional costs of purchase, instructions, or maintenance for each student.

Considering all the benefits Twitter can provide, as well as the earlier research regarding this matter, we can see that Twitter can help teachers to expand their lectures, depending on what currently the world's hottest topic is, when it comes to the research field. In addition, the teacher can analyze the student groups, and based on their discussions, comments, pictures, the teacher can obtain results about the potential problems, suggestions, hot topics and students wishes; which can be useful in preparation of teaching material.

3. CHARACTERISTICS OF SOCIAL NETWORKS

Social networks imply the formation of virtual communities, where users can exchange their opinions, collaborate and discuss topics related to their common interests, *etc*. The use of social networks has changed the way we spend our free time, do business and communicate, as well as the ways of learning and organizing teaching process. The tendency of people to exploit the possibilities offered by social networks are increasingly being used in education, in order to adapt to the needs of the learning process of students and making it interesting and appealing. We have noticed that the students started creating groups on social networks in order to connect and share information about the contents and academic obligations.

This has led the teachers to start creating and actively using social networks, so they can give students more accessible contents, encourage discussion among them and provide them access to variety of helpful examples and videos, which can help them review the content in a new way, taking a proactive approach in studying and critical understanding of the content.

Twitter as a source of information

Twitter represents a social network that allows users to send and read short messages of up to 140 characters. Twitter was created in March 2006 and was officially launched in July 2006. Twitter is being developed at an incredible rate, reaching over 200 million users with more than 200 million tweets a day. Users create their own Twitter account, and from the moment they receive the account information they can start "tweeting", a term for sending messages on Twitter. Users can follow the other users' tweets, a process known as "following". These users are also called "followers". Tweets that the user is sending are visible to all. However, users can choose the option of sending tweets to their "followers" in particular so they will not be visible to the public.

The goal of Twitter is very simple. It is designed for short information and communication and the users are usually between 18-34 years old [6]. About 300 participants participated in a poll conducted on the Internet, of which 70% from Europe and the rest from North

America and Asian countries. All respondents answered the set of questions on how they were using Facebook and Twitter, and which of the these two social networks they preferred.

In order to connect the questionnaire results with the subject's character, each subject was given a personality test - standard measuring instrument that assesses personality traits based on openness (for new experiences), conscientiousness, extroversion and neuroticism (emotional (in) stability). In addition, intellectual curiosity, need for cognitive stimulation and acquiring new skills were also measured among the respondents. Social use of Twitter is associated with high scores on tests of sociability and openness as a dominant trait (but not with neuroticism), while negatively correlated with conscientiousness. This finding suggests that Twitter is not used to deal with loneliness, but a source of information. Limited with short statuses, Twitter users use the social network for cognitive stimulation, without socializing. Taking all of this into consideration, Twitter was selected as the main source of updated information which can improve the quality of education.

Social analytics is a term that has become very popular in recent years, mostly because by using these sites on a daily basis, a large amount of data is being created, and the information is publicly available via API (Application Programming Interface). Therefore, using the techniques of analysis with social networks, we have come up with some interesting new information that may be of important use in various areas such as: computer science, economics, marketing, politics, sociology, *etc*.

In Belgrade Business School, there are 7 professional studies departments of all the above mentioned areas, so in addition to the realization in the computer science department, the idea can also be implemented to other departments, where the professors would able to enrich their lectures with the latest trends using Twitter analysis.

4. BIG DATA

Big data represents the ability to manage a large amount of different data at reasonable speed and within an appropriate time frame, in order to allow data analysis in real time. The data is supplied from many different sources and can be found in various forms. With fast of development of sensors, smart devices and social networks, generated data has become complex, primarily because it includes not only traditional structured data, but unstructured or semi-structured data as well.

Big data analytics allows you to work through massive amount of information in real time and previously collected information as to find the unseen patterns and nonsense to the eye, that can lead to new findings, and point out to opportunities for new services and products. Moreover, it allows the development of more effective ways of functioning, improving the transparency and accountability of institutions. Big data has a huge impact on society. Everyday people are offered a variety of opportunities, as digital users, in order to enhance the value of information. Data is already a source of power in the modern world and a very valuable product for those who can analyze it. All people are the users of digital data. It allows us to have the access to all the knowledge that was once considered out of reach or was the privilege of the few.

Big data technology - IBM InfoSphere BigInsights

It is estimated that one third of the global information is stored in the form of alphanumeric text and "Still image" data format, which is extremely useful for most Big data technologies. Big data technology and the benefits it brings have been recognized by the leading software companies that deliver commercial software. IBM BigInsights is a software platform that can help in analyzing large amount of a broad spectrum data - the data that is often ignored or rejected because they are too impractical or difficult to process using traditional methods.

In order to effectively carry out the value of such data, BigInsights included several "open source" projects, including Apache Hadoop and several technologies developed by IBM, including BigSheets tool. Hadoop and its related projects provide efficient software framework for "data-intensive" applications that exploit distributed computing environment to achieve a high level of scalability. IBM technologies enrich this "open source" frame with analytical software, integrated software, platforms and tools with the extension.

IBM InfoSphere BigInsights Quick Start Edition is a free technology that allows new solutions on how to efficiently convert large and complex amount of data, [7] by combining the Apache Hadoop (including MapReduce and Hadoop Distributed File Systems) with a unique IBM technologies such as Big SQL, text analysis and BigSheets. To install the IBM InfoSphere BigInsights Quick Start Edition, it is necessary to fulfill several conditions. As for the hardware prerequisites for starting and installation, a minimum of 4GB of RAM, 40GB of hard disk space, and two processors are required. The operating system by which BigInsights works is Linux. IBM BigInsights technology is suitable for the analysis of data from the Twitter social network.

The aim is to create an application that will, using this technology, successfully analyze Twitter data and thus help modernize the educational system in high schools. With this in mind, several studies have been made on how to implement this improvement. Studies have shown that most students use social networks on a daily basis, and regarding the fact that social networks offer a large amount of updated information that can be used for educational purposes, it was concluded that the best solution was to create an application that could analyze data obtained from social networks, in order to facilitate the search and therefore help teachers to improve the topicality of their content.

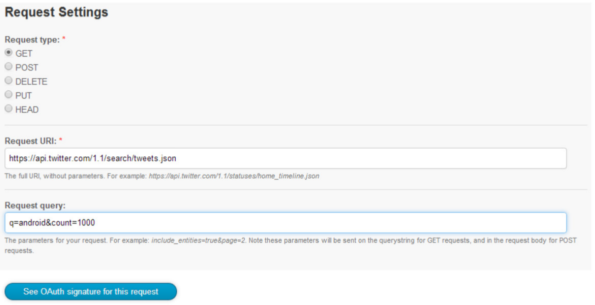
5. PROCESS OF CREATING A FILE FOR ANALYSIS

Twitter maintains an open platform that supports millions of people around the world who share the discovery of what is happening in real time. Twitter wants to empower its partners and customers to build a valuable information network, by allowing a free creation of various applications for different purposes.

Twitter social network for free creation and development of various Twitter applications requires the creation of a Developer's account. This makes it possible to successfully analyse the data and that what the Twitter API v1.1 is for. In its first version, Twitter API v1.0 used to bring back data to JSON format by using Ajax calls. However, now a new version of Twitter API v1.1 supports a variety of methods and data search, retrieving data in the following formats: XML, JSON, RSS and Atom. The new version also requires user authentication with built-in time limit for each code used when accessing the API, so now OAuth authentication and verification are being used.

The next step is to create the application, put personal data and location where the files for analysis will be stored. After that, you should write a request that will allow Twitter to know what needs to be searched. Requests may be different, depending on what information the user wants to receive. They can be simple, consisting of a few phrases or a hashtag. It is possible to search for tweet queries from certain locations, using a certain language. In addition, it is possible to search for tweets that are most popular or have multiple queries. In this case, the search is related to a word "android" and the query looks like this: **q=android&count=1000**

By this query, Twitter will generate last 1000 tweets containing the word "android" to the code which is necessary for the following step in order to create a file from which the data will be further analyzed.



Request Settings

Request type: *
● GET
○ POST
○ DELETE
○ PUT
○ HEAD

Request URI: *
https://api.twitter.com/1.1/search/tweets.json
The full URI, without parameters. For example: *https://api.twitter.com/1.1/statuses/home_timeline.json*

Request query:
q=android&count=1000
The parameters for your request. For example: *include_entities=true&page=2*. Note these parameters will be sent on the querystring for GET requests, and in the request body for POST requests.

Figure 1. Creating request for analysis inside TwitterAPI

After that, we receive the command called "cURL command" that needs to be used within a few minutes. This is the information which is obtained after the query generates code that is required in the first step.

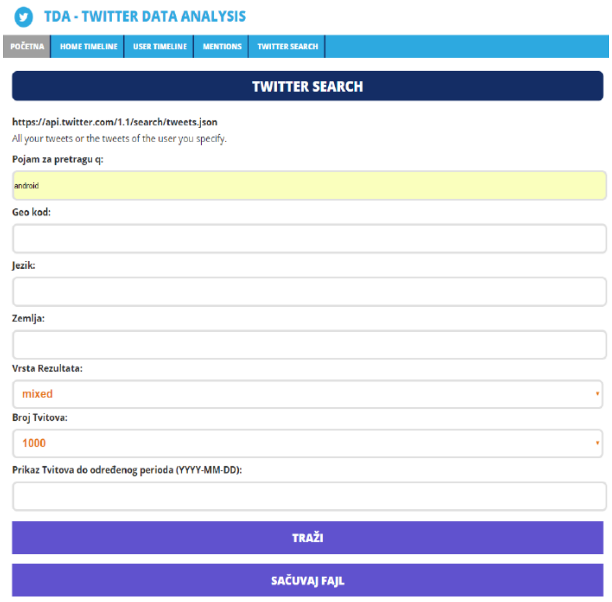
The biadmin documents will then locate a file called "and.json". "JavaScript Object Notation" (JSON) is a simple, text-based open standard designed for human-readable data exchange which is derived from the JavaScript language to represent simple data structures and associative arrays, *i.e.* objects. Despite its close ties with JavaScript, JSON is an independent language. It is often used for serialization and transfer of structured data over the network connections, especially between the server and web application, serving as an alternative to XML.

Loading data into IBM BigInsights is performed by typing an address http://BIVM:8080 in Web console. We need to find a file that has been created and upload it to the tab "Files" where they will be recognized as a text CSV file. In the following procedure we need to save the file as a Master workbook called "analysis of android," which is then automatically opened in BigSheet tool. BigSheets the analysis tool based on the browser which was originally developed by a group of new technologies of IBM (Emerging Technologies group).

Today, BigSheets is included in BigInsights in order to enable users to explore and analyze data. BigSheets represents tabular interface so that users can create, filter, combine, explore, and schematically show the data collected from various sources. BigInsights Web console includes a tab at the top to access BigSheets. BigSheets is a tool for the analysts provided by the IBM InfoSphere® BigInsights, and it is a platform based on open source Apache Hadoop project. BigSheets is the spreadsheet tool that is mainly used for analysis of social media and structured data collected through the application of samples obtained from a BigInsights.

When BigInsights collects and enters the data from different sources, BigSheets tool allows interactive exploration and managing the data from distributed data system. It is easy to form and manage data in BigSheets tool by using the built-in macro functions. It is also possible to create charts for a visual presentation of work and export analysis results into one of several popular output formats.

6. USE OF APPLICATION – TDAEDU IN EDUCATION



TDA - TWITTER DATA ANALYSIS

TWITTER SEARCH

https://api.twitter.com/1.1/search/tweets.json
All your tweets or the tweets of the user you specify.

Pojam za pretragu q: android
Geo kod:
Jezik:
Zemlja:
Vrsta Rezultata: mixed
Broj Tvitova: 1000
Prikaz Tvitova do određenog perioda (YYYY-MM-DD):

Figure 2. TDAedu application example

Data access allowed to Twitter users is displayed by the application in 17 columns. This is the meaning of these data:

header1 = (created_at) - Date and time the tweet was created; header2 = (id_str) - ID number; header3 = (geo) - Geographic location; header4 = (coordinates) - Coordinates; header5 = (location) - Location; header6 = (user.id_str) - ID of the user; header7 = (user.name) - Username; Header8 = (user.screen_name) - Twitter Username; header9 = (user.location) - The location of the user; Header10 = (user.description) - A brief description of the user; header11 = (user.url) - Link to the user; header12 = (user.followers_count) - The number of followers; header13 = (user.friends_count) - The number of "friends" that a user follows; header14 = (retweet_count) – The number of tweets; header15 = (FAVORITE_COUNT) - The number of favorites; header16 = (lang) - Language used by a user; header17 = (text) - Complete tweet in a form of text.

According to all these data, it is possible to perform several types of virtualizations:

The most commonly used language in tweets

The first example shows the total percentage of language coverage - the language mostly used in writing the tweets that are being analyzed. The application displays the column header 16 which is the column showing in which language the tweet is written, and the Total column that uses the COUNT function to calculate the number of the same languages from the column header16. Moreover, it is possible to make a virtual display using the chart where you can see that among the last 1000 tweets almost two-thirds were written in the Japanese language, followed by English, and German, *etc*. By using this analysis, the teacher can see which languages are mostly used to talk about the word from our query. In this case, we made a search for the word "android". After the analysis, a lecturer can tell students to create an android application in the Japanese language, in order to have better results on the market. On the other hand, they can visit most popular Japanese websites for android programming, since this analysis shows that this topic is very popular in Japan.

Influential people

Social media have gained great popularity among marketing teams, and Twitter represents an effective tool for a company and helps draw attention to their products. Twitter facilitates the involvement of users and direct communication with them, and in return users can provide the "word of mouth" marketing to the company by talking about their products.

This type of analysis could be of interest for students studying marketing and trading. A teacher instructs students to find influential users, and show them the potential utilization of users for marketing purposes.

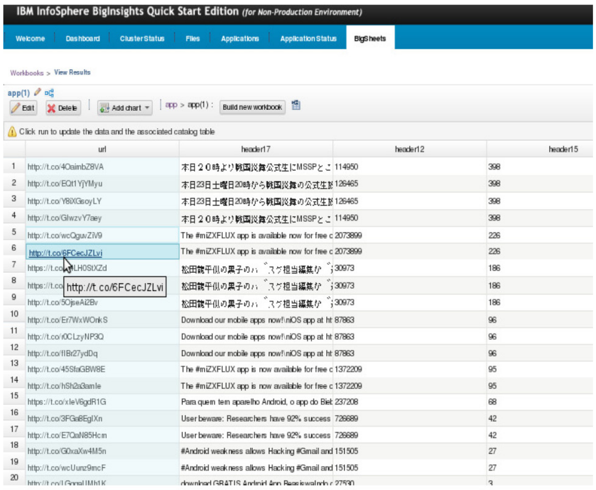
In addition to the analysis of certain concepts in marketing to perceive the reaction of users to the specific product, it is possible to detect a new advertising idea. Marketing departments can be more efficient by selecting who they want to reach to, and it can done by discovering the most influential person on Twitter, because by retweeting the messages can go much further than the followers of the person who originally sent the tweet. Bearing this in mind, it is necessary to try to engage users whose posts tend to generate a lot of retweets. Since Twitter keeps track of all retweets, it is possible to reach the target group through the analysis of Twitter data.

Showing links

In order to use Twitter analysis for school subjects related to information technologies, the example of request about android applications has been made. Right before the lecture, a professor who teaches android programming analyses Twitter data with students and obtains information that are trending and are connected to that area, and therefore adapt the lecture to the trending information.

In order to obtain good insight into the most popular android applications, the application sorts out the tweets by the most influential users, those who have the largest number of followers, meaning these users are always the first to publish information on current events in the world. After sorting the users, the application will use the URL function from tweets in the column 17 (column showing the entire tweet) to show only the links to most updated blogs or go directly to updated applications. There is also a display of the column 17, which contains a complete tweet to have a better insight into the link that is displayed. The application then inserts the column with the number of followers to sort the links by most influential users, and at the end, there is the column 15, which shows the number of users who liked the tweet content. With this display it is possible to see the links that are most preferred by the users.

The application provides the ability to quickly open a link, so that users do not have to wait for the link to open in a new window. Specifically in this case, after the first 5 links open, we could already get the information about currently most popular android applications. The most popular tweet included a link to the application released on that day, which is called Hacking Gmail App and which was posted on the "The hacker news" blog.



Click run to update the data and the associated catalog table

	url	header17	header12	header15
1	http://t.co/4OaimbZ8VA	本日２０時より戦国炎舞公式生にMSSPとこ	114950	398
2	http://t.co/EQt1YjYMyu	本日23日土曜日20時から戦国炎舞の公式生放	126465	398
3	http://t.co/Y8iXGsoyLY	本日23日土曜日20時から戦国炎舞の公式生放	126465	398
4	http://t.co/GlwzvY7aey	本日２０時より戦国炎舞公式生にMSSPとこ	114950	398
5	http://t.co/wcQguvZiV9	The #miZXFLUX app is available now for free c	2073899	226
6	http://t.co/6FCecJZLvi	The #miZXFLUX app is available now for free c	2073899	226
7	https://t.co LH0StXZd	松田龍平似の黒子のハ゛スケ担当編集か゛;	30973	186
8	https://t.co http://t.co/6FCecJZLvi	松田龍平似の黒子のハ゛スケ担当編集か゛;	30973	186
9	http://t.co/5OjseAi2Bv	松田龍平似の黒子のハ゛スケ担当編集か゛;	30973	186
10	http://t.co/Er7WxWOnkS	Download our mobile apps now!\niOS app at ht	87863	96
11	http://t.co/r0CLzyNP3Q	Download our mobile apps now!\niOS app at ht	87863	96
12	http://t.co/fIBr27ydDq	Download our mobile apps now!\niOS app at ht	87863	96
13	http://t.co/45SfaGBW8E	The #miZXFLUX app is now available for free c	1372209	95
14	http://t.co/hSh2a3amle	The #miZXFLUX app is now available for free c	1372209	95
15	https://t.co/xleV6gdR1G	Para quem tem aparelho Android, o app do Biet	237208	68
16	http://t.co/3FGa8EgIXn	User beware: Researchers have 92% success	726689	42
17	http://t.co/E7QaN85Hcm	User beware: Researchers have 92% success	726689	42
18	http://t.co/G0xaXw4M5n	#Android weakness allows Hacking #Gmail and	151505	27
19	http://t.co/wcUunz9mcF	#Android weakness allows Hacking #Gmail and	151505	27
20	http://t.co/LGoqaUMh1K	download GRATIS Android App Beasiswalndo c	27530	3

Figure 3. Sorted links from the most popular tweets

Tweets analysis from student groups

One of the ways Twitter data can be used is for the analysis of student tweets that use a certain hashtag, in a conversation related to this subject. To successfully apply this analysis, it is required to make an agreement with a group of students (in this case the subject "information technologies in business") so that students can frequently discuss the subject on Twitter, during the preparation for the test or exam. The agreement is that after each set of questions related to the subject, or after each statement, assistance or answer to someone's question, they post a twe*et al*ong with a hashtag #bpsitb (Belgrade Business School - information technology in business), as to be easily accessible. If you search for #bpsitb entry, the application will display a window with the most commonly used words in those tweets.

In order to have the best analysis of the tweets with hashtag #bpsitb, it is necessary to count the number of words that were commonly used, and based on those results we could see what was mostly written about and what could be a potential problem for students. The applications WordCount, which is located within the IBM BigInsights, manages the collection of text files and gives back the total number of words located in that file. After tweets analysis, the application will show a picture with the number of words, most commonly used, being increased and vice versa. The results obtained from the analysis can be used to focus on what caused the most problems to students while preparing for the exam, and to make it clear for future generations, in order to avoid the same problem, that they need to pay more attention to that part of studying material.

7. CONCLUSION AND FUTURE WORK

Based on previous experiences of scientific research in this area, as well as the research in this paper, it can be concluded that the implemented methods and techniques for the analysis of social networks contribute to the improvement of learning outcomes and interaction between students and professors. Electronic education and Distance learning in the future will be an integral part of any education system. Efficiency, effectiveness and competitiveness of educational institutions will depend on the ability to adapt to the demands of students, and implement actions that will improve the learning performance. The results of this research offer the possibility for further work in this area, primarily in the area of upgrading and expanding the proposed solutions for the application of analysis of data from social networks; in creating and developing e-learning courses where content would be created according to data on student behavior and activities, both at the university and on the social networks. This collected data would be processed in the system and applied in real time. The main goal of future research would be to develop applications for creating adaptive courses which would not require additional skills from a teacher. Creating modern applications for analysis could execute queries offering more options which would also provide possibilities for advanced visualization.

REFERENCES

[1] Paul C. Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, James Giles - „Harness the Power of Big data“, McGrawHill, 2012

[2] Xin Chen „Web-Based Tool for Collaborative Social Media Data Analysis“ - 978-0-7695-5114-2/13 © 2013 IEEE

[3] Seyed-Ali Bahrainian „Sentiment Analysis and Summarization of Twitter Data“ – 978-0-7695-5096-1/13 © 2013 IEEE

[4] Vadim Kagan and Andrew Stevens, Sentimetrix V.S. Subrahmanian, University of Maryland „Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election“ 1541-1672/15/$31.00 © 2015 IEEE

[5] Yeongjun Kim „Smartphone Response System Using Twitter to Enable Effective Interaction and Improve Engagement in Large Classrooms“ - 0018-9359 © 2014 IEEE

[6] https://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/bi_qse.html

[7] David John Hughes “A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage” - Manchester Business School East, The University of Manchester