ISSN 2347 - 3983

International Journal of Emerging Trends in Engineering Research (IJETER), Vol. 3 No.1, Pages : 19 – 24 (2015)
Special Issue of ICAET 2015 - Held on February 23, 2015, Cochin, India
http://warse.org/pdfs/2015/icaet2015sp04.pdf

Big Data Analysis, Algorithms and Applications: A Survey

Purti Jain[1], Dr. Shruti Kohli[2]
[1]Department of Computer Science
Birla Institute of Technology, Mesra
Ranchi, India
purti.jain@yahoo.com

[2]Department of Computer Science
Birla Institute of Technology, Mesra
Ranchi, India
shruti@bitmesra.ac.in

Abstract—This paper is written for researchers seeking to analyze the wealth of information available from social media. Since the social networking aeon is rising, there has been an escalation in user generated content.Every day millions of people share their opinions on micro blogging sites as it is one of the shortest and simplest way to express one's thought. Analyzing social media, in particular Twitter feeds for sentiment analysis (SA), has become a major research field due to availability of application programming interfaces (APIs) provided by Twitter, Facebook and News services. The study focuses on analyzing the probable solutions pointed out by several eminent researchers for sentiment analysis.

Keywords— Big Data, Sentiment Analysis, Twitter, Healthcare

INTRODUCTION

Not many people enjoy talking about health and fitness, especially when it concerns their own health problems. But with the advent of various social media avenues like Twitter, Facebook, etc., people have started sharing in the public domain expressing feelings, reporting and updating activities and whereabouts. The data made public on social media sites, such as Twitter, provide a plethora of information about individuals, groups, and neighborhoods.

Twitter is an online social networking medium, popular since October 2006, where registered users share or post messages under 140 characters known as tweets. It is a social medium for people to communicate and stay connected through the exchange of quick, frequent messages [1]. They share information, news and opinions with followers and seek knowledge and expertise through public tweets. Tweets have been composed from daily conversations, updates, and critiques on news, movies, politics, life, etc.

An incredible quote by Eric Schmidt Google CEO[2, 4] "There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing...People aren't ready for the technology revolution that's going to happen to them." All this accumulation results in continuous generation of an immense volume of data, which if analyzed intelligently, can be of extreme value, as it can give us a variety of critical information to make smarter decisions as shown in fig 1.

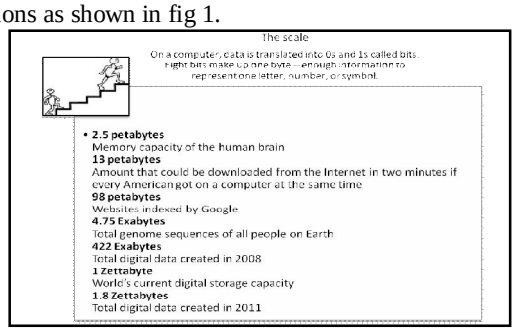

Fig 1: The level of data generated so far [3]

BIG DATA: A GROWING TORRENT

Definition

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit into the structures of traditional database architectures. Big data refers to the large, complex volume of data captured from multiple sources that cannot be processed using traditional methods and holds massive cognizance which can to be explored by great efforts.

According to McKinsey & Co [5, 6] Big Data is "the next frontier for innovation, competition and productivity". The impact of Big Data gives not only a huge potential for competition and growth for individual companies, but the right use of Big Data also can increase productivity, innovation, and competitiveness for entire sectors and economies. Big data—large pools of data that can be captured, communicated, aggregated, stored, and analyzed—is now part of every sector and function of the global economy [6]. Fig 2 is an evidence that data is being generated every second and thus it has become essential to build tools that can handle such a large amount of data in an effective and efficient way. Big Data is a pool of large-sized datasets to capture, store, search, share, transfer, analyze, and visualize related information or data within an acceptable elapsed time [7].

ISSN 2347 - 3983

International Journal of Emerging Trends in Engineering Research (IJETER), Vol. 3 No.1, Pages : 19 – 24 (2015)
Special Issue of ICAET 2015 - Held on February 23, 2015, Cochin, India
http://warse.org/pdfs/2015/icaet2015sp04.pdf

Make up of Big Data

Big data is composed of three types of data: Structured, Unstructured and Semi-structured.

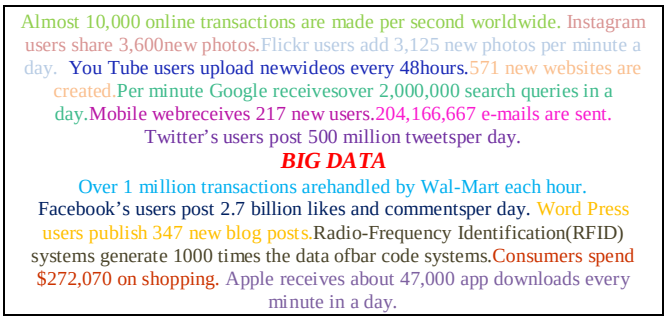

Fig 2: Information about the popular organizations that hold Big Data [7]

1. *Structured data*

Structured data is an organized data in a predefined format. The formatted data resides in fixed fields within a record or file and has entities and attributes mapped. Structured data is used to query and report against predetermined data types [7].
Some sources of structured data include:

- Relational databases
- Flat files in record format
- Multidimensional databases
- Legacy databases

2. *Unstructured data*

Unstructured data is a set of data with a complex structure that might or might not have a repeating pattern [7]. It:

- Consists typically of meta data
- Comprises inconsistent data
- Consists of data in different formats such as e-mails, text, audio, video, or image files

Some sources for unstructured data include:

- Text Internal to an Organization: Comprises documents, logs, survey results, and e-mails within the organization's database and data warehouse.
- Data from Social Media: Comprises data from social media platforms including YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- Mobile Data: Comprises data such as text messages and location information.

3. *Semi-structured data*

Semi-structured data, also known as schema-less or self-describing structure, refers to a form of structured data that contains tags or mark up elements in order to separate semantic elements and generate hierarchies of records and fields in the given data. Such type of data does not follow proper structure of data models as in relation databases [7].
Some sources for semi-structured data include:

- Database systems
- File systems like Web data and bibliographic data
- Data exchange formats like scientific data

Big Data Dimensions

Gartner analyst Doug Laney introduced the 3Vs concept in 2001 MetaGroup research publication, 3D data management: Controlling data volume, variety and velocity.

- *Volume*

Volume is the amount of data generated by organizations or individuals [7, 8]. Big data means an ample amount of data-terabytes or even petabytes (1000 terabytes). So it is the most immediate challenge of big data, as it requires scalable storage and support for complex, distributed queries across multiple data sources.

- *Velocity*

Velocity is the frequency and speed at which data is generated, captured and shared. Consumers as well as businesses now generate more data and in much shorter cycles, from hours, minutes, seconds down to milliseconds [8]. Velocity also includes quick business decisions such as in stock exchange.

- *Variety*

Variety is the proliferation of new data types including those from social, machine and mobile sources. New types include content, location or geo-spatial, hardware data points, log data, machine data, metrics, mobile, physical data points, process, radio frequency identification (RFID), search, sentiment, streaming data, social, text and web. Also, variety includes traditional unstructured clinical data (i.e., free text) [8].

- *Value*

Value is a way of exploiting the large amount of data to extract insight and draw valuable conclusions out of it which can help an organization, enterprise to grow and expand. Big Data technologies are now seen as enablers to create or capture value from otherwise not fully exploited data. In essence, the challenge is to find a way to transform raw data into information that has value, either internally, or for making a business out of it [6].

- *Veracity*

Veracity isn't just about data quality, it's about data understandability, accuracy, fidelity or truthfulness.Veracity refers to trust ability of data. Now days, there are large number of spams occurring on internet so it has become very essential to distinguish between genuine data and noise. Veracity is a process of filtering the noise and providing the right data for usage. Veracity refers to the trust into the data and is to some extent the result of data velocity and variety.

Blending data from multiple sources

Big data can be said to comprise six different categories, or streams, of information [9]:

1. Web and social media data: Clickstream and interaction data from social media such as Facebook, Twitter, LinkedIn, and blogs. It can also include health plan websites, smart phone apps, etc.
2. Machine-to-machine data: Readings from sensors, meters, and other devices.
3. Big transaction data: Health care claims and other billing records increasingly available in semi-structured and unstructured formats.

ISSN 2347 - 3983

International Journal of Emerging Trends in Engineering Research (IJETER), Vol. 3 No.1, Pages : 19 – 24 (2015)
Special Issue of ICAET 2015 - Held on February 23, 2015, Cochin, India
http://warse.org/pdfs/2015/icaet2015sp04.pdf

4. Biometric data:Fingerprints, genetics, handwriting, retinal scans, and similar types of data. This would also include X-rays and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data.
5. Government agency sources: such as census data.
6. Human-generated data: Unstructured and semi-structured data such as electronic medical records (EMRs), physicians' notes, email, and paper documents [10].

Big Data Applications in different domains

1. Transportation:Big Data has transformed transportation by providing improved traffic information and autonomous features.
2. Education:Big Data has transformed the modern day education process through innovative approaches for teachers to analyze the students' ability to comprehend and thus, impart education effectively in accordance with each student's needs. The analysis is done by studying responses to questions, the time taken when attempting those questions, and other behavioral signs in the classroom.
3. Travel:With the help of Big Data, the airline companies can track customers who fly between specific routes so that they can make the right cross-sell and up-sell offers and even shape their inventories. Some airlines also apply analytics to pricing, inventory, and advertising to improve customer experiences, which leads to more customer satisfaction, and hence, more business.
4. Government:The study and analysis of available data is allowing governments to make informed decisions for fraud management, discover unknown threats, ensure security of global supply chain by monitoring global cargo traffic, use budgets more judiciously, analyze risks, and lots more.
5. Healthcare:In healthcare, physicians can make use of Big Data to determine the best clinical protocols that will ensure the best health outcome for patients at specific locations. The pharmaceutical and medical device companies use Big Data to improve their research and development decisions, while health insurance companies use it to determine patient-specific treatment therapy modes that promise the best results. Big Data also helps researchers to spot and work toward eliminating healthcare-related challenges before they become real problems [7].

LITERATURE REVIEW

Methodology

The six articles studied presented in this survey are summarized in Table 1. The first, second and third column describes the title, author and year respectively. The objectives of the articles are illustrated in fourth column. The task is divided into categories: Feature Selection as FS, Sentiment Analysis as SA and Sentiment Classification as SC in fifth column. The sixth column specifies whether the article is domain oriented by means of Yes/No. The seventh column shows the algorithms used. The eighth column specifies whether the article uses SA techniques for general analysis of text(G) or solves binary classification problem(Positive/Negative). The ninth and tenth column illustrates the data scope and the benchmark dataset respectively. The last column specifies if any other languages other than English are analysed.

Mathiesen et al. [11] analyze online human behaviour on social media like Twitter by analyzing the occurrence and co-occurrence frequency of keywords in user posts. International brand name like Starbucks is used for which the occurrence rate is estimated and persistence or memory effect of brand-users is observed. The positive and negative sentiments of humans are analysed and their affect on a particular brand is analysed. The co-occurrence rate describes how individual brands are linked to each other and a corresponding relationship network is constructed. Fluctuations in successive tweets are also considered to analyse human behaviour.

Santos and Matos [12] uses Naïve Bayes classifier to identify tweets mentioning flu or flu-like illness or symptoms. Further multiple linear regression model is used to estimate the health-monitoring data from Influenzanet project. This work is presented in Portuguese language and additionally it was investigated whether the predictive model created can be applied to data from the subsequent flu season.

Bahrainian and Dengel [13] shows product-based sentiment summarization of multi-documents with the purpose to inform users about pros and cons of various products. Different algorithms for SA polarity detection and sentiment summarization are compared and it is found that hybrid polarity detection system outperforms and could be an advantage over other methods when used as a part of a sentiment summarization system.

Xiang et al. [14] focuses on sentiment classification of Twitter messages to measure DOC of the Twitter users. In order to achieve this goal, first personal tweets are separated automatically from news and then negative tweets are identified. Multinomial Naïve Bayes achieved best results over others.

Lima and de Castro [15] proposes an automatic sentiment classifier for Twitter messages to reduce human intervention, complexity and cost of the entire process. To assess the performance of proposed system tweets related to Brazilian TV show were captured in 24h interval and it shows that the proposed technique achieves an average accuracy of 90% using the hybrid approach.

Lampos and Cristianini [16] presents a method for tracking the flu epidemic in the UK by using the contents of Twitter for 24 weeks during H1N1 flu pandemic. They compare the flu score with data from HPA, obtaining an average which is greater than 95%.

Tools available for analysis

Different tools are available in the market for the purpose of analyzing data which mainly include: R, Matlab, SAS, SPSS, Excel and Stata. R is day by day gaining popularity among the other tools as it is freely available and easy to use and learn. A comparison among the various analytical tools is depicted in Table 2 and a comparative graph is shown in fig 3.

ISSN 2347 - 3983

International Journal of Emerging Trends in Engineering Research (IJETER), Vol. 3 No.1, Pages : 19 – 24 (2015)
Special Issue of ICAET 2015 - Held on February 23, 2015, Cochin, India
http://warse.org/pdfs/2015/icaet2015sp04.pdf

Table 1: Articles related to Sentiment Analysis

Title	Author	Year	Objective	Task	Domain Oriented	Algorithm Used	Polarity	Data Scope	Data Set/Source	Other Language
Statistics of co-occurring keywords in confined text messages on Twitter [11]	J. Mathiesen, L. Angheluta, M.H. Jensen	2014	To analyze the occurrence and co-occurrence frequency of keywords in user posts on Twitter.	FS, Statistical	Y	Chi-square	Pos/Neg	International brand names like Starbucks, Apple, IBM etc product review	Twitter	English
Analysing Twitter and web queries for flu trend prediction [12]	José Carlos Santos,Sérgio Matos	2014	Infodemiology study to estimate and predict the incidence rate of influenza like illness in Portugal.	FS, SC	Y	NB, SVM, Random Forest, Decision Tree, Nearest Neighbor, Linear regression model	Pos/Neg	Tweets, Search engine logs related to flu	Twitter, Web search queries, Influenzanet project	Portuguese
Sentiment Analysis and Summarization of Twitter Data [13]	Seyed-Ali Bahrainian, Andreas Dengel	2013	To compare various SA polarity detection algorithms and introduce a new aspect-based sentiment summarization	SA	Y	Supervised PD(SVM, NB, ME) Unsupervised PD, Hybrid PD, Aspect detection	G	Product-based (review of smartphones)	Twitter	English
Monitoring Public Health Concerns Using Twitter Sentiment Classifications [14]	Xiang Ji, Soon Ae Chun, James Geller	2013	To measure DOC[1] about disease outbreak, the location and speed with which they appear. An early warning tool ESMOS[2] is developed to monitor DOC caused by various epidemics.	SC	Y	Clue-based algorithm, NB, MNB, SVM	G	Tweets related to listeria, measles, swine flu and tuberculosis	Twitter, Phirehose library, Google Map	English
Automatic Sentiment Analysis of Twitter Messages [15]	Ana C. E. S. Lima, Leandro N. de Castro	2012	Automatic sentiment analysis for Twitter messages that reduces human intervention, complexity and cost of the whole process.	SA	Y	NB, 3 approach for automatic classification(emoticon-based, word-based, hybrid)	Pos/Neg	Tweets related to TV show “Agora é Tarde” from Brazilian station	Twitter, Twitter4J library	Portuguese
Tracking the flu pandemic by monitoring the Social Web [16]	VasileiosLampos, NelloCristianini	2010	To measure the prevalence of disease in a population by analysing the contents of social networking tools.	FS, Statistical	Y	Porter’s algorithm, LASSO LARS algorithm	G	Tweets related to H1N1 flu pandemic in UK	Twitter, Health Protection Agency	English

1 DOC Degree of Concern
2 ESMOS Epidemic Sentiment Monitoring System
SVM: Support Vector Machines
NB: Naïve Bayes
ME: Maximum Entropy
MNB: Multinomial Naïve Bayes
PD: Polarity Detection
LARS: Least Angle Regression

ISSN 2347 - 3983

International Journal of Emerging Trends in Engineering Research (IJETER), Vol. 3 No.1, Pages : 19 – 24 (2015)
Special Issue of ICAET 2015 - Held on February 23, 2015, Cochin, India
http://warse.org/pdfs/2015/icaet2015sp04.pdf

Table 2: Data Analysis tools

Features	Data Analysis Tools					
	R	**SAS**	**SPSS**	**Stata**	**Excel**	**Matlab**
Developed By	Ross Ihaka and Robert Gentleman		Norman H. Nie, Dale H. Bent, and C. Hadlai Hull			Cleve Moler
Place	University of Auckland	North Carolina State University	IBM Corporation	Statacorp	Microsoft	MathWorks
Year	1993	1976	1968	1985		1984
Stands for	Rumored to stand for its original creators	Statistical Analysis System	IBM Statistical Package for the Social Sciences	Name based on word combination of "statistics and data"	Microsoft's spreadsheet program	MATrixLABoratory
Open source	Yes	No	No	No	No	No
Data Extension	*.Rdata	*.sas7bcat, *.sas#bcat, *.xpt (xport files)	*.sav, *.por (portable file)	*.dta	*.xls	*.m,*.mat, *.dat
User Interface	Programming	Programming	Mostly point-and-click	Programming/point-and-click	Point-and-click	Programming
Learning Curve	Steep	Steep	Gradual	Moderate	Flat/Gradual	Steep
Data Manipulation	Very Strong	Very Strong	Moderate	Strong	Weak/Moderate	Very Strong
Statistical Analysis	Very Broad Scope High Versatility	Very Broad Scope High Versatility	Moderate Scope Low Versatility	Broad Scope Medium Versatility	Moderate scope	Limited Scope High Versatility
Graphics	Excellent	Very Good	Good	Good	Very good	Excellent
Specialties	Packages for Graphics, Web Scraping, Machine Learning & Predictive Modeling	Large Datasets, Reporting, Password Encryption & Components for Specific Fields	Custom Tables, ANOVA & Multivariate Analysis	Panel Data, Survey Data Analysis & Multiple Imputation	Calculation, graphing tools, pivot tables	Simulations, Multidimensional Data, Image & Signal Processing
Advantages	Library support, visualization	Large datasets	Like Stata	Easy statistical analysis	Easy, visual, flexible	Elegant matrix support, visualization
Disadvantages	Steep learning curve	Expensive, outdated programming language	More expensive and worse	Large datasets	Large datasets	Expensive, incomplete statistics support
Ease of Learning	Not easy	Easy	Very Easy	Very Easy	Easiest	Easy
Operating System(Mac/Windows)	Both	Windows	Both	Both	Both	Both
Typical Users	Finance, Statistician, mathematician, scientist	Business, Government, Statistician, financial specialist	Business data analyst, nonprofessional statistician, and financial analysis specialist	Science	Business, Financial personnel and even those without technical background	Engineer, statistician
Official Website	http://www.r-project.org/	http://www.sas.com	http://www-01.ibm.com/software/analytics/spss/	http://www.stata.com/	http://office.microsoft.com/en-us/excel/	http://www.mathworks.com/products/matlab/

1) NYU Data Services, https://sites.google.com/a/nyu.edu/statistical-software-guide/summary
2) http://www.slideshare.net/ILRI-Jmaru/module-4-data-analysis
3) Getting Started in R~Stata Notes on Exploring Data, Oscar Torres-Reyna otorres@princeton.edu
4) Exploring Data and Descriptive Statistics (using R) Oscar Torres-Reyna Data Consultant otorres@princeton.edu

ISSN 2347 - 3983

International Journal of Emerging Trends in Engineering Research (IJETER), Vol. 3 No.1, Pages : 19 – 24 (2015)
Special Issue of ICAET 2015 - Held on February 23, 2015, Cochin, India
http://warse.org/pdfs/2015/icaet2015sp04.pdf

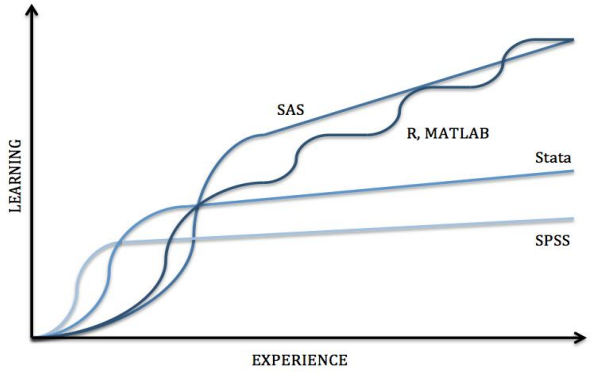

Fig 3: Illustrating the popularity of R among other analytical tools [17]

CONCLUSION

In this survey paper, we investigated the influence of social media on the life of people. The positive negative influence can be used to depict a real-world phenomena or can be associated with real-world phenomena.

Perhaps, the biggest concern among the public is regardinghealth and therefore it has become necessary to analyse the opinions of the users posted on the networking sites so that an upcoming epidemic can be subsided. As there is an increase in the usage of social media among users, through the tweets generated we can depict the correlation between the popular hospitals and their patients and through the sentiments generated we can create a framework for hospital recommendation system.

REFERENCES

[1] Bae,Y. and Lee, H.,"Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Twitterers", Journal of the American Society for Information Science and Technology, vol. 63, issue 12, December 2012, pp. 2521–2535.

[2] Derek Mead, "Big Data Explained Brilliantly in One Short Video", April 23, 2013. URL: http://motherboard.vice.com/blog/big-data-explained-brilliantly-in-one-short-video

[3] Popular Science and Katie Peek, "The Promise of Big Data", Harvard Public Health, 2012, pp. 15-19, 42-43, URL: http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/

[4] Baker, H., Daschle, T., Dole, B. and Mitchell, G., "A Policy Forum on the use of Big Data in Health Care", Bipartisan Policy Center (BPC), 2013.

[5] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, May 2011, URL: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

[6] Networked European Software and Services Initiative (NESSI) Big Data White Paper, "Big Data A New World of Opportunities", December, 2012.

[7] Kogent Learning Solutions, Bill Franks, Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman, JorisMeys, Andrie de Vries, Mark Gardener, Dr. Murray Logan, Michael J. Crawley, Deborah J. Rumsey, Johannes Ledolter, StephaneTuffery, Dean Abbott, Wrox Certified Big Data Analyst (WCBDA), "Introducing Big Data Analytics and Predictive Modeling", Wiley Publishers.

[8] Canada Health Infoway (Infoway), "Emerging Technology Series: Big Data Analytics in Health", 2013

[9] Cottle, M., Hoover, W., Kanwal, S., Kohn, M.,Strome, T. and Treister, N.W., "Transforming Health Care Through Big Data", Institute for Health Technology Transformation(IHTT), 2013.

[10] Soares, S., IBM Data Management, 2012, June 13, URL: http://ibmdatamag.com/2012/06/a-framework-that-focuses-on-thedata-in-big-data-governance/

[11] Mathiesen, J., Angheluta, L., Jensen, M. H., "Statistics of co-occurring keywords in confined text messages on Twitter", The European Physical Journal Special Topics, vol. 223, issue 9, September 2014, pp. 1849-1858.

[12] Santos, J. C., Matos, S., "Analysing Twitter and web queries for flu trend prediction", *Theoretical Biology and Medical Modelling* 2014, vol. **11**(Suppl 1):S6, May 2014, pp. 1-11.

[13] Bahrainian, S.-A., Dengel, A. "Sentiment Analysis and Summarization of Twitter Data", 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE), December 2013, pp. 227 – 234.

[14] Ji, X., Chun, S. A., Geller, J., "Monitoring Public Health Concerns Using Twitter Sentiment Classifications", 2013 IEEE International Conference on Healthcare Informatics (ICHI), September 2013, pp. 335 – 344.

[15] Lima, A.C.E.S., de Castro, L.N., "Automatic Sentiment Analysis of Twitter Messages", 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), November 2012, pp. 52-57.

[16] Lampos, V., Cristianini, N., "Tracking the flu pandemic by monitoring the Social Web", 2010 2nd International Workshop on Cognitive Information Processing (CIP), 2010, pp. 411-416.

[17] NYU Data Services, URL: https://sites.google.com/a/nyu.edu/statistical-software-guide/home.