International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 8, Issue 6 [June 2019] PP: 10-16

Big Data Analytics in Healthcare: Social & Biological Data

Reza Shokri[1], İlker Kocabaş[2]
[1,2]*International Computer Institute, Ege University, İzmir, Turkey*

Abstract:- Big data technology bring new opportunity to discovering, predicting and decision making in biomedical and healthcare domain. Understanding the health impact of big data and its value help people and government make better decision and drive performance in physical and mental health. Although big data analytics has the potential to provide useful insight in healthcare, however, without the efficient tools and talent in such a domain it is impossible to stay on top of necessary information. Leveraging technology such as parallel processing and cloud computing lead to better data quality in cost and time effective manner. This study provides a baseline to assess the proliferation of the use of big data in two healthcare sub-domains: (i) Bioinformatics, and (ii) public healthcare.

Keywords: - *Big data, bioinformatics, healthcare, social media*

I. INTRODUCTION

Biological and social data analytics demands the latest advancement in computational intelligence, data mining, machine learning and statistical methodologies. Applying big data platforms and analytics in the realm of natural science not only has the potential to change lives, but also to save them [1]. Advancement of big data analysis offers a cost-effective opportunity in critical decision making area such as health care, security, economic, crime, natural disaster and recourse management [2, 3]. Bioinformatics is the algorithm and application of information technology to the management of biological big data. In the biomedical informatics domain, big data is a new paradigm and an ecosystem that transforms case-based studies to large-scale, data-driven research [4].

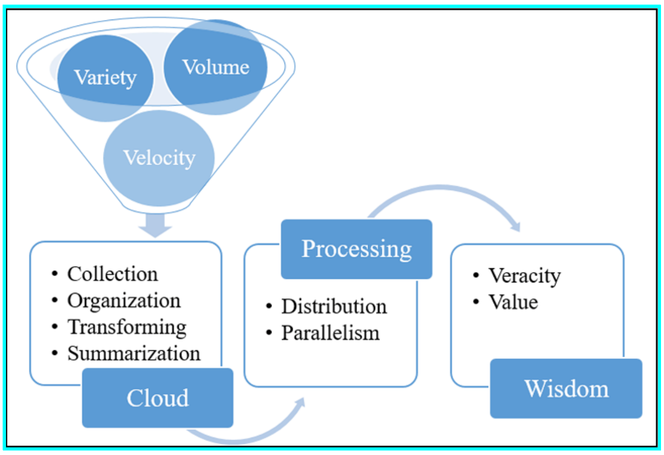

Fig.1. Big data processing scenario

The majority of social network mining is extraction valuable, user-generated data. Recently, many public health organizations, such as the United States Centers for Disease Control and Prevention (CDC), the US Food and Drug Administration (FDA), the World Health Organization (WHO), the American Public Health Association (APHA), have established a social network presence. Although, social media mining has the positive impact on public health and safety, however, traditional text processing technique and tools are not more suitable for this masterwork. Because of dynamic nature of big data, preparing data for analysis is a labor task. Real time, short length and abbreviation form of content in social media bring new challenge in text classification, information retrieval, data quality, and data privacy issue. To address this challenge we can leverage big data concept.The concept of big data analytic is successful application and talent in different domain, which provide pattern in order to make human life more comfortable, agile and secure. Some of the most significant potential of big data comes from the bridge among resource and data streams. For instance, in the healthcare system, the outcome of combination multiple data comes from patient behavior, clinic, service and cost, medical research and even weather report is more desirable and cost effective. Big data can expose people’s hidden behavioral patterns and even shed light on their intentions [5].Big data is not just about the volume of data. Bigger data is not always better data and just because it is accessible does not make it ethical [6]. From the best knowledge of the authors, the common definition has three V's: (1) Volume, (2) Variety and (3) Velocity of data (Fig.1). Furthermore, the main aim of big data is (4) Veracity and finally big data analyzing lead to big (5) Value. While the first three V's are more technical and related to data engineering such as collection, summarizing, storage or transforming, the last two V's focused on data quality and data mining. The real value of big data is wisdom which obtains by data analytics, knowledge extraction, and decision making. Big data analytics are an algorithm and set of advanced technologies designed to work with large volume of data. Performing analytics on large volumes of data requires efficient methods to store, filter, transform, and retrieve the heterogeneous data. Therefore, the big data solution must be able to offer deep analytic, high agility, massive scalability with low cost and latency [7].

The rest of this study is organized as follows: We begin with the introduction of social media analytics. In Section 3 we share recent research in some key areas of public health in which social media mining has been especially popular, with an emphasis on how these areas are evolving to increase public health impact. Section 4 focuses on big data in Bioinformatics. Section 5 includes a brief introduction of the big data analytics frameworks and introduce MapReduce paradigm shortly. Finally, we draw conclusion.

II. SOCIAL MEDIA ANALYTICS

Online technology as a new form of resource comes to change resource, food, water, education, lifestyle, and healthcare. Social networks shrinking our world, thus mining social media has a potentially in preventing disease. Such an impact comes from the sharing of new personal experience related to lifestyle or the environment, and also helps people to successfully modify their risk behaviors. Since the Internet has been an integrated part of our lives, it can be part of new fundamental resources. In fact, people work as a data agent and creating data while using online technology.Social networks are a combination of online people that works as a data agent. This organic sensor network sense and collect huge volumes of activities which reported by individual agent. Moreover, in the shadow side, social life is captured by surrounding smart devices. This data is useful to government agencies as well as private companies to support decision making in areas ranging from law enforcement to social services and business pattern [8].

Technology as a new resource is continuously evolving to meet market demands, rather than safeguarding the status quo [9]. Data generally created from communication device, consumer transaction, and specifically from online behaviors. Big data is going to change lifestyles more and less in the whole world. Social media has become one of the most relevant data sources for big data [10]. Observing personal behavior in social media as a new opportunity is used in monitoring and improving population health. Such a transform has more impact on rich and industrial countries. As more behaviors are recorded with social technologies, the set of knowledge includes: domain knowledge, programming skills, computing and communication infrastructure are required for exploring the effects of social interactions on health outcomes.

One of the challenges that differentiate social media analysis from existing tasks in normal form of text and graph mining is the data life time. Social media data is the temporal perspective. Thus, texts are usually time-sensitive and the networks evolve over time. Furthermore, the integration and convergence of all data related to healthcare which comes from different resource is labor task.

2.1 Social Networks and Healthcare

Healthcare data come from different resources. In the past, heath data were obtained from clinical result and patient records usually in structured format. Nowadays, the online population creates a vast organic sensor network composed of individuals reporting on their activities, their social interactions, and the events around them [11]. People share and discuss their views and opinions, and many share their health-related information both in general-purpose social media and in health-related social networks [12]. In some sense, social text streams are the sensors of the real world [13].Sharing data make them more effective. If health is to become a fundamental element of our everyday lives, we must responsible about the social healthcare and try to strengthen the connections between our health care and the other components of health in our communities. Social media is virtual space make feel better while sharing experiences. For example, social networks are allowing patients diagnosed with depression to share their thoughts and connect with other patients and doctors [14]. Monitoring and mining these networks help us to predict and prevent abnormal behaviors.

2.2 Social Media and Text Analytics Challenges

Major part of big data is unstructured and semi-structured data sets which come from social media. Social media are media for social interaction, and use of electronic and Internet tools for the purpose of sharing and discussing information and experiences with other human beings in more efficient ways [15]. Among the various formats of data exchanged in social media, text plays an important role [16]. Thus, text analytics provide an effective way to meet users' diverse information needs. For example, in business intelligence, the most common use of social media analytics is the opinion mining. The opinion mining or customer sentiment is a type of natural language processing for tracking the public attitude about a particular product or service. Text analytics techniques include approach and algorithms such as data mining (DM), information retrieval (IR), natural language processing (NLP), and machine learning (ML). In this section, we briefly review state-of-the-art text analytics.Social media as a new resource in NLP tasks like text summarization, named entity recognition, and relation extraction, bring new opportunity and challenges for public health monitoring and surveillance. The nature of the microblogging, as a form of social media, forces user to use the short format of text in chatting and blogging. Real time, short length and abbreviation form of content in social media bring new challenge in text classification and information retrieval. The proposed solution has to be responsible about gathering, managing, analyzing, and visualization of huge volume of semi-structure and unstructured data sets. Unfortunately, traditional text processing technique and tools are not more suitable for this masterwork. In this section, we aim to address these gaps. Nowadays, the popularity of big data transfer information among asymmetric resource. Thus, in the first step we need for convergence and integrate all of them before using. Convergence means fusion or interaction of different resources to create one new target or domain. The difficulty is that, sometime same term has different manning and sometime the different terms point to the same things. Big data should have a comprehensive view not separately to draw a roadmap for augmented future. In addition, the traditional resource and tools require convergence and adapting with the new big data paradigms leverage for massive data processing. Although, it is clear that there is no single system that would be most suitable for every needs. However, MapReduce (see section 5) approach and well-known big data open source frameworks such as Apache Hadoop is a good address to these challenges.

2.3 Text Processing Methodology

The majority of text processing is converting text to indexing terms which make the best use of resources. The term text processing refers to the automatic approach and tools for modifying or manipulation of electronic text. The token is a sequence of characters in document. Tokenization of raw text is a standard pre-processing step for many NLP tasks. The base form of traditional tokenization includes (i) elimination of stop words: filtering out words with very low discrimination value, and (ii) stemming: conflate syntactic variations of words.*Text presentation: In text mining, information retrieval, and machine learning, a vector space model (VSM) is the most common way to transform textural information into sparse features. The* $TF \times IDF$ *(TF: the number of term occurrences in a document, IDF: Inverse document frequency) is a well know method to evaluate word importance in document. The formula is given in Eq. 1, where* DF *is the number of documents in the collection for a given term and N is the total number of documents in the collection.*

$$TF \times IDF = TF \times log_2\left(\frac{N}{DF} + 1\right), (1)$$

Knowledge discovery: It is well known that words have the history about the other. Information content of a message is dependent on the receiver's prior knowledge as well as on the message itself. Classification and clustering methods are commonly used to knowledge discovery. For example, cosine similarity widely used for measuring similarity between two V1 and V2 messages.

$$\text{Similarity (V1, V2) =cosine } \frac{V1 \cdot V2}{||V1|||V2||} \qquad , (2)$$

III. RELATED WORKS AND RESEARCH TOPIC IN SOCIAL MEDIA MINING

How social media can be harnessed to best achieve public health outcomes is a topic of much research in the public health community. In this section we briefly review research area and goal of automatic methods for the collection, extraction, analysis, and validation of big social data for public health. Understanding the health impact of big data and its value help people and government make better decision and drive performance in physical and mental health. Decision making based on collective shared experience of patients in social media is a new non-negligible resource. Research topics for health monitoring and surveillance include but not limited:

- Event detection- extract patterns of the abnormal behaviors of groups or individual [17, 18, 19].
- Disease control and prevention- surveillance and monitoring disease such as Influenza [20], Ebola [21] and foodborne.
- Health monitoring- including drug interactions [22], drug safety [23] substance abuse and smoking [24], and depressing [25]

IV. BIG DATA IN BIOINFORMATICS

Big data sources are no longer limited to search-engine logs and indexes. Bioinformatics research is another big data domain which is increasing dramatically in last decade. Bioinformatics is the field of molecular biology that used advantage of the computer science for biological data processing.

4.1 History

Although in 1977 that the Bioinformatics expression was used by Dutch theoretical biologist “Paulien Hogeweg”, however the seed of this field return to, elucidation of the structure of DNA (1953), determined the amino acid sequence of the first peptide hormone (1955), and determining first three-dimension structure of protein (1963) [26]. The key challenge in Bioinformatics data analytics related to size and ease of molecular purification. As a simple example, chromosomal DNA molecules, containing many millions of nucleotides. A single human’s DNA contains around 3 billion base pairs representing approximately 100 GB of data. It was clear that communication delays could be eliminated if servers held copies of data locally. Hence, distributed form of national biological database established in 1988 under an umbrella of European Molecular Biology network (EMBnet), which every node has last updated information. Today International Nucleotide Sequence Database (INSD) has three primary databases: DNA Data Bank of Japan (DDBJ), National Center of Biotechnology information (GenBank) in USA, and EMBL in Europe, which are repositories for nucleotide sequence data from all organisms. All three databases have daily synchronization. They also collaborate with Sequence Read Archive (SRA), which archives raw reads from high-throughput sequencing instruments.

4.2 Challenges in Biological Dataset

Bioinformatics was applied in the creation and maintenance of a database to store and used biological information in research, include: (i) Genome Database (structured database and sequences database), (ii) Protein Database (structural database), and (iii) Complex Database (Protein Nucleic acid Complex Database). Undoubtedly processing and management such a complex domain needs excellent computing facility.

Regardless of storing a massive volume of data, a much bigger challenge is about processing such a data it in a timely manner. The truth is, the results produced by one tool are not always in a format that can be used by the next tool in a workflow. Therefore, tools and data need to be more than close than ever been. Since Bioinformatics is geographical distributed data, therefore cloud computing technology has been used for data storing and processing. There is no need to move data out of cloud for processing. Thus, the EBI is building a cloud-based infrastructure called Helix Nebula.

4.3 Bioinformatics Cloud and Big Data Service Models

The old generation of tools was not built for the massive volume and a variety of data that we face today. Parallel programming and computing capability are necessary for analyzing massive Bioinformatics data. The cloud provides a scalable and cost efficient solution to the big data analytics by coupling data and software into the cloud and delivering them as services. Beside state-of-the-art cloud services (e.g. SaaS, PaaS, IaaS), we can define three sublevels of service models in the big data era: (i) Data as a service (DaaS): providing customers a way to mine their own, (ii) Information as a Service (IaaS): providing insights based on the analysis of processed data, and (iii) Analytics as a Service (AaaS): providing higher-level answers to specific questions.

4.4 Big Data Migration

Transferring vast volume of biological data to the cloud is a significant bottleneck in cloud computing. On the other hand, state of art TCP is not efficient in today’s networks since it just support single path. Per-packet load balancing causes reordering and TCP thinks there was packet loss. Since routing and transportation mechanism have a significant effort on throughput, providing redundant paths between different end-system improves efficient transfer in lack of the bottleneck link. To address this problem, resource polling via multipath routing significantly improves transfer time (Fig. 2). Multipath TCP (MPTCP) is a TCP extension, which enables single data flow to be separated across multiple paths in order to maximize network utilization [27]. The resource pooling mechanism is making a collection of resources behave like a single pooled resource [28]. Making a collection of path behave like a single pooled path leads to reliability, flexibility and efficiency. If a path dies, it is just dropped from the pool. Such a capability strongly depends on the end system behaves that could spread their load across multiple paths. Main advanced of such a solution includes:

- Maximum network utilization
- Load balancing through traffic engineering
- Robustness through dynamic alternative routing

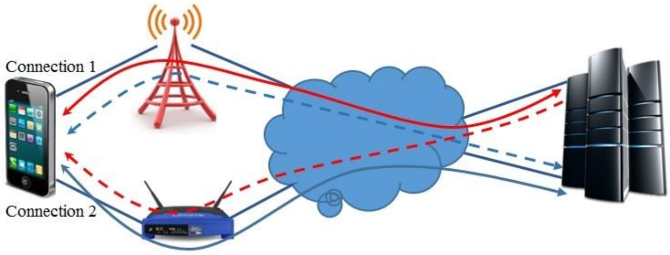

Fig.2. MPTCP and resource pooling

4.5 Data Quality

The DaaS provides data quality. Quality has a dynamic nature and conformance to valid requirements. As shown in Fig. 3, data quality has many dimensions like accuracy, completeness, consistency, and auditability. Technically the aim of data quality is delivering right data to the right customer at the right time. Data quality is measured to determine data veracity whether or not data can be used as a basis for decision making. New big data analytics platforms like Hadoop integrated with infrastructures and application provide data quality.

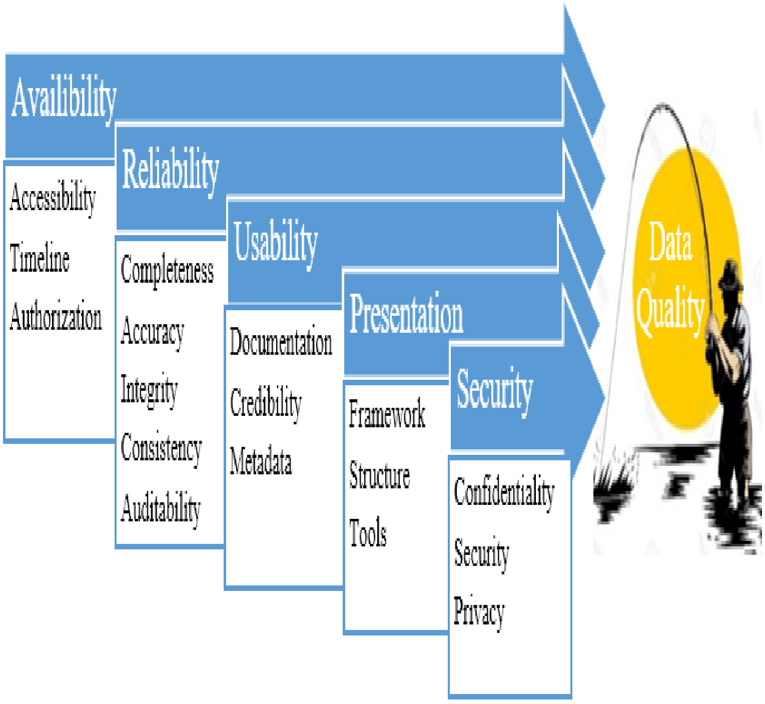

Fig.3. Big data quality schema

Data quality has been investigated, focusing especially on data as represented in the relational model, traditionally adopted in the database. Decision making and action based on data or events which comes from online resources, is not more trustworthy without cleaning and integration. The aim of building a data warehouse is to have an integrated, single source of data that can be used to make decisions. Data quality as an internal part of data integration play important role in decision making.The majority of data quality is veracity. Toward value, there is need of trusted data source, advanced technology, flexible and secure system. The value of social and biological data analytics comes in three forms: enhance heath knowledge, predict disease and improve treatment, qualify and augment human life style.

V. BIG DATA ANALYTICS FRAMEWORKS

Biological data and social media cover billions pieces of data, however, such a data is not self-explanatory. The mythology of big data is offering a higher form of analytic paradigm that can generate wisdom in new way that was previously impossible. Integration such a system has to cope with lack of:

- Availability: data should be accessible in authorized form.
- Reliability: heterogeneous and multi relational and dynamic nature of social networks makes data gathering and analyzing difficult.
- Usability: include metadata and documentation.
- Presentation: including hardware, database, operating system, programing language, monitoring dashboard.
- Security and privacy**:** Confidentially of personal information

Supervised, unsupervised, and hybrid machine learning approaches are the most widely used tools for big data analytics. Cloud base and clustered approach such as MapReduce and Apache Hadoop is used for developing novel solutions on massive data sets such as web analytics, relational data analytics, machine learning, data mining, and real-time analytics. MapReduce is a good platform for unstructured data set, because the upfront investment for using MapReduce is small: no schema, no index, no normalization, and NoSQL.

The high scalability of the MapReduce paradigm allows for massively parallel and distributed execution over a large number of computing nodes [29]. The key to efficient MapReduce processing is that, wherever possible, data is processed locally- on the slave node where it's stored. Such a distributed architecture generally configured on commodity servers. MapReduce paradigm works based on two main functions- Map and Reduce. The Map function splits text into words and assign a (key, value) pair for each of them. While the Reduce function responsible for aggregation. Technically, the whole process completes in four stages as follows:

- Split: input files are split into HDFS blocks and the blocks are replicated.
- Map: slave node applies the Map () function on the local data and save the output in temporary storage.
- Shuffle: output of Map phase is forwarded to related group based on specific key.
- Reduce: slave nodes process each group of output in parallel mode and forward final result to output.

The existing tools for many bioinformatics problems are still not adequate for big data. The highly efficient fault tolerant nature, flexibility, platform independent, and easily scaling over a large set of commodity servers are the main characteristic of MapReduce. In addition, a few Apache Hadoop and other MapReduce base implementations developed to address the MapReduce performance (Table I).

TABLE I. Table MapReduce base implementation projects

Category	Project	Description
Standard	Hadoop	Since does not support index, so inefficient
Distributed Programming & File system	HDFS	Hadoop distribution file system
	Storm	Processing large-volumes of high-velocity data
	Pig	High-level language for analysis programs.
Big Graph Models	Cassandra	Manage structured/unstructured data set
	Giraph	A framework for analyzing social graph.
Document	MongoDB	Popular *NoSQL* solution with high availability
Machine Learning (Algorithm & Library)	Spark	Framework for machine learning algorithm.
	Mahout	Provides tools to find patterns in big data
	MLlib	Provides libraries for classification, clustering
	MLbase	Provides tools for ML applications
Application & Business Intelligence	Nutch	Extensible interfaces for custom implements
	Pentaho	Data integration, data mining, and visualization
	SparkR	A light-weight front-end to use R on spark

VI. CONCLUSION

We are at the beginning of a big data revolution which needs a big transformation in the use of tools, talent and resources. Big data as information extraction paradigms is a new approach of thinking about technology, economy, culture and healthcare. However, big data is not enough to discover the truth. Thus, we need to find new ways to interact with big data and find the pattern of the universe. Undoubtedly, there will be a need for experts' person with trained skills to effectively discover and uses the hidden pattern in big data. The future could be better if we change our behavior and improve public awareness as a human kind which pay attention more positive to challenge as a new opportunity. We can and should to use big data to understand today and design tomorrow. This opportunity prepares augmented reality -valuable, safe and secure life- for global citizenship.

REFERENCES

[1] A. O'Driscoll, J. Daugelaite, R.D. Sleator," *Big data', Hadoop and cloud computing in genomics"* Journal of Biomedical Informatics 46 (2013) 774–781

[2] R. Tinati, S. Harford, L. Carr, and C. Pope," Big data: methodological challenge and approaches for sociological analysis," Sociology 48(1),2014, pp.23-39

[3] M. Helbert." Big Data for Development: From Information- to Knowledge Societies, 2013

[4] J. Luo, M. Wu, D. Gopukumar, Y. Zhao,"Big Data Application in Biomedical Research and Health Care: A Literature Review", Biomed Inform Insights. 8: 1–10. Published online 2016 Jan 19

[5] R. Abbas, "The Social Implications of Location-Based Services: An Observational Study of Users," J. Location- Based Services, vol. 5, 2011, pp.156-181.

[6] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, N. Christakis," Tastes, ties, and time: a new social network dataset using Facebook.com," Social Networks, vol. 30, 2008, pp 330-342.

[7] R.S. Kalan, İ. Kocabaş," Adaptive tools and technology in big data analytics," Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 3 Issue 1, January – 2016

[8] K. Michael, Keith W. Miller," Big Data: New opportunity and new challenge," Computer, 46 (6), 2013, pp. 22-24.

[9] K. Iwai,"A contribution to the evolutionary theory of innovation," imitation and growth. J. Econ. Behav. Organ. 43 (2), 2010, pp.167–198.

[10] B-O. Gema, J.J Jason, D. Camacho," Social big data: Recent achievements and new challenges," Information Fusion, 2015

[11] H. Kautz," Senior member presentation: Data mining social media for public health applications," 23rd International Joint Conference on Artificial Intelligence *(IJCAI 2013)*, Beijing, China, 2013.

[12] M.J. Paul, A. Sarker, J.S. Brownstein, A. Nikfarjam, M. Scotch, K.L. Smith, Gr. Gonzalez," Social media mining for public health monitoring and surveillance," Pacific Symposium on Biocomputing 2016

[13] Q. Zhao, P. Mitra, and B. Chen. "Temporal and information flow based event detection from social text streams". In Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, 2007,pp.1501-1506

[14] A. Akay, A. Dragomir, B-E. Erlandsson,"Mining social media big data for health" The advantages of harvesting big social media data, IEEE PLUSE, 2015

[15] S. Moturu," Quantifying the trustworthiness of user-generated social media content. PhD thesis," Arizona State Uni., 2009.

[16] X. Hu, H. Liu," Text analytics in social media, "Mining Text Data, Springer, 2012, pp.385–414.

[17] Y.L. Yu, "A survey on social media anomaly detection," ACM Trans. Knowl. Discov. Data. 9, 4, Article 39 (2014), 18 pages.

[18] H. Sayyadi, M. Hurst, and A. Maykov," Event detection and tracking in social streams," Proceedings of the Third International ICWSM Conference, 2009

[19] J. Weng, Y. Yao, E. Leonardi, F. Lee," Event detection in Twitter, HP Laboratories, External Posting Date: July 06, 2011

[20] A. Culotta,"Towards detecting influenza epidemics by analyzing Twitter messages," ACM Workshop on Soc.Med. Analytics, 2010.

[21] M. Odlum, "How Twitter can support early warning systems in Ebola outbreak surveillance," Annual Meeting of the American Public Health Association, 2015.

[22] R. B. Correia, L. Li and L.M. Rocha, "Monitoring potential drug interactions via network analysis of Instagram user timelines," Pacific symposium on Biocomputing (PSB), 2016.

[23] D. Hand, "Principles of data mining," Drug Safety, vol. 30, 2007, pp. 621–622

[24] M. J. Paul, A. Sarker, J.S. Brownstein, A. Nikfarjam, M. Scotch, K. L. Smith, G Gonzalez, "Social media mining for public health: monitoring and surveillance," Pacific Symposium on Biocomputing (PSB), 2016.

[25] M. Park, C. Cha, M. Cha," Depressive moods of users portrayed in twitter," In *Proc. ACM SIGKDD* Workshop on Healthcare Informatics (HI-KDD), 2012

[26] T.K. Attwood, A. Gisel, N-E. Eriksson, E. Bongcam-Rudloff," Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective, Bioinformatics" 2011

[27] R. Stewart, Stream Control Transmission Protocol. RFC 4960 (Sep 2007)

[28] D. Wischik, M. Handley, M. Bagnulo Braun," The resource pooling principle," ACM SIGCOMM Computer Communication Review, Volume 38, Issue 5, 2008, pp.47-52

[29] T. Eytan, J. Benabio, V. Golla, R. Parikh, and S. Stein, "Social Media and the health system," The Permanente Journal/ Winter 2011/ Volume 15 No.1