JCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

A Survey on Realms and Applications of Social Media Data Analysis

Aishwarya Rajamani[1*], Alpha Vijayan[2]

[1] Department of Computer Science, New Horizon College of Engineering, Bangalore, India
[2] Department of Computer Science, New Horizon College of Engineering, Bangalore, India

**Corresponding Author: aishraja12@gmail.com, Tel.: +919980577928*

Available online at: www.ijcseonline.org

Accepted: 18/Nov/2018, Published: 30/Nov/2018

***Abstract*—** The information era witnesses the creation of multimedia data, transfers and transactions in the order of millions. This data by virtue of their formats comes in varying sizes and differing temporal characteristics. The wealth of information carries potential both in terms of explicit content that is expressed and the implicit or hidden content. Processing the former is quite developed while the procedures and applications of working with the implicit knowledge are growing steadily. This paper aims to present a range of techniques from recent works pertaining to the processing and applicability of such data. The purpose of the survey is to bring to light the specific methods of social media data analysis in a concise and organized manner. Specifically, natural language processing, topic modelling, sentiment analysis and affective analysis have been identified as the overarching heads taken up by several recent researches. Finally, some observations pertaining to social media data analysis identified from several works are enlisted.

***Keywords*—** Data Mining, Machine Learning, Natural Language Processing, Social Media

I. INTRODUCTION

The onslaught of social networks being a driving force in helping ease the process of expressing oneself and the issues one cares about has resulted in massive amounts of data being generated. The impending need for analyzing such data is not only advantageous but will also prove to be useful to the users of social media. The advantages of data analysis are: utilization of the knowledge gained for the purpose of better advertising; understanding the pulse of product trends amongst the market populace; leveraging the insights to strategize business, etc.

A study by Hootsuite, a popular social media management portal, in the month of January, 2018 shows that the total internet user population amounts to over 4 billion and that the population that actively makes use of social media amounts to over 3 billion. With the users and their data burgeoning thus, this data can be of great value if novel approaches of their utilization are researched.

This paper is an attempt to provide a comprehensive overview of prevalent methods of analysing social media data. The contents have been organized into the following sections: Section II describes a summarized version of the procedural methodology adopted in social media data analysis; Section III discusses the sources and types of raw data from social networks; following that in Section IV, various realms of analysis on aforementioned data are discussed; Section V enlists in brief a few application scenarios for the utilization of such information; finally the observations made from studying the analysis of data sourced online are listed in section VI. Section VII briefly states the concluding remarks.

II. RELATED WORK

A systematic plan for analysis is required for any kind of analysis. However, it is of prime importance with respect to data sourced from online portals as it is usually unstructured. In industries and in research, mining of structured data has always followed the Knowledge Discovery (KDD) steps. The following list has been identified from several works to be the general procedure adopted in various custom ways according to the unstructured-data problem domain:

- Understanding of Problem Domain and Overall Plan
- Choosing a platform for analysis based on planned data and the capacity of platform to handle the data
- Selection of variables in relation to problem domain
- Data acquisition, i.e., collecting required raw-data
- Filtering of unwanted data
- Algorithmic/Methodical Analysis of acquired data
- Creation of reports and visualization
- Plan of Action for future

III. TYPES AND SOURCES OF INFORMATION

In relation to gathering of information from the social realm, the sources can be categorized as given by A.Gandomi & H.Murtaza [1]. They are as follows:

- Social networking portals such as *Facebook* and *LinkedIn*
- Blog portals such as *Blogger* and *WordPress*
- Microblogs such as *Twitter* and *Tumblr*
- Social news portals such as *Digg* and *Reddit*
- Social bookmarking portals such as *Delicious* and *Stumble Upon*
- Media sharing portals such as *Instagram* and *YouTube*
- Encyclopedia-like portals such as *Wikipedia* and *Wikihow*
- Question-and-answer portals such as *Answers* and *Quora*
- Review portals such as *Yelp*, *TripAdvisor*

The types of data, as can be assessed from usage of the above portals, include: text in formats of html, xml, json, txt; images in popular encodings like jpg, png, gif; videos in encodings such as mp4, mpeg, AVI, FLV, and so on. The usefulness of the types of data however depends upon various factors characteristic to each type of data. For example, a higher resolution picture can reveal a lot more in terms of insight that can be gained.

IV. REALMS AND METHODS OF ANALYSIS

A. Natural Language Processing

Natural Language Processing or NLP as it is popularly called is a class of techniques under the Artificial Intelligence (AI) umbrella. While AI consists of systems that can perform intelligently, NLP consists of systems that can understand the language of humans. The majority of data that is generated online consists of informal, conversational style of text and audio. Thus, it is meaningful to process such data obtained from the medium of internet portals using NLP.

It is useful to discuss a recent NLP based study by G.Coppersmith et al, in the field of mental health [2]. While information specific to hospital visits is recorded in medical institutions the state of a person, especially one affected with a mind disorder has to be monitored at all times. To this end, the aforementioned study uses what is called as linguistic signals produced from the data obtained from patient’s chat communications. A classification approach is taken to label the emotions and aggregation is performed. To factor out the overbearing of volume of per day communications, aggregation of emotion related data has been emphasized in this study.

B. Sentiment Analysis

Sentiment analysis is a popular research area currently given the trend of ricocheting comments and opinions on platforms such as *Twitter*.

Consider the example of hateful expressions online. A study by A.Schmidt & M.Wiegand presents a survey of NLP techniques which encompasses a wide variety of procedures such as word generalization and sentiment analysis [3]. Specifically, analysis of the expressed sentiment is useful in classification as the generated proportion of positive, negative or other such indicative labels can be used as features.

In a recent study by N.Mamgain et al, microblog data on the top colleges in India were analyzed using several techniques such as: lexicon based analysis by making use of Nielson provided lexicon that assigns a valence value to sentiments; Naïve Bayes approach by selecting features by means of Chi-Squared test; Support Vector Machine in addition to comparing of polynomial, Gaussian Radial Basis, sigmoid and linear kernels is performed; finally, the Neural network approach by employing multi-layer perceptrons [4]. Of all techniques, their results state that the neural network approach is highly accurate and that despite its simplicity Naïve Bayes performs very well too.

Native languages represent a psychological haven for all members of humanity. In a survey work by C. Nanda and M. Dua, there features several instances of native languages being analyzed ranging from the Indian dialect of Hindi, to dialects of countries like Nepal and Tibet [5].

C. Affective Analysis

Affective analysis is a type of computing which combines the disciplines of computer science, psychology and science of cognition. Its potential is seen from the artificial intelligence point of view where, for instance, a system is able to replicate the emotion of empathy or perhaps respond to an environment with a result that stems from some psychological-cognitive basis.

While processing of natural language is one option, neural networks especially the recent developments in the area, are conducive to short length data such as *Tweets*. A neural network in the most basic level consists of perceptrons that take weighted inputs and returns a result based on a function. Neural network technologies have come a long way and the current trends see the use of deep neural networks and convolutional neural networks.

A recent research work published by S.Poria et al, describes a novel framework that uses multiple algorithms and data sources for an efficient affective analysis [6]. Multi-modal analyses of sentiments are performed, i.e. videos, audios and text are used for analysis. Videos annotated with labels indicating emotions are employed. The annotations are given distinct value encodings to separate the emotions and thus the emotion feature from videos is obtained. Audio features such as Pitch and Voice intensity are obtained and for textual data a hybrid technique combining Support Vector Machines and Convolutional Neural networks is used. The proposed architecture which they term as Decision Level Fusion Framework is the defining feature of the paper. Feature set fusion is performed on two types of input and the final decision takes as input the former as well as a third mode of input. An important result of the research is that extracting emotions from videos is a harder problem than just resolving the polarities.

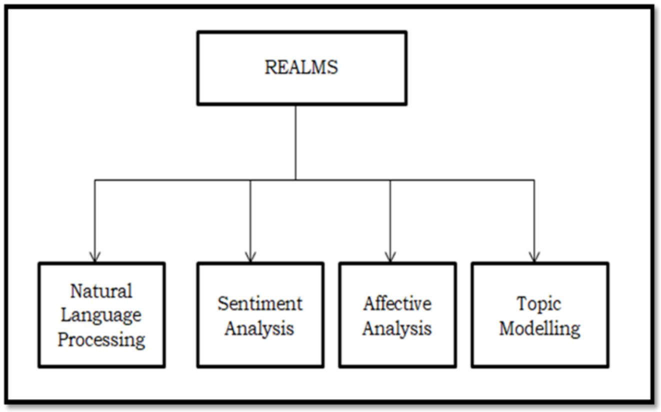

Figure 1: Realms of Analysis

D. *Topic Modelling*

A topic model can be defined as a statistical construct of hidden topics in a piece of literature or a set of documents. This method can be used to decipher hidden patterns based on an overarching topic. For example, documents pertaining to a particular subject may contain references to topic quite unrelated to the subject on the surface.

A research work by K.H.Lim and others, proposes a spatial based Latent Dirichlet Allocation algorithm that utilizes the location tag that the microblog Twitter provides [7]. Although only a less percentage of microblog data are tagged with the geological information it is still useful to perform analysis. The paper finds that the S-LDA approach finds relevant topics and maps them to places that are related.

V. APPLICATIONS OF SOCIAL DATA ANALYSIS

A few recent applications are discussed in this section:

A. *Pharmacy and Medicine*

The pharma industry is ever at work in finding drugs and improving upon existing ones. Adverse Drug Reactions or ADR is a popular field of research in the pharma world. To monitor automatically these adversarial responses to medicines is of focal interest.

A paper by A.Sarker et al, describes a framework for detecting the reactions based on the responses of patients on social media [8]. The key challenges they identify are: while pharma related data can be obtained with ease from social platforms that are health oriented there is still a dearth of information from popular microblog platforms; additionally, the imbalance in collected data is a sore spot. The authors feel that more social media data related to the problem at hand will prove useful to researchers. However, some privacy questions are impending in this regard.

B. *Disaster and Crisis*

While some natural calamities can be precisely predicted others take their own course. It is however useful to learn from incidents for several reasons. Firstly, their characteristics can help avert the same situations in times thereafter. Secondly, during early hours of a proliferating event user generated content on social media can be useful in providing relief or in working out strategies to curb it. Micro-blogs can be of some help in such a study as reported by D.T.Nguyen et al [9]. In the field of image processing, a type of deep-feed-forward neural networks known as convolutional networks is found to work well. Dat Tien Nyugen et al, employ both non-CNN and CNN methods on crisis related data-sets. The former included techniques such as Support Vector Machines, Logistic Regression and Random Forest. The latter was implemented using Theano, which supports both normal CPUs and Graphics Processing Unit (GPUs). The results show that the latter approach is better than SVM approach.

C. *Tourism Enhancement*

Surveys are the traditional manner in which candidate preferences for assessment of subject or object of interest. With the populace engaged constantly to social portals by expressing interest or otherwise through text, images, etc., it can prove to be an alternative way to collect survey data. This has been explained in a paper by A.Hausmann et al, where the authors compare results of analysis performed on the visits to Kruger National Park located in South Africa [10]. Such an analysis is possible due to geo-tagging feature in services like *Instagram* and *Flickr*. This has been proposed as a cost-effective manner of surveying preferences. Also, the study finds different user groups for the different platforms and differing focus with each platform.

Another study by Z.A.Hamstead et al, finds correlation between usage of parks in New York and the characteristics as indicated by analyzing *Flickr* and *Twitter* data that contains geographical information [11]. Predictors of rate of visitation were first formed and tested post which modelling through regression is performed. The study interestingly finds that visitation is negatively correlated to green factor which is contrary to popular belief and other studies.

D. Business and Industries

A.S.Halibas, in their paper that discusses classification and clustering algorithms for microblog data, indicate that decision-making based on evidence can increment brand value [12]. They clearly outline the operators to be used for analysis with Rapid Miner software which can be acquired by businesses. For the purpose of quality management they suggest the use of customer feedback and reviews. Additionally, real time micro-blog data analysis is suggested as an area of future work as it requires lots of processing.

VI. Observations

The following are some challenges faced by researchers during analysis using social media sourced inputs:

- Micro-blogs containing geographical data are very few in comparison to the total amount of posts
- For nuanced and sensitive areas of application such as health care, data valid and worthy of research were hard to obtain
- Often several sources of data are required for analysis to yield good results
- The informal nature compounded with little or no structure calls for thorough pre-processing and NLP procedures
- Machine learning and statistical methods can be employed effectively and adapting the same to big data architectures can provide more efficiency in terms of speed and volume of processing

Additionally, it is observed that there is evident interest amongst the enthusiasts of unstructured media to formulate frameworks for easy, systematic and organized application to various domains. These are useful starting points when starting the development of such an application from the scratch.

For instance, there is a proposal for an architecture where the general procedure of analysis as described previously has been modularized into three units [13]. Each one takes care of a logical chunk of analysis. Firstly the initial acquiring of data forms one logical chunk. Following it are introduced the modularized chunks of its immediate preliminary processing and final analysis that will contribute to the results.

It is important to note how these attempts at creating frameworks aid in lending some form of systematism to the immensity that the combination of structured and unstructured data analysis poses

VII. Conclusion

Social Media data is a valuable asset in the field of analysis. This is made possible due three important factors: the growth of social portals and its usage; the development of analysis infrastructures in the form of frameworks in Python, etc.; the application programming interfaces provided by social portals. As the social media data grows and the analysis on the same increase, it is important to suggest that the privacy and security aspects of the population be preserved and individuals are away by a long shot from risk of breaches and invasions.

As a final remark, it is worth reiterating that this data can be used as creatively as the application scenarios discussed portray and can provide valuable insights for betterment of systems in place by lending to it an authentic people-centric approach.

References

[1] A.Gandomi, and H.Murtaza "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* Vol. 35, No. 2, pp.137-44, 2015.

[2] G.Coppersmith, C.Hilland, O.Frieder, and R.Leary. "Scalable Mental Health Analysis in the Clinical Whitespace via Natural Language Processing." *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Orlando, FL, pp. 393-396, 2017.

[3] A.Schmidt and M.Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing." *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, pp. 1-10, 2017.

[4] N.Mamgain, E.Mehta, A.Mittal, and G.Bhatt, "Sentiment Analysis of Top Colleges in India Using Twitter Data." *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, New Delhi, pp. 525-530, 2016.

[5] C. Nanda, M. Dua, "A Survey on Sentiment Analysis", International Journal of Scientific Research in Computer Sciences and Engineering, Vol 5, Issue 2, pp. 67-70, April, 2017.

[6] S.Poria, E.Cambria, A.Hussain, and G.B. Huang. "Towards an Intelligent Framework for Multimodal Affective Data Analysis." Neural Networks, Vol. 63, pp. 104-16, 2015.

[7] K.H.Lim, S.Karunasekera, A.Harwood, and L.Falzon. "Spatial-based Topic Modelling Using Wikidata Knowledge Base." *2017 IEEE International Conference on Big Data (Big Data),*Boston, MA, USA, pp. 2009-2018, 2017.

[8] A.Sarker, R.Ginn, A.Nikfarjam, K.O.Connor, K.Smith, S.Jayaraman, T.Upadhaya, and G.Gonzalez. "Utilizing Social Media Data for Pharmacovigilance: A Review." *Journal of Biomedical Informatics* 54 , pp.202-12, 2015.

[9] D.T.Nguyen, K.A.A.Mannai, S.Joty, H. Sajjad, M.Imran,P.Mitra, “Robust classification of crisis-related data on social networks using convolutional neural networks”, Proceedings of the 11th International Conference on Web and Social Media, ICWSM, Montreal, Canada, pp. 632-635, 2017.

[10] A.Hausmann, T.Toivonen, R.Slotow, H.Tenkanen, A.Moilanen, V.Heikinheimo, and E.D.Minin. "Social Media Data Can Be Used to Understand Tourists’ Preferences for Nature-Based Experiences in Protected Areas." *Conservation Letters* Vol. 11, Issue. 1, pp.1-10, 2017.

[11] Z.A.Hamstead, D.Fisher, R.T.Ilieva, S.A.Wood, T.Mcphearson, and P.Kremer. "Geolocated Social Media as a Rapid Indicator of Park Visitation and Equitable Park Access." *Computers, Environment and Urban Systems* Vol. 72, pp.38-50,2018.

[12] A.S.Halibas, A.S.Shaffi, and M.A.K.V.Mohamed. "Application of Text Classification and Clustering of Twitter Data for Business Analytics." *2018 Majan International Conference (MIC)*, Muscat, Oman, pp. 1-7, 2018.

[13] R.S.Shirsath, V.A.Desale, A.D.Potgantwar, “Big Data Analytical Architecture for Real-Time Applications”, International Journal of Scientific Research in Network Security and Communication, Vol 5, Issue 4, August, 2017.

Authors Profile

Aishwarya Rajamani pursued Bachelor of Engineering in Computer Science at M V Jayaraman College of Engineering, Bangalore, India between 2010 and 2014. She is currently pursuing her Master of Technology in Computer Science at New Horizon College of Engineering, Bangalore, India. Her technical areas of interest are Cloud Computing, Data Science and Machine Learning.

Alpha Vijayan completed Bachelor of Technology in Computer Science at College of Engineering, Chengannur, Kerala in the year 2000. She holds a Master of Engineering degree in Computer Science from Jerusalem College of Engineering, Chennai. Currently, she is pursuing PhD. in Computer Science and Engineering discipline from Jain Deemed to be University, Bangalore. She has a total of 16.8 years of teaching experience.