Ambient Science, 2019: Vol. 06(1); 01-04
DOI:10.21276/ambi.2019.06.1.ga01

Published by: National Cave Research and Protection Organization, India

ambient SCIENCE
Vol. 06(1):01-04
Year 2019

GENERAL ARTICLE

A Perspective on Text Classification, Clustering, and Named-entity Recognition in Social Media

Kia Jahanbin, Fereshte Rahmanian*, Vahid Rahmanian, Masihollah Shakeri, Heshmatollah Shakeri, Zhila Rahmanian1, Abdolreza Sotoodeh Jahromi

Research Center for Social Determinants of Health, Jahrom University of Medical Sciences, Jahrom, Iran

Study Area: Jahrom, Iran
Coordinates: 28°30'00"N; 53°33'38"E

Key words: Big social media, Text mining, Data mining, Event prediction

Abstract

Nowadays, the use of social media is increasing continuously and these media have become a significant source for unstructured data, including business, governments, and public health. The growing confidence in social networks calls for data mining techniques that are likely to assist reforming the unstructured data and place them within an organized pattern. We are presenting here a brief review on text mining in social networks; focusing on text classification, text clustering, and Named-entity recognition (NER). Firstly, discussed on knowledge discovery in texts and big data, later explained big data in social media and describe its important aspects and finally, reviewed several articles that show the different perspective of using this knowledge in another science.

Introduction:

Knowledge Discovery is the process of extracting implied effective, and fresh valuable information from data (Fayyad *et al.*, 1996; Frawley *et al.*, 1992). Data Mining has used specific algorithms for inventive patterns from data. KDD objects at discovering concealed patterns and connections in the data.

Knowledge discovery from the text (KDT or Text Mining) was first introduced by Feldman & Dagan (1995), refers to the process of extracting high quality of information from structured; such as RDBMS data (Akbari *et al.*, 2018; Chen *et al.*, 1996), semi-structured; such as XML and JSON, and unstructured text resources; such as word documents, videos, and images (Pouriyeh & Doroodchi, 2009; Pouriyeh *et al.*, 2010).

After the emergence of social media, an enormous amount of data started to generate. The researchers call this, the Big Data which has three key features, known by "3V": Volume, Variety, and Velocity. Some researchers add two other characteristics: Value and Veracity, and thus speak about "5V" (Jahanbin & Mehrjoo, 2017; Lomotey & Deters, 2014).

Mukkamala *et al.* (2014) and Nguyen *et al.* (2015) use interchangeably the terms Big Social Data and Social Big Data to refer the overall data created by social media. To demonstrate the huge amount of data generated by social media affirms that through Facebook, 10 million photos are uploaded every day. Marrin (2014) highlights the fact that more than 300 million tweets are sent to Twitter every day too, and 3000 photos are uploaded from Flickr every minute without forgetting the 153 million blogs posted daily. The exploitation of Social Big Data is very valuable in many fields such as sociology, psychology, politics and the commercial area (Farahbakhsh *et al.*, 2013; Kooshkaki *et al.*, 2015; Rabade et al., 2014).

Nevertheless, the Social Big Data are known to be very large, dynamic and unstructured. Therefore, it seems interesting and essential to proceed with the mining of social media to extract useful and hidden knowledge within these wide masses of data (Bello-Orgaz *et al.*, 2016).

The science of text mining can be examined in two aspects: applications and theoretical (in paper). Further, theoretical of text mining can be categorized as follows:

Text categorization or classification, text clustering, Named-entity recognition (or concept/entity extraction), production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

According to Mukkamala *et al.*, (2014) classification, social media mining presents six most representative research issues in the social media mining: i) community analysis or detection, ii) opinion mining and sentiment analysis, iii) recommendation, iv) analysis of influence, v) diffusion or propagation of information and vi) privacy.

In this article, we tried to review the latest articles in text mining in social media and we focus on Text categorization (or text classification), text clustering, and Named-entity recognition (or concept/entity extraction). We arranged it mainly in two sections and several sub-section which are as follows. The volume of data on social networks is increasing every year, and an analysis of these

*Corresponding Author: *fereshte.rahmanian@gmail.com*

ISSN- 2348 5191 (Print) & 2348 8980 (Electronic)

Ambient Science, 2019: Vol. 06(1); 01-04
DOI:10.21276/ambi.2019.06.1.ga01

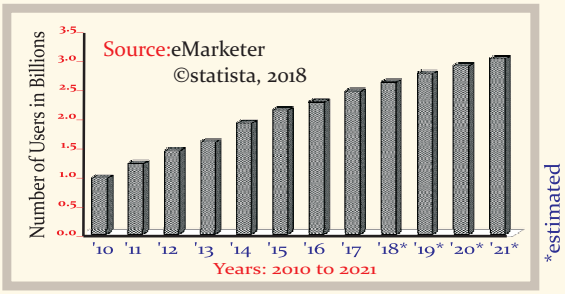

Figure-1: Number of social network users in the world

data is important in many sciences (Fig.-1).

Text categorization/classification in social media:

It is the task of allocating predefined categories to free-text documents. For example, news stories are normally organized by topics or geographical codes; academic papers are often categorized by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, etc. Another widespread application of text categorization is spam filtering (Chen *et al.*, 1996).

Text categorization refers to the automatic process of assigning predefined categories or classes to a body of the text. Iglesias *et al.* (2016) introduce a framework for real-time classification of web documents based on an evolving fuzzy system. This method starts with a term extraction step to clean up the text and extract keywords and then uses an evolutionary fuzzy logic to assign news to a related category (e.g. science-health category) based on the extracted terms.

Rumour stance classification: it is defined as classifying the social posts into one of supporting, denying, querying or commenting on an earlier post. Zubiaga *et al.* (2018) examined in depth the use of so-called sequential methods to the rumour stance classification task. Sequential classifiers are able to utilize the discursive nature of social media (Tolmie *et al.*, 2018), learning from how 'conversational threads' evolve for a more precise classification of the stance of each tweet. This work makes three core contributions on rumour stance classification: i) focus on the stance of tweets towards rumours that appear while breaking news stories unfold. ii) extend the stance types considered in previous work to include all types of responses observed during breaking news. iii) the method uses the emergent structure of interactions between users on Twitter to determine if each tweet is supporting, denying, querying or commenting on rumour's truth value.

Event Classification: a large number of social media on platforms like Flickr and Instagram is related to social events. The duty of social event classification refers to the difference of event and non-event-related contents as well as the classification of event types. Zeppelzauer *et al.* (2016), presented a widespread study of textual, visual, as well as multimodal representations for social event classification.

This study encompasses two tasks: (i) social event relevance recognition, (ii) social event type classification. For both the tasks, authors appraise the potential of the textual, visual modalities, and different information synthesis schemes. This paper evaluates on benchmark dataset from the SED 2013 (Reuter *et al.*, 2013).

Text clustering in social media

Cluster analysis (Clustering) is the task of grouping a set of objects in the same group (cluster) which are more similar to each other than to another cluster. Clustering is unsupervised classification and effective clustering maximizes intra-cluster similarities and minimizes inter-cluster similarities (Chen *et al.*, 1996). Newly user clustering has become more and more dynamic regarding the topic since in the high popularity of social media like Twitter, Weibo, and Facebook. Qiu & Shen (2017) explored efficient approaches for user clustering on short text streams with social data, aiming to infer topic sharing of users over time and dynamically cluster users based on their topic in the same cluster share similar topic.

Unlike the traditional approach, this method not only regards the dynamism of the topic distribution of the users but also includes social relations among users than to handle with the sparsity of short text, allocate the same topic for all words in one short text during Gibbs Sampling.

Co-citation analysis: it is a method developed by bibliometric research has a semantic similarity measure for documents that make use of citation relationships or the frequency with which two documents are cited by other articles (Small, 1973). Shiau *et al.* (2017) collected studies on social- related issues that were published between January 1996 and December 2014, collecting a total of 2565 papers and 81,316 citations from the ISI database. Co-citation investigation and cluster analysis were applied to prove seven main factors: (i) the degree of complex social networks; (ii) community structure; (iii) strong and weak ties; (iv) the progress of social media; (v) network structure and relationship; (vi) value concept and quantity strategies; and (vii) social principal.

Supply chain management (SCM): it is the management of the flow of goods and services includes the drive and storing of raw resources, of work-in-process inventory, and of finished goods from point of source to point of consumption (Ray *et al.*, 2016).To explore the information structure of SCM, Shiau *et al.* (2018) collected data (4652 articles and 145,242 references) from the WoS online database. In this study, SCM was taken as the main word, and the search period ranged from 1996 to 2013. To ensure the carefulness of the data, books and conference papers were eliminated from the 5034 research papers, and then, 4652 academic papers and 145,242 reference articles were

Ambient Science, 2019: Vol. 06(1); 01-04
DOI:10.21276/ambi.2019.06.1.ga01

nominated. Their outcomes revealed four core themes regarding SCM: (i) sustainable supply chain management, (ii) strategic competition, (iii) information value (iv) the improvement of SCM.

Negative links detection: the negative links (or harmful connections) are typically founded by fake profiles with destructive aims. Distinguishing bad (or suspicious) links within online users can better help in the mitigation of fake profiles from online social networks (OSNs).

The clustering coefficient has been changed as *Mutual Clustering Coefficient* denoted by *MCC*, to measure the connectivity between the joint friends of two linked users in a group (Wani & Jabin, 2018). Profile information aids to discovery the similarity between users. The study objects to show a method to recognize suspicious (negative) links created by the opponents by abusing the mutual friend trait in a group or a page on Facebook. Recognizing suspicious links can improve aid in designing the fake user detection system. The suggested method is founded on the composition of mutual clustering coefficient and profile information of a user which essentially helps in detecting suspicious connections in a group or a page on Facebook.

Named-entity recognition (NER) in social media:

NER is a subtask of data mining that pursues to discover and classify named entity indications in unstructured text into predefined classifications like person names, organizations, locations, medical codes, time expressions, etc (Ritter *et al.*, 2011).

Maximum research on NER has been structured as taking an unannotated block of text, such as this:

John sold 500 shares of Apple. in 2018.

And producing an annotated block of text that remarks the names of entities:

[John] Person sold 500 shares of [Apple] Organization in [2018] Time.

Sampling approach: Tran *et al.* (2017) presented a sampling approach based on self-learning to develop the training data. It goals to minimize annotation charges while maximizing the favorable efficiency from the model. The key concept was that the training data is updated with informative and very reliable samples. Thus, a classifier is provided with improved performance as a result of the variety and the plenty of samples in the training data. To address these problems, the authors suggest the sampling approaches founded on the uncertainty and variety of instances. The ambiguous samples are inquired from the outcomes of a classifier practical to unlabeled data, and diversity of instances is tested based on their context and content.

Vector models are applied to measure the liking score of the instances. Furthermore, the samples whose predicted label sequence is highly confident are also nominated to provide the training data. The conditional random fields (CRFs) are selected as a primary model to train the classifier. Both manually-labeled and machine-labeled instances are added to the training data to retrain a new classifier. Experiments were directed on the tweet streams to appraise the efficiency of the suggested method, and the specific query approach applied to the offered algorithm. The analyses disclose that the suggested method with the variety of the context and content is effective and dramatically improves the performance of the NER system (Tran *et al.*, 2017). Weather prediction Rossi *et al.* (2018) evaluated the possibility to found an automatic set of services aimed at connecting weather predicting with event detection and data extraction using social media. They took a case study, the information generated within Twitter, before and throughout a latest weather-induced flood in north Italy, evaluating the dynamics of the data generation process and the extraction of valued data for the key stakeholders of emergency management: meteoro-logical organizations, who subject weather prognostications and alerts, first responders, have to act in the response phase.

Annotations: Ghiasvand & Kate (2018) presented a new method for training clinical NER systems that do not require any manual annotations. It only requires a raw text corpus and a resource like UMLS that can give a list of named entities along with their semantic types. Using these two resources, annotations are automatically achieved to train machine learning approaches. The technique was evaluated on the NER shared-task datasets of i2b22010 and SemEval 2014.

Unlike Zhang & Elhadad (2013) where machine learning was not considered, we applied machine learning techniques which are trained on automatically annotated corpus followed by self-training iterations. This method can learn to identify named entities that could happen in changing contexts. Another primary distinction is that approach overlooks the probability that the seed terms could have multiple senses and may not be of the aim entity class when creating in a corpus, but this method ensures that only explicit terms are used for generating automatic annotations. Another primary distinction is that approach overlooks the probability that the seed terms could have multiple senses and may not be of the aim entity class when creating in a corpus, but this method ensures that only explicit terms are used for generating automatic annotations. One more difference; while the method regards only noun phrase pieces in order to define candidate named entities, a restriction also pointed out by them (Zhang & Elhadad, 2013), proposed method considers all noun phrases, containing nested ones, as achieved via complete parsing.

References:

Akbari, E., Jahanbin, K., Afroozeh, A., Yupapin, P., & Buntat, Z.

(2018). Brief review of monolayer molybdenum disulfide application in gas sensor. *Condens. Matter Phys*., 545:510-518.

Bello-Orgaz, G., Jung, J.J. & Camacho, D. (2016): **Social Big Data**: Recent achievements and new challenges. *Infor. Fusion*, 28:45-59

Chen, M.S., Han, J. & Yu, P.S. (1996): Data mining: an overview from a database perspective. *IEEE T. Knowl. Data Eng*., 8(6):866-883.

Farahbakhsh, R., Han, X., Cuevas, A., & Crespi, N. (2013): Analysis of publicly disclosed information in Facebook profiles. Proceedings of the **International Conference on Advances in Social Networks Analysis and Mining**, pp 699-705.

Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P. (1996): Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the **Second International Conference on Knowledge Discovery and Data Mining**, pp 82-88.

Feldman, R., & Dagan, I. (1995): Knowledge Discovery in Textual Databases (KDT). Proceedings of the **First International Conference on Knowledge Discovery and Data Mining,** pp. 112-117.

Frawley, W.J., Piatetsky-Shapiro, G., & Matheus, C.J. (1992): J.A.m. (1992). Knowledge discovery in databases: An overview. *AI Mag*., 13(3):57-70.

Ghiasvand, O., & Kate, R. (2018): Learning for clinical named entity recognition without manual annotations. *Inform. Med. Unlocked*, 13:122-127.

Iglesias, J.A., Tiemblo, A., Ledezma, A., & Sanchis, A. (2016): Web news mining in an evolving framework. *Info. Fusion*, 28:90-98.

Jahanbin, K., & Mehrjoo, S. (2017): Signal reconstruction through compressive sensing and principal component analysis in wireless sensor networks. *Int. J. Com. Sci. Net. Secur.*, 17(6): 147-154.

Kooshkaki, F.R., Mehrjoo, S., & Jahanbin, K. (2015): Video transmission by the use of bayesian compressive sensing in wireless sensor network. *Int. J. Com. Tech. Appl.*, 8(2):1675-1685.

Lomotey, R.K., & Deters, R. (2014): Towards knowledge discovery in big data. Proceedings of the **2014 IEEE 8th International Symposium on Service Oriented System Engineering,** pp 181-191.

Marrin, W. (2014): **Media Studies** 2.0. Pub. by: Routledge, Taylor & francis Group, London & New York. P. 214.

Mukkamala, R.R., Hussain, A., & Vatrapu, R. (2014): Fuzzy-set based sentiment analysis of big social data. Paper presented at: **2014 IEEE 18th International Enterprise Distributed Object Computing Conference**.

Nguyen, D.T., Hwang, D. & Jung, J.J. (2015): Time-Frequency Social Data Analytics for Understanding Social Big Data, pp 223-228. In: Camacho, D., Braubach, L., Venticinque, S. & Badica, C. (eds) **Intelligent Distributed Computing VIII. Studies in Computational Intelligence**, vol 570. Pub by: Springer, Int. Pub., Switzerland.

Pouriyeh, S.A., & Doroodchi, M. (2009). Secure SMS Banking Based On Web Services. Proceedings of the **2009 International Conference on Semantic Web & Web Services**. Las Vegas, Nevada, USA.

Pouriyeh, S.A., Doroodchi, M., & Rezaeinejad, M. (2010). Secure Mobile Approaches Using Web Services. Proceedings of the **2010 International Conference on Semantic Web & Web Services**, Las Vegas, Nevada, USA.

Qiu, Z., & Shen, H. (2017): User clustering in a dynamic social network topic model for short text streams. *Infor. Sci*., 414:102-116.

Rabade, R., Mishra, N. & Sharma, S. (2014) Survey of Influential User Identification Techniques in Online Social Networks, pp. 359-370. In: Thampi, S., Abraham, A., Pal, S. & Rodriguez, J. (eds.) **Recent Advances in Intelligent Informatics. Advances in Intelligent Systems and Computing**, vol 235. Pub. by: Springer International Publishing, Switzerland.

Ray, S.K., Basak, A., Fatima, K., & Seddiqe, M.M.I.S.(2016): Study on Supply Chain Management of Industries in FMCG Sector in Bangladesh. *Global J. Res. Eng. (G)*, 16(2): 29-34

Reuter, T., Papadopoulos, S., Petkos, G., Mezaris, V., Kompatsiaris, Y., Cimiano, P., de Vries, C., & Geva, S. (2013). Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. Proceedings of the **MediaEval 2013 Multimedia Benchmark Workshop.** Barcelona, Spain.

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. Proceedings of the 2011 **Conference on Empirical Methods in Natural Language Processing**, pages 1524–1534,

Rossi, C., Acerbo, F.S., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., Tarasconi, F., Cristoforetti, M., & Alikadic, A. (2018): Early detection and information extraction for weather-induced floods using social media streams. *Int. J. Disaster Risk Reduct.*, 30(A):145-157

Shiau, W-L., Dwivedi, Y.K., & Lai, H. (2018): Examining the core knowledge on facebook. *Int. J. Info. Manag*., 43:52-63.

Shiau, W-L., Dwivedi, Y.K., & Yang, H.S. (2017): Co-citation and cluster analyses of extant literature on social networks. *Int. J. Info. Manag.*, 37(5):390-399.

Small, H.J. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Info. Sci. banner.* 24(4):265-269.

Tolmie, P., Procter, R., Rouncefield, M., Liakata, M., & Zubiaga, A. (2018): Microblog Analysis as a Program of Work. *ACM T. Soc. Computing Arch.* 1(1):1-43.

Tran, V.C., Nguyen, N.T., Fujita, H., Hoang, D.T., & Hwang, D. (2017): A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields. *Knowledge-Based Sys*., 132:179-187.

Wani, M.A., & Jabin, S. (2018): Mutual Clustering Coefficient-based Suspicious-link Detection approach for Online Social Networks. *J. King Saud Univer. - Com. Info. Sci.,* (in Press)

Zeppelzauer, M., Schopfhauser, D. (2016): Multimodal classification of events in social media. *Image Vision Comp. J.,* arXiv.org > cs > arXiv:1601.00599

Zhang, S., & Elhadad, N. (2013): Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *J. Biomed. Inform*., 46(6):1088-1098.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., Cohn, T. & Augenstein, I.J. (2018): Discourse-aware rumour stance classification in social media using sequential classifiers. *Infor. Process. Manag.* 54(2), 273-290.