A review of the use of machine learning techniques by social media enterprises

Fazle Rabbi

Member Australian Computer Society

Abstract

In this review, the current status of research on machine learning techniques by social enterprises was explored. First the popularity of social media, the type of big data generated by them, how they are potentially and actually are used were outlined. The general concept of machine learning, components, methods and its application were discussed. On this background, specific methods and stages used in machine analytics of social media were reviewed with explanatory diagrams reproduced from the works of various authors. Bayesian Network and Support Vector Machine are most commonly used methods due to their advantages over other methods. The methods are used mostly for opinion mining, sentiment analysis and trend analysis. Social network analysis and business management are the most frequent applications, although it has been applied in many other fields also. Many new methods and modifications of current methods are being developed almost continuously to solve the problems in current methods.

There are also limitations in current machine learning techniques. Samples drawn from big data may not always represent of the population. The gigantic size of data, with low value intensity, spread over several sources and their dynamic nature can cause certain biases. Large computing capacities and sophisticated methods of sampling, extraction and analysis are required to handle big data. Accuracy and objectivity of social media data are not always assured and the problem multiplies with use of diverse data sources. Access and ethical issues may arise. Heterogeneity, noise accumulation, spurious correlations and incidental endogeneity are other problems.

Keywords: Machine Learning, Statistics, Social Media

Introduction

In 2017, there were 4.6 billion users of social media globally. Facebook (2234 million users), YouTube and WhatsApp (1500 million users each) were the top three social media in 2017 (Statista, 2018). Most users spend more than 20% of their time on social media. Their use in mobile phones is increasing faster. Billions of data are generated every day due to the massive scale of social media use. Likes, dislikes and comments posted in these media can provide valuable information for businesses on customer preferences, choices and buying behaviour of products and services. Such information can be used by firms to devise strategies to maximise their sales.

Mathsworks (Matlab producers) defines machine learning as a data analytics technique used for teaching computers to learn from experience, similar to humans. In the algorithms used in machine learning, computational methods are applied to learn information directly from data free of dependence on a predetermined model equation. Adaptive improvements in the performance of algorithms happens with increasing frequency of learning increasing the number of samples or instances (Mathworks, 2018). In their paper, Fan and Gordon (2014) cited Zeng, et al. (2010) to define social media analytics "is concerned with developing and evaluating informatics tools and

frameworks to collect, monitor, analyse, summarize, and visualize social media data usually driven by specific requirements from a target application.

Emergence of big data has made machine learning an important tool for solving problems in many areas. Machine learning algorithms find natural patterns in data that generate insight and help you make better decisions and predictions. They are used every day to make critical decisions. An example of its social utility is media sites relying on machine learning to analyse several numerous options to give recommendations to users. Another example is retailers using it to obtain insight into purchasing behaviour of people.

Machine learning can be used when solution to a complex task or problem involving a large amount of data and lots of variables is required in the absence of any formula or equation. Face/speech recognition, fraud detection in records of transactions, dynamic nature of data requiring continuous adaptation like online shopping. Two types of machine occurs as explained ii Fig 1 reproduced from Mathworks (2018).

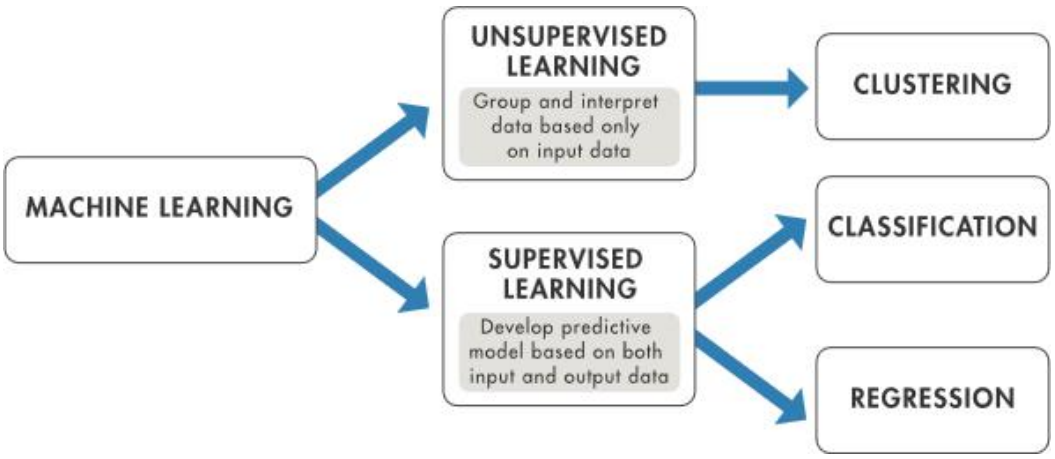

Figure 1 Two types of machine learning (Mathworks, 2018)

In supervised learning both input and output data are used for developing a predictive model. In the unsupervised learning, only input data are used for identifying patterns and intrinsic structures in the input data. The differences between the algorithms used in the two types are shown in Fig 2 reproduced from Mathworks (2018).

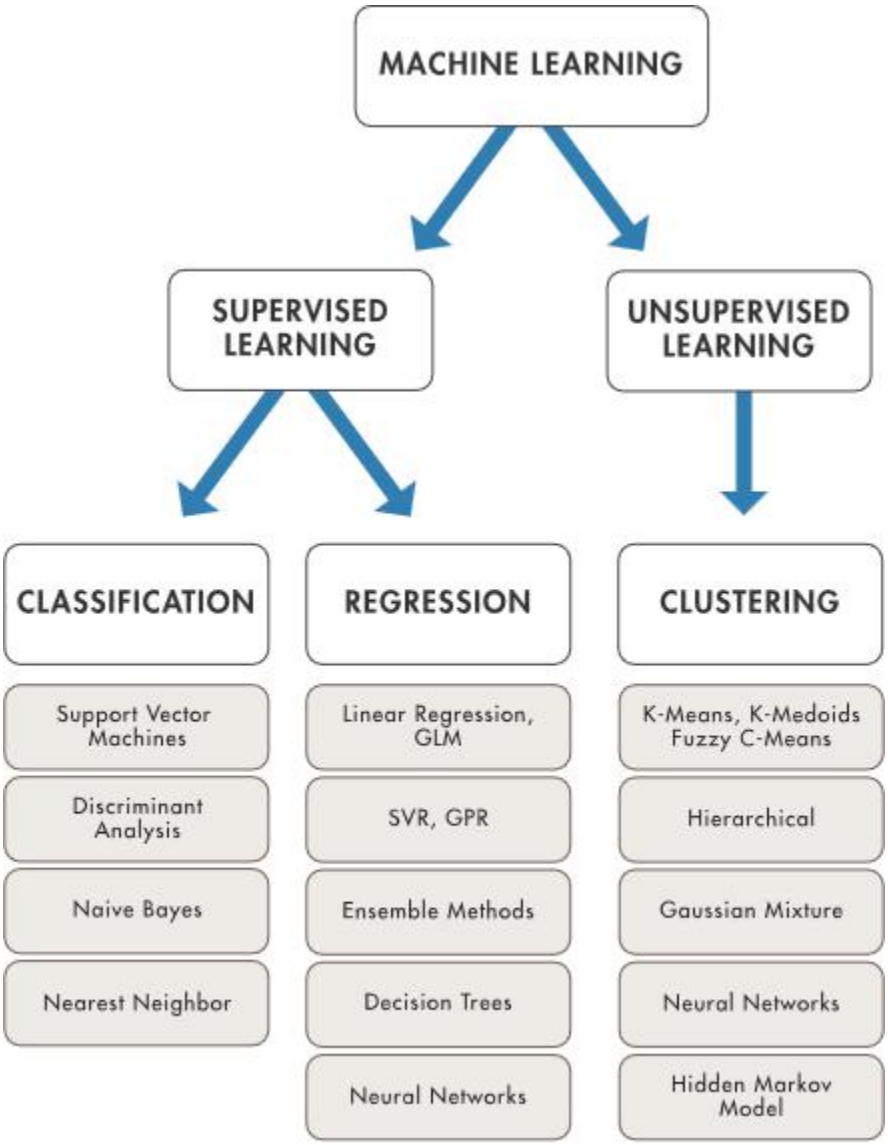

Figure 2 Methods used in the two types of machine learning (Mathworks, 2018)

With this background information on machine learning techniques in general, how these or other techniques are applied by social media enterprises is examined in the next section.

Machine learning techniques used by social media enterprises

The discussion on this review topic is limited to evaluation of current status rather than critical review, due to sheer size and diversity of research reports on this topic. Out of several paper published during the last decade, some important ones are selected for discussions on methods and applications of machine learning techniques in social media.

The overall picture

Using search terms in search engines and selecting on quality of research, Injadat, et al. (2016) reviewed available data mining techniques of social media. In all, 19 types of data mining were

identified as used by researchers. An alphabetical order of the methods given by the authors (p 5) is: AdaBoost, Artificial Neural Network (ANN), Apriori, Bayesian Networks (BN), Decision Trees (DT), Density Based Algorithm (DBA), Fuzzy, Genetic Algorithm (GA), Hierarchical Clustering (HC), K-Means, k-nearest Neighbours (k-NN), Linear Discriminant Analysis (LDA), Linear-Regression (Lin-R), Logistic Regression (LR), Markov, Maximum Entropy (ME), Novel, Support Vector Machine (SVM) and Wrapper. SVM (22%) and BN (20%) were the most used methods. Others were used only to the extent of less than 10%. The most common applications were in Business and Management (BM), Education (EDU), Finance (FIN), Government and Public (GP), Medical and Health (MH) and Social Networks (SN) (p 6 of the authors). Social network heads the list with 62%, followed by Business Management with 17% and the rest are below 10% of reported applications. Within social media, microblogging topped with 31%, followed by social networks (12%) and videos and photos sharing (11%). Research objectives were distributed among (p 7): Biometric, Content Analysis, Cyber Crime, Disease Awareness, Geolocating, Quality Improvement, Risk Management, Semantic analysis and Sentiment analysis, with quality improvement an sentiment analysis as the most frequent objectives. Machine learning methods, especially SVM, BN and DT, were more popular among researchers, although non-machine learning methods are also important. Best method of data mining in social media were SVM and BN.

Methods used in machine learning techniques

A brief discussion of machine learning techniques in social media is available in Brooke (2018). What was started as an interacting environment, social media like Facebook saw the opportunity of the big data available on users' preferences and dislikes of products and services in the market. Thus, these social media have become both marketing channels and information sources for firms to identify their customer needs and problems and devise marketing strategies to maximise sales. Postings in social media are unstructured data and do not follow any rules. Social media contain postings of a mixture of texts, images, sounds and videos. Advantages of using machine learning in social media are: actionable insights, scalable, differentiating between text and contexts and relevance and authority, speaking for the customers in their own language.

A review of commonly used machine learning techniques for big data analyses was done by Magesh and Nirmala (2016). In his article, Narula (2018) reviewed and listed the pros and cons of various commonly used machine learning techniques for different purposes. The reviewed techniques were: Bayesian classifier, decision tree classifier, support vector machines, neural networks, logistic regression, linear regression and K-nearest neighbours.

In their review, Fan and Gordon (2014) identified three overlapping stages of social analytic processes as capture, understand and present (Fig 3). Businesses capture those data from millions of data available in social media, based on relevance to their activities. Relevant information from several different social media like Facebook, Twitter, LinkedIn and YouTube and blogs are collected using Application Programming Interfaces (API), newsfeed or crawling methods. In the capture stage inclusion and exclusion criteria need to be clearly defined to avoid collection of unnecessary data. Derivation of meaning and metrics to use in business decisions is are included in understand stage. Eliminating noisy data using rule-based classification leads to cleaning of data to the more pertinent ones. Now meaning is derived using statistical techniques like text and data mining, network analysis, machine translation and natural language processing. These techniques help to identify the sentiments (feeling and behaviour) of users. Some of the useful

metrics possible to derive here are: background, network relationships, interests, problems and concerns of social media users. Such information are vital to the business. The results of understand stage affects present stage and influences the decisions taken. Both offline and online processing pf data are employed. In visual analytics, human participation may also occur. Summarised results from understand stage are presented using different presentation techniques. Visual dashboard is very commonly used. The authors also listed and discussed opinion mining (sentiment analysis), topic modelling, social network analysis, trend analysis and visual analytics as the main social analytical methods. Models created in understand stage can support capture stage. Visual analytics is useful both in understand and in present stages. Sentiment analysis is most useful technique for businesses. Text analytics is done using text analytics methods like computational linguistics and natural language processing to automatically extract user sentiments or opinions from text sources at any level of granularity from mere words or phrases extending to whole document. However, sampling biases and flawed techniques can lead to wrong results.

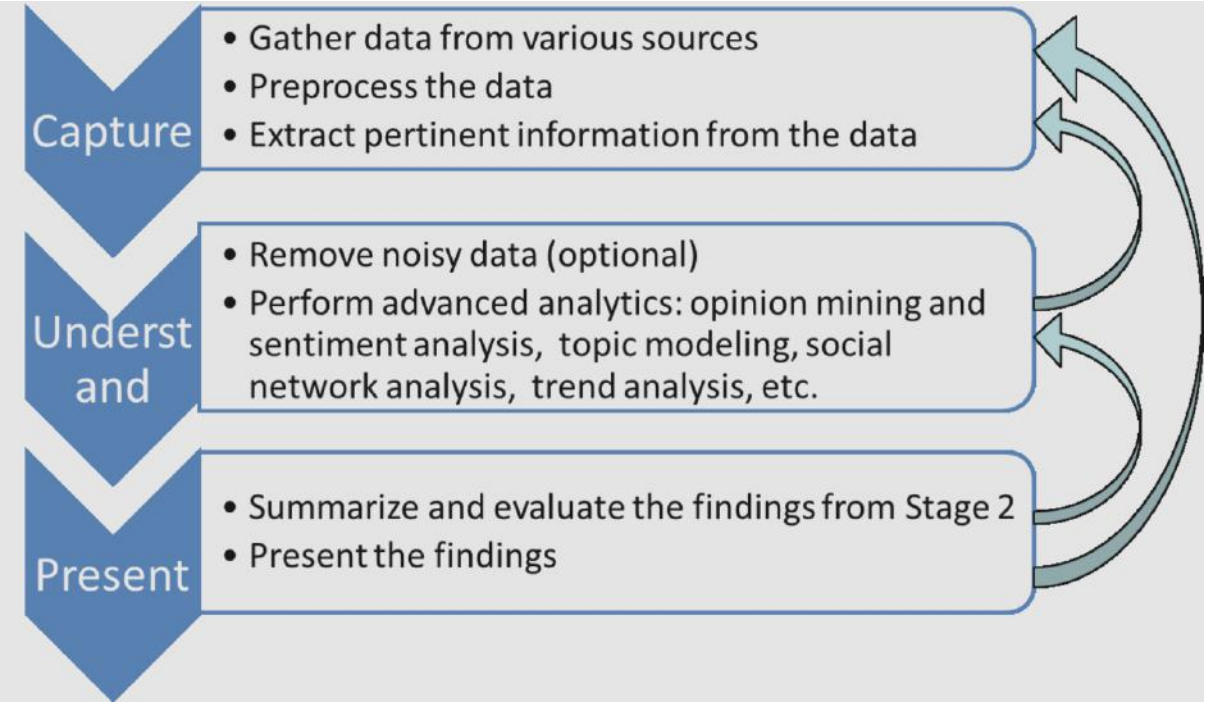

Figure 3 Stages of social media analytics (Fan & Gordon, 2014)

Large data samples does not mean unbiased. Situations may occur when satisfied customers are silent and dissatisfied are vociferous and express their negative views immediately and more frequently. Topic modelling is a type of thematic analysis. Themes around which large number of users cluster are identified and tabulated. Deviations are especially noted. The identified themes can be used in some other analytics like interests and concerns. Modelling algorithms allow using thematic analysis in streaming and continuous data as well. Advanced statistical methods can be used to identify latent topics between words and texts. Social network analysis informs about social relationships of users. Influential relationships can be used for business promotion. Real time analytical techniques are slowly replacing conventional methods of social network analysis. Trend analysis is useful to forecast sales growth, effectiveness of promotional campaigns and assessing risks. Apart from conventional time series and regression analysis,

neural analysis and support vector machines are now available. In visual analytics, reasoning using interactive visual interfaces is done. Synthesis, exploration, discovery and confirmation of results obtained from big data is done through visual analytics. A combination of human and machine efficiencies are used. Both hidden structure and details need to be revealed by visual analytics. Dashboards are used for visual analytics and presentation. In crisis management and breaking events situations, dashboards are useful. Businesses uses in increasing sales, product and market diversification and segmentation, new product design and development are obvious.

A master thesis titled "Machine learning techniques for social media analysis" by Korda and Vidanovic (2018) described a machine learning model with spam filters for evaluating sentiment of users commenting on company postings and for clustering using unsupervised learning techniques to process user data.

Big data in social media was termed "social big data" and some recent methods of social big data analytics were discussed by Bello-Orgaz, et al. (2016). Importantly, the interface between big data, social media and analysis was depicted diagrammatically, reproduced in Fig 4. At the interface, all the three merge.

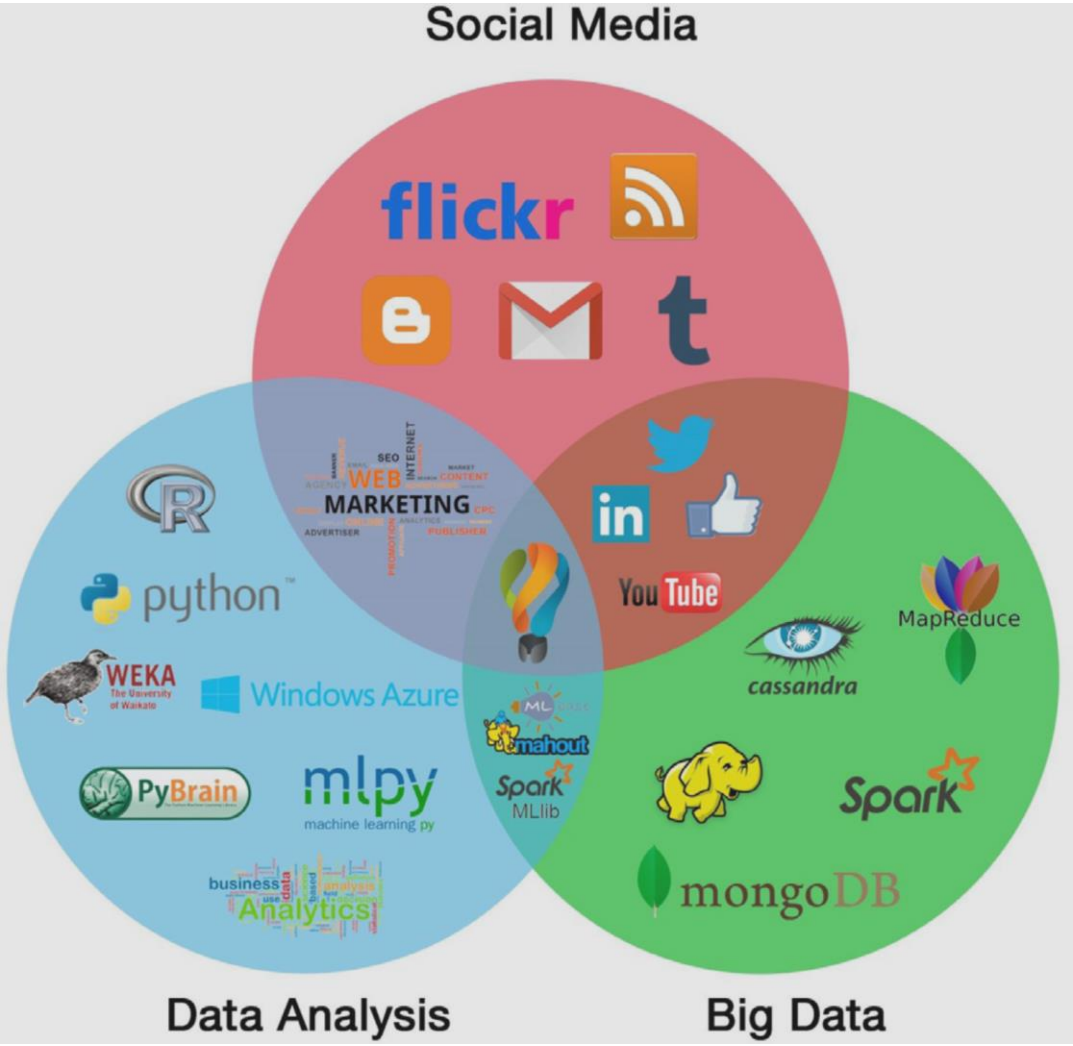

Figure 4 Interface of big data, social media and analytics (Bello-Orgaz, et al., 2016)

New analytics frameworks based on MapReduce concept like Apache Hadoop and Spark for processing of large data have been developed. New libraries (Mahout and SparkMLib) for efficient storage and machine learning algorithms for massive data analytics have been developed. The framework of MapReduce given by the authors is reproduced in Fig 5. Many other Apache-frameworks for various uses, including machine learning, have also been discussed.

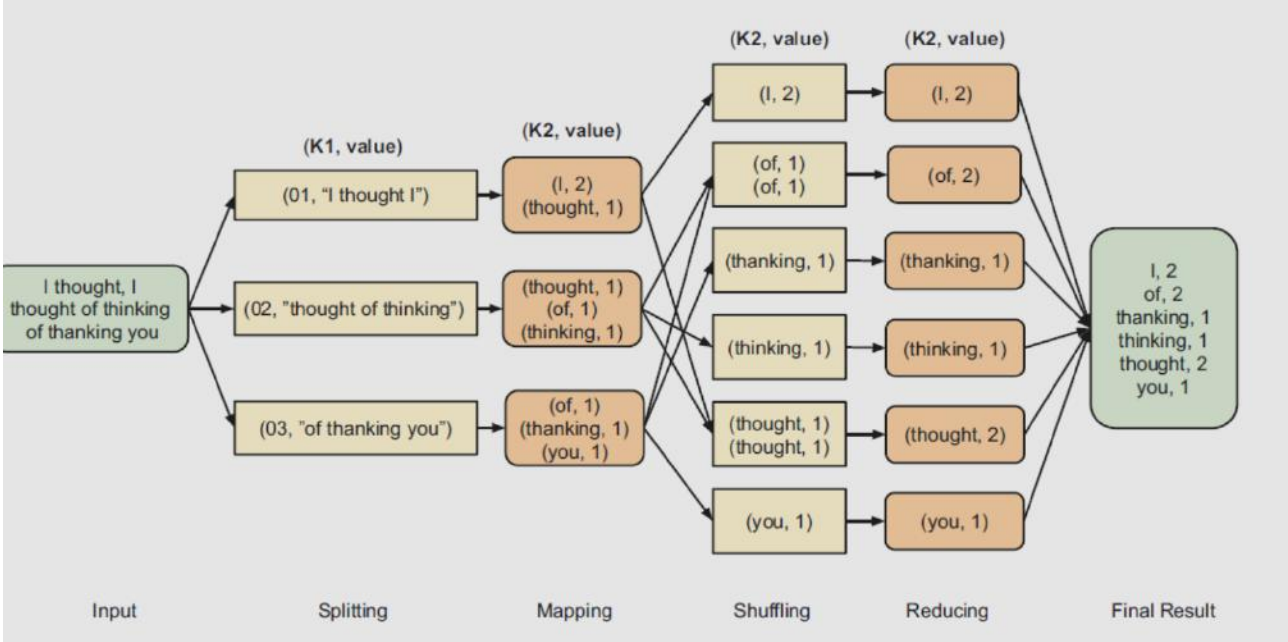

Figure 5 Framework of MapReduce for text word count (Bello-Orgaz, et al., 2016)

Competitive intelligence is the process of acquiring information about the competitor's strategies and market information from social media and other web sources. Machine learning methods are applied here.

The methods of acquiring competitive intelligence from social media were described by Dey, et al. (2011) The sources of data, framework (Fig 6) and analytics tools for use in various stages have been described and discussed. Case study examples have been provided to prove its utility.

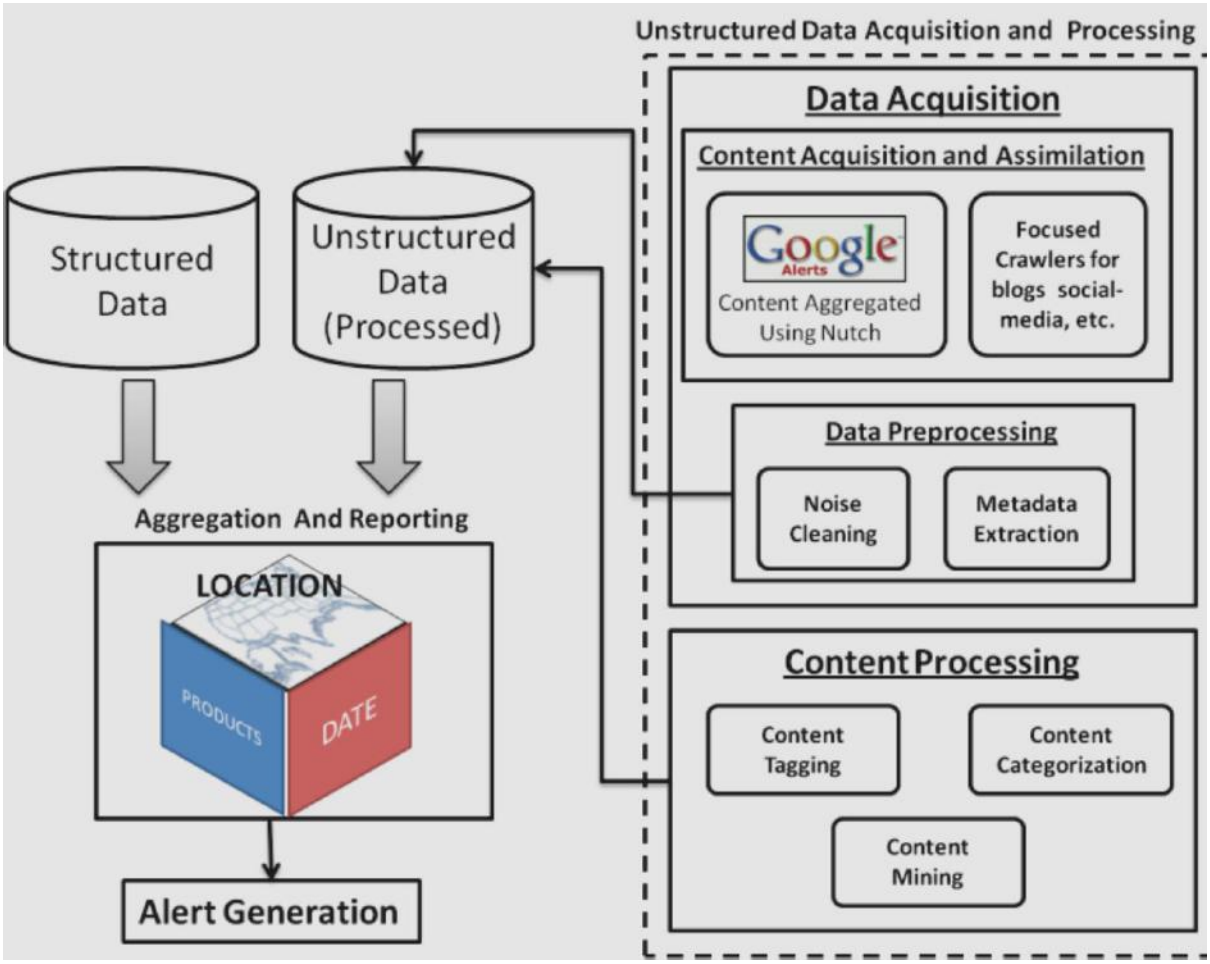

Figure 6 Framework of competitive intelligence analysis (Dey, et al., 2011)

In their work, Khan, et al. (2015) noted that sentiment analysis was originally developed for large texts like product reviews. In the current form, its applicability to twitter may be limited. The authors proposed a new entity-level sentiment analysis of twitter. It consists of two steps of sampling and analysis. In the first step, a lexicon-based method is used. This method can give high precision, but it gives low recall. So, another set of additional twitter samples were taken in the second step. The information from the first step was used for automatic identification of opinionated tweets. Labelling was done manually, but it can also be done automatically.

Stressing on the interdisciplinary nature of social media analytics (SMA), Stieglitz, et al. (2014) presented its framework as reproduced in Fig 7. Compared the frameworks given in Fig 1 and Fig 2 (Mathworks, 2018), there is no categorisation of supervised an unsupervised learning. However, compare to the three stages in Fig 3 (Fan & Gordon, 2014), this framework is more detailed. Some challenges are also discussed. Being an inter-disciplinary field in its early stages, SMA borrows core theories from other disciplines such as statistics, sociology, network science, and graph theory. Also, interdisciplinary cooperation is an absolute necessity for SMA. There are methodological issues related to each type of data. No common method is possible to handle all types of structured and unstructured data. The gigantic size of data spread over several sources and their dynamic nature introduces certain biases however carefully the method is chosen. Large size requires large computing capacities and sophisticated methods of sampling, extraction

and analysis. Deriving meaningful insights from such large and dynamic information sources is a serious challenge for data mining. Accuracy and objectivity of social media data are not assured always. When different data sets are used together, these problems multiply. Collecting data from such numerous sources which vary in the nature of textual content itself is a challenge. Access to the data and differences in data formats may vary between social media. Ethical issues also arise when data from social media are collected. Currently used computational methods and algorithms have many weaknesses.

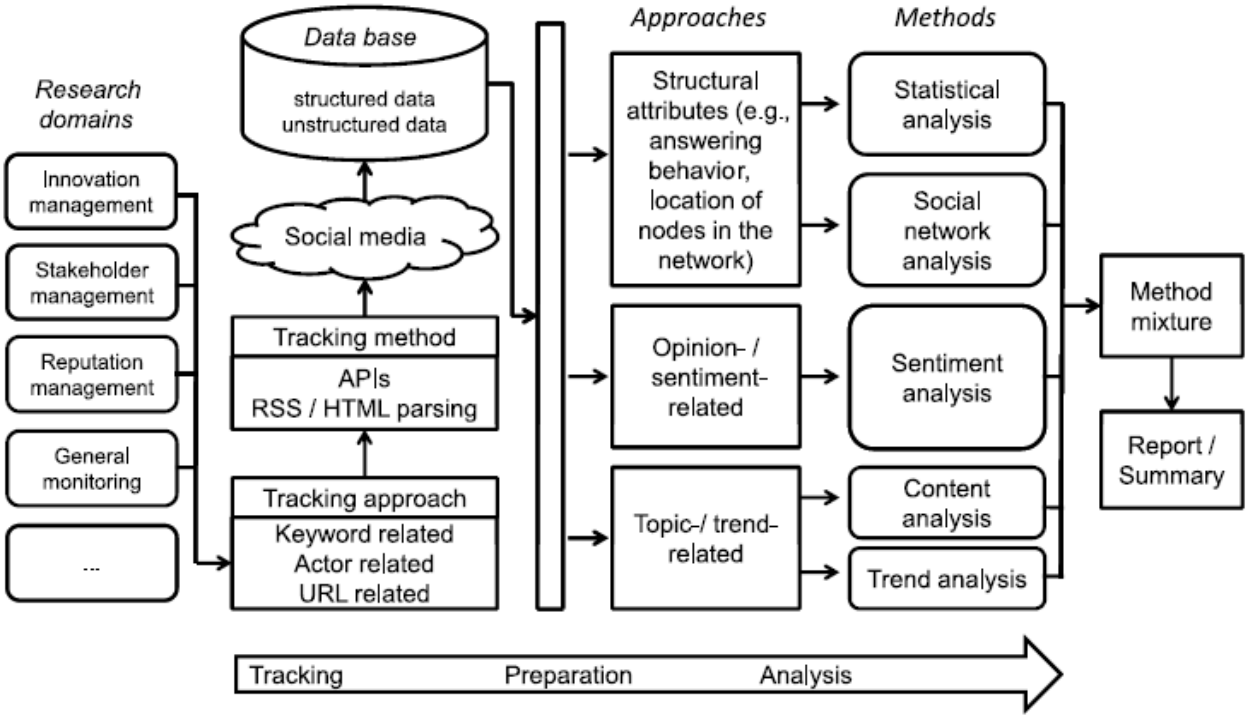

Figure 7 Framework of social media analytics (Stieglitz, et al., 2014)

Approaches and methods used for data mining and analysis were critically evaluated by Cambria, et al. (2013). Most of the points were similar to the points discussed many workers cited above.

User information available from one social media may not be complete. It is better to seek information on the same individual from multiple social media sites. To address problems associated with cross-media user identification, a methodology called MOBIUS, was introduced by Zafarani and Liu (2013). A behavioural modelling approach was used in this model for the purpose of systematic feature construction and assessment, which can facilitate integration of additional features, if required. The authors tested some of its applications for user name verification and suggested a few other applications of their model.

In their review, Gandomi and Haider (2015) discussed the concept and definitions of big data. Volume, variety and velocity (the 3 Vs) are the characteristics of big data. In addition, terms like veracity (inherent unreliability of some data), variability (referring to the large variations within data and its flow rates), complexity (big data are generated through many complex sources), value (value intensity of big data in relation to the volume is generally low). A lower limit for the

size of big data cannot be prescribed as the storage capacities of social sites increase over time. Types of data are also important in defining the size of big data. Textual data and video data of the same size require different processing methods. Structural heterogeneity is a characteristic of big data across the social sites. Only 5% of the big data are currently structured. Analytical methods for various types of big data, including audio and video data, and social media analytic, are discussed. The massive nature of big data samples makes them representative of the majority of population, if not entirely. Then, statistical significance becomes irrelevant in big data analysis. Some inherent distinctive features of big data such as heterogeneity, noise accumulation, spurious correlations and incidental endogeneity pose challenges to the analytics.

In order to make machine learning more accessible to non-experts, Tim, et al. (2013) proposed a system called MLbase. The system has at its core an optimizer (Fig 8). This optimiser transforms a declarative ML task into a sophisticated learning plan. During this process goes on, the optimizer tries to identify a plan aimed at delivering quick returns a high quality answer to the user. Thus, MLbase tries to improve the result through several iterations. The design of MLbase is for fully distributed activities. The run-time is adequate to use the characteristics of machine learning algorithms very efficiently. The potential of the optimizer and performance advantages of algorithm-specific execution strategies were demonstrated in the paper.

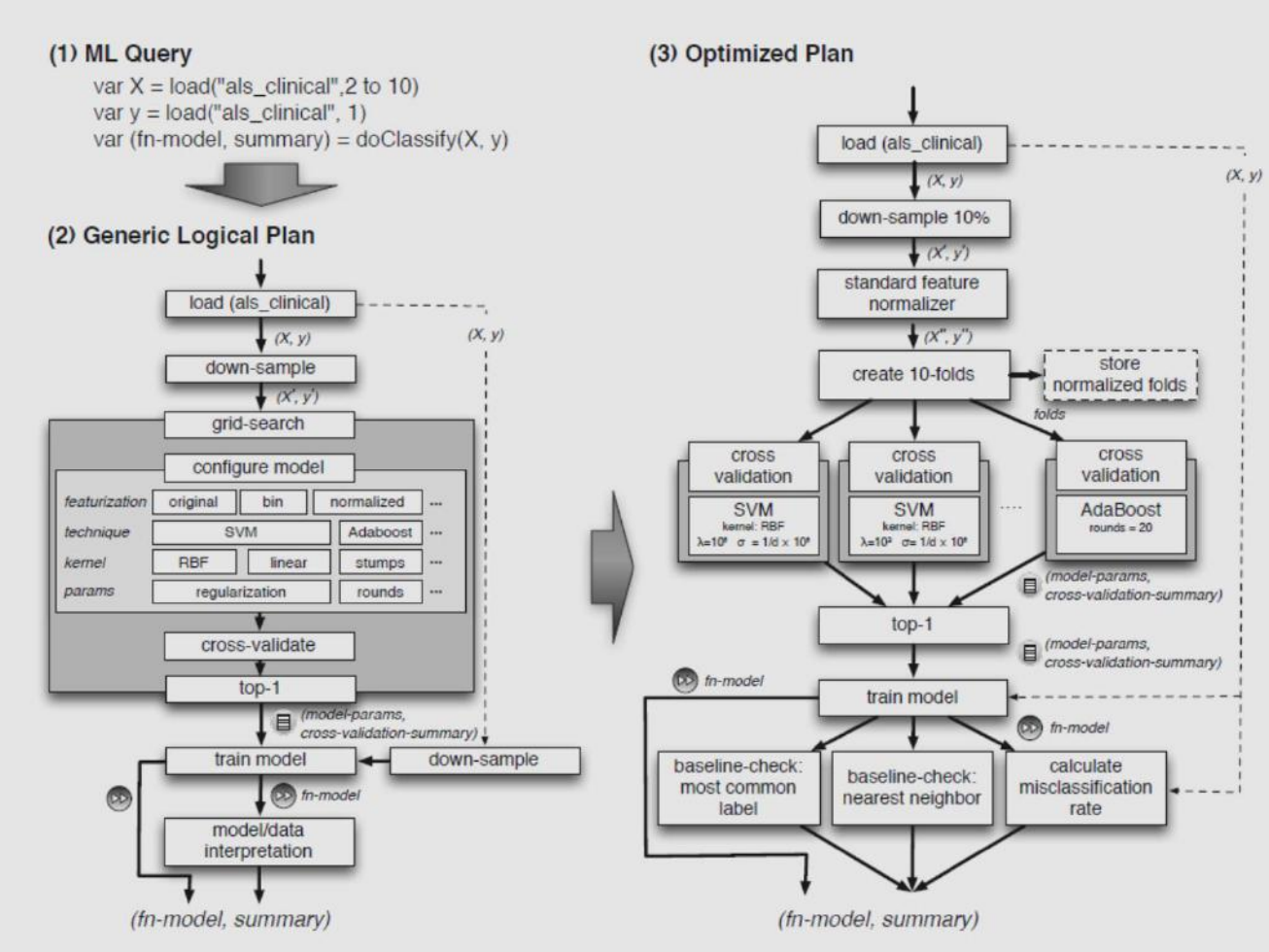

Figure 8 Optimisation process of MLbase (Tim, et al., 2013)

A theorem on big data, HACE theorem was proposed by Wu, et al. (2014), The HACE theorem states that, "Big data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data". Three research challenges of big data were also identified as Tier I, Tier II and Tier III. Tier I is the big data mining platform focusing on low level of data accessing and computing challenges. Tier II has the challenges of information sharing and privacy issues, and application domains and knowledge to be gained from big data. Finding suitable mining algorithms is the Tier III challenge.

Applications of machine learning techniques in social media

Social media messages which work best for restaurant industry-consumer interactions were determined by Kwok and Yu (2013) analysing 982 Facebook messages. Information about the restaurant like menu descriptions was more popular than market-related words like winner or check. Photo and status media types of were liked more than link or embedded videos. Conversational messages were received more favourably compared to sales and marketing messages. Thus, clear pattern of likes and dislikes of social media users emerged from this study.

From the comparison of three pizza firms done by He, et al. (2013), it appears that the relative numbers of followers for different firms may vary for different social sites. Thus, there may be more followers for a firm in Facebook and for another firm in Twitter. The total followers in all social sites may be more significant. All the firms used social media to actively engage their customers in social media to promote their services and sustained bond with the customers.

Text mining of a random sample of 3516 tweets by Mostafa (2013) revealed positive consumer sentiments towards famous global brands like Nokia, IBM, KLM, DHL etc. The authors thought web texts as noisy due to lexical and syntax problems. For the analysis, the author used, a lexicon predefined by experts, which included about 6800 seed adjectives with their orientation known.

Sentiment analysis and classification of reviews on travel destinations was done by Ye, et al. (2009) using three supervised machine learning algorithms, viz., Naïve Bayes, SVM and the character based N-gram model. About 80% accuracy was obtained for all three approaches. However, SVM and N-gram methods were superior to Naïve Bayes.

Noting the widespread use of social media for political communications, Conover, et al. (2011) found use of SVM on Twitter hashtag data was superior to its use on full texts to predict political affiliations of people with 91% accuracy when tested on 2010 US midterm polls. Latent semantic analysis was used to identify hidden structure of with good correlation for political affiliations. Classifier-based methods like clustering (95% accuracy) were better than content-based methods.

Advanced semantic analysis methods were used by Yu, et al. (2013) to compare social media and conventional media on short term stock market performances of firms. Social media influenced stock return and risk more than conventional media. Both types of media also interact to offer higher strength to predictability. Different types of social media impacted the stock performance variously. Thus social media marketing strategies are important in stock market performances also.

Social media bots, algorithmically driven entities that appear to be genuine users, were used to the extent of at least 20% of the entire users, to manipulate the 2016 US presidential election according to the research done by Bessi and Ferrara (2016). Using the latest bot detection

algorithms, the authors found political partisanships in hashtag adoption, spatial and temporal communication, political support dynamics and influence mechanisms.

Posts in the brand's pages of social media like Facebook can be used for prediction of performance metrics and how they impact brand building. In a research work on cosmetics, Moro, et al. (2016) obtained the best results with minimum error (26.9%) for lifetime people liked the page and engaged with the post. The lifetime post consumers also had a brand building predictivity with a mean error of 27.2% with maximum effect for type of the content in the Facebook post among the seven model features. The support machine vector method was used by feeding seven input variables. The authors have used several diagrams to describe the procedure adopted at each stage.

Machine learning algorithms may not work in situations and this opacity can happen in three ways: intentional opacity created as to protect corporate or state secrecy, opacity related to technical illiteracy and opacity as a characteristics of machine learning algorithms and arising from scale of applications required (Burrell, 2016).

Many more papers deal with methods and applications of machine learning techniques in social media analytics for diverse objectives related to different sectors of activities.

Conclusions

There are varieties of techniques of social media analytics using machine learning or non-machine learning methods. Decisions on whether machine learning is better, which methods to be used and how to use them depend on the research context, area of research, objectives of the study, social media considered, types of big data available, machine limitations of learning and human skills limitations.

References

Bello-Orgaz, G., Jung, J. J. & Camacho, D., 2016. Social big data: Recent achievements and new challenges. *Information Fusion,* 28(March), pp. 45-59.

Bessi, A. & Ferrara, E., 2016. Social Bots Distort the 2016 US Presidential Election Online Discussion. *First Monday,* 21(11), p. 14 pp.

Brooke, S., 2018. *How to Make Sense of Social Media Using Machine Learning.* [Online] Available at: https://towardsdatascience.com/how-to-make-sense-of-social-media-using-machine-learning-8a3db1506d03
[Accessed 27 April 2018].

Burrell, J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society,* 3(1), p. 2053951715622512..

Cambria, E., Schuller, B., Xia, Y. & Havasi, C., 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems,* 28(2), pp. 15-21.

Conover, M. D. et al., 2011. *Predicting the political alignment of twitter users.*. s.l., IEEE, pp. 192-199.

Dey, L., Haque, S. M., Khurdiya, A. & Shroff, G., 2011. *Acquiring Competitive Intelligence from Social Media.* s.l., ACM, p. 3.

Fan, W. & Gordon, M. D., 2014. The power of social media analytics. *Communications of the ACM,* 57(6), pp. 74-81.

Fan, W. & Gordon, M. D., 2014. Unveiling the Power of Social Media Analytics. *Communications of the ACM,* 57(6), pp. 74-81.

Gandomi, A. & Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* Volume 35, pp. 137-144.

He, W., Zha, S. & Li, L., 2013. Social media competitive analysis and text mining: A case study in the pizza industry.. *International Journal of Information Management,* 33(3), pp. 464-472.

Injadat, M. N., Salo, F. & Nassif, A. B., 2016. Data mining techniques in social media: A survey. *Neurocomputing,* 214(November), pp. 654-670.

Khan, A. Z., Atique, M. & Thakare, V., 2015. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering.*

Korda, P. & Vidanovic, P., 2018. *Machine Learning Techniques for Social Media Analysis,* s.l.: POLITECNICO DI MILANO.

Kwok, L. & Yu, B., 2013. Spreading social media messages on Facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly,* 54(1), pp. 84-94.

Magesh, S. & Nirmala, K., 2016. A Survey on Machine Learning Approaches to Social Media Analytics. *International Journal of Applied Engineering Research,* 11(4), pp. 2411-2416.

Mathworks, 2018. *What is machine learning.* [Online]
Available at: https://www.mathworks.com/discovery/machine-learning.html
[Accessed 27 April 2018].

Moro, S., Rita, P. & Vala, B., 2016. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research,* 69(9), pp. 3341-3351.

Mostafa, M. M., 2013. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications,* 40(10), pp. 4241-4251.

Narula, G., 2018. *Machine Learning Algorithms for Business Applications – Complete Guide.* [Online]
Available at: https://www.techemergence.com/machine-learning-algorithms-for-business-applications-complete-guide/
[Accessed 27 April 2018].

Statista, 2018. *Most popular social networks worldwide as of April 2018, ranked by number of active users (in millions).* [Online]
Available at: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[Accessed 27 April 2018].

Stieglitz, S., Dang-Xuan, L., Bruns, A. & Neuberger, C., 2014. Social media analytics. *Business & Information Systems Engineering,* 6(2), pp. 89-96.

Tim, K. et al., 2013. MLbase: A Distributed Machine-learning System. *Cidr* , 1(1-2), p. 7 pp.

Wu, X., Zhu, X., Wu, G.-Q. & Ding, W., 2014. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1), pp. 97-107.

Ye, Q., Zhang, Z. & Law, R., 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications,* 36(3), pp. 6527-6535.

Yu, Y., Duan, W. & Cao, Q., 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems,* 55(4), pp. 919-926.

Zafarani, R. & Liu, H., 2013. *Connecting users across social media sites: a behavioral-modeling approach.* s.l., ACM, pp. 41-49.

Zeng, D., Chen, H., Lusch, R. & Li, S.-H., 2010. Social media analytics and intelligence. *IEEE Intelligent Systems,* 25(6), pp. 13-16.