IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Special Issue || Dec. 2019 || PP 21-26
International Conference on Management Practices, Innovations & Research 2019

Data Mining Techniques for Social Media Analysis: A Review

Charanjeet Dadiyala, Neha Mogre, Priyanka Mogre, S. Kiran Kumar
Assistant Professor RCERT, Chandrapur
Assistant Professor TGPCET, Nagpur
Assistant Professor BIT, Ballarpur
Assistant Professor CMR, Hyderabad

Abstract— *Social Media has now become one of the most important applications that almost everyone uses and manages on a daily basis, results in generation of tremendous amount of personal data. There are many popular websites like Twitter, Facebook, Instagram, Quora, etc., daily stores and records our personal data in terms of millions of chats, likes, comments, answers etc. This tremendous amount of personal data can also be referred as Big Data, contains almost everything from making a purchase, booking a movie ticket, viewing online products or planning an international trip. This phenomenon gave birth to the idea of Social Media Online Data Tracking. This tracking can help companies to learn and know better about their consumers about their choices. The amount of Big Data that gets generated can reach up to Petabytes. As Big Data is most likely to be generated in an unstructured format which can be considered raw to be the input to any system. Here, researchers and academicians come into play to develop a method to covert the unstructured data into structured ready to use format. These methods are various data mining techniques used for better analyzing social media and its generated data to understand and target the concerned audience. These tools can prove vital in understanding how social media reflects the real world people and their choices, buying selling trends and preferences. This paper aims to present a walk-through of various data mining techniques used for the massive amount of data utilization in social networks.*

Keywords— *Big Data, Social Media, Data Mining, Big Data Analysis, Social Media Analysis*

I. Introduction:

Social Media generates humungous amount of data that contains information about the internet users. This data is termed as Big Data. This data generation has become a continuous process, like every second it is being recorded and stored. Now, this big data generated by the internet users can be used for the internet users to identify their choices, their buying patterns, their preferences, their news feeds, etc. So, this data is analyzed and used to develop customized feed for every single Internet User. This customized feed can be further customized to get the feed for a single user and his particular item or brand. Big Data proves beneficial in predicting and planning future media campaign for the target consumers. To accomplish the above said tasks, Data Mining come into picture, which helps to extract hidden information, analyze relationships and patterns, etc among the transactional data generated on daily basis.[1][5]

The data being generated by Social Media includes user’s photos, videos, likes, dislikes, comments about a certain topic, feedback regarding a product, expectations for the product to be launched in near future, etc. These data also includes the detailed information about their demographic whereabouts, which is further used to plan, develop campaigns, offers and products based on this geographic information. This planning and developing various offers based on Big Data can help companies to compete better and have a winning edge in the market. This paper is divided into sections as follows: Section II & III establishes the relationship between Big Data & Data Mining and Social Media Analysis on Big Data respectively. Section IV & V describes the process and various data mining methods for Social Media Analysis respectively. Finally, Section VI has a overview of popular tools of SMA, Section VII has a brief discussion over the previous sections and concluding the paper with Section VIII with a conclusion.

II. Big Data and Data Mining

The process of extracting important usable data from humungous unstructured raw data is called Data Mining. This process can easily identify the recurring patterns present in the large dataset and establish relationships to solve problems with the use of data analysis. In simpler terms, data mining is a mining of data having the recurring patterns. Data Mining includes the collection of data and storing it into a data warehouse and later the concept of computer processing come into play. Whereas, Big Data can be termed as the collection of humungous amount of structured, semi-structured and unstructured. Big Data comprises of 5 V’s, namely, Volume, Variety, Velocity, Veracity and Value. [7]

Here’s how these two concepts coincide with each other to become something more meaningful: Data Mining uses various mathematical algorithms to extract concerned information from Big Data.

KDD stands for Knowledge Discovery in Data, which is a systematic sequential process of generating the useful information from the enormous amount of unstructured raw big data. This KDD process has many levels for its operation such as data preprocessing, data preparation, data selection, data cleansing, etc.

Broadly, there are two functions involved in Data Mining namely, Descriptive and Classification & Prediction. In Descriptive, the generic attributes of data are being handled, such as frequent pattern mining, association mining, correlation mining, cluster mining, etc. [6][8] Whereas, in Classification, the data is being handled to label all the unlabeled classes, and Prediction uses the data to predict future values or expectations such as trend, price, supply, demand, etc.

III. Social Media Analysis on Big Data

Social Media is now-a-days not limited to be possessed by Companies, Big Organizations or Brand Managers, but it is been managed and maintained by almost everyone. [9] These SM websites contains the data which can be used and studied to better understand people and their choices, which goes beyond race, sex, color, belief or societal status. People love to share their daily routine, their music preferences, their preferred shopping brands, and their favorite place to gym, their preferred place to go to, to have food, including their thoughts on the current affairs, decisions by the Government, whether they agree or disagree on some ongoing debate, and these reactions are not limited to their own Geographic location, but it has become global now. Indians react on US political scenarios, Americans reacting to Indian Prime Minister, etc. [2] These generated data can be very useful in making decisions regarding businesses, political moves, launching a new product, investing money on certain brand marketing, re-releasing a film, etc. These kinds of decision making are not limited to the usual monitoring or analysis of likes, comments, re-tweets; but it considers an in-depth study of a particular domain to reach to a point where we can take clear and certain decision.

Hence, Social Media Analysis (SMA) can be termed as the approach to collect humungous amount of raw data from Social Media websites, which gets used and evaluated to make certain profit making decisions.

IV. Process of SMA

Every analysis of Social Media has to go through a standard process, i.e. few steps to follow, namely, Capturing & Tracking the data, preparation of the data, analysis & evaluation of the data and finally summarization of the output, which is depicted in Figure 1. These predefined steps may vary with many in-between micro-steps depending upon the application, objective, or question being answered during the analysis. But, broadly speaking, every SMA has to obey these four steps in the process of analysis.

Pre-requisite for SMA:
Before starting the process of SMA, one must prepare the core process with few pre-requisite data.

- **Clearly-Defined Questions**: Before starting gathering data from various sources, one must define the objectives for the analysis in a very crystal clear manner. This step ensures the clarity in the questions to be answered eventually at the end of the analysis. This step further ensures that the overall analysis will not sidelined by any other “not-that-important” objective or question at this point. If questions are clearly identified and defined before starting the process of SMA, it solidifies that it will be a specified and much certain process of analysis.
- **Set your Goal**: This speaks how important it is to clearly set the goals to achieve at the end of the analysis. When the analysis process identifies a well-defined end point, the transition from Start State to End State may result in a quality process. Having a well-defined, well-identified end state ensures that the process may not get lost somewhere else but will definitely reaches its destination as identified and desired.
- **Define your Target**: As social media is well-spread, well-distributed among several domains, this step clearly identifies the domain where the analysis will get takes place. This ensures that the data will be gathered only from the pre-decided domain only, solidifying the scope of the analysis to be narrowed-down to only that one (or multiple) pre-decided domain.

This pre-requisite data eventually enhances the overall performance of SMA resulting in quality analysis, as this primary step identifies and shapes the analysis process in better defined way. [3] This step gives sharp vision to the objective of the overall Social Media Analysis.

1. Capturing & Tracking

This is the very first step of Social Media Analysis. It covers capturing and gathering data from various sources such as Social Media websites, News Articles, blogs, video blogs, video hosting websites, historical data, freely available databases, etc. This step prepares the foundation of the overall process; hence it is very important to identify the concerned sources to be used from extraction of data. This step also covers the idea of tracking the data such as tracking data using data access via tools, data access via APIs, RSS feeds, tracking live tweets, tracking real-time stock prices and so on. This captured, extracted and tracked data is in raw format, which gets processed and made ready for use in the next step.

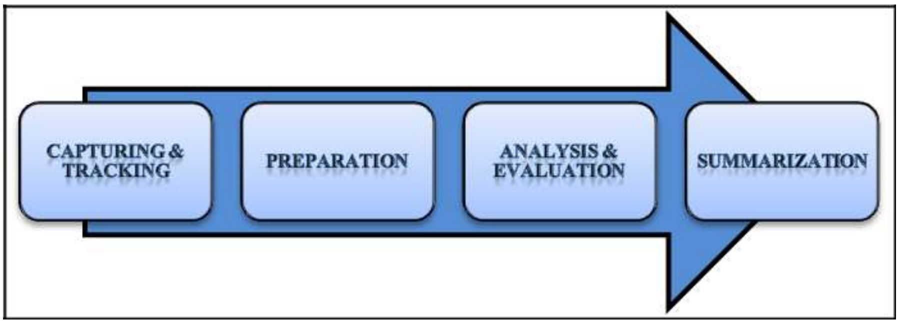

Figure 1: Process of Social Media Analysis (SMA)

2. Preparation

As the data gathered in the first step of Capturing & Tracking is in the raw format, it needs to process the data before using it. Here comes this step of Preparation of data which basically is getting the raw data ready to use by pre-processing it and extracting the information. This is very important step as any data which gets fed to any system needs to be correct and valid, as they say Garbage-In-Garbage-Out. This includes the process of text cleaning, tagging and storing. It also underlines scrubbing, cleansing, removing and validating errors, etc. This also covers the process of removing noise from the data such as irrelevant data, missing values, fixing dates and numbers, inconsistent data etc.

3. Analysis and Evaluation

After successfully extracting the information from the raw data, it gets utilized in the core step of analysis and evaluation of results. The information extracted in the previous step gets stored in relational databases, SQL databases, Open Source Databases and many other tools and using this stored information, important findings are being evaluated. This is the step where actual advanced analytics gets applied and it results in the generation of analysis findings. These findings are then evaluated. Various analysis algorithms and tools are popularly gets used and readily available across the globe. These tools may vary depending upon the domain, application or objective of the overall analysis process.

4. Summarization

The final step of SMA is summarization which basically is to present the findings evaluated in the previous step of Analysis & Evaluation. This summarization can be in different format and it may vary depending upon the domain, application or objective of the overall analysis process. For example, summarization in bars, graphs, charts, numbers, percentages, yes/or, classified labels, etc. This is also popularly referred as Visualization.

V. Methods of SMA

a. Computational Statistical Learning

It basically refers to computationally intensive statistical methods including re-sampling methods, Markov chain Monte Carlo methods, local regression, kernel density estimation and principal components analysis (PCA). Computational Statistical Learning is purely based on probability spaces, where probability spaces are defined using a sample space, a set of events and the probabilities of events. The main difference between Computational Statistical Learning (CSL) and Machine Learning is that where CSL refers to infer relationship between different variables, ML refers to predict the future values. [9] As previously, due to lack of having computational requirements, CSL were mostly used. But with time, ML bounced back and now it taking over and gripping to almost every single domain.

b. Machine Learning:

It refers to the techniques which can learn from the experience using systematic acquisition of data from various sources. ML is basically a subset of AI, which got extended from Computational Statistical Learning. In CSL, as the numerical statistical manipulations were bit complex, difficult and time-consuming, with ML, the same mathematical manipulations can be implemented comparatively easily. There are mainly two categories in Machine Learning i.e. Supervised and Unsupervised Learning.

2a. Supervised Learning:

It is called supervised learning as here training data is available and by developing a predictive model based on both input and output, prediction is performed. Here the classes are pre-labeled and clearly identified.

Classification: It is a method which is used to create or build the system that can identify different classes or labels and classify the data accordingly. It generates categorical output. For example, you can use this to find how people are reacting to the ongoing debate and classify their opinions into two classes of a yes or a no.

Naïve Bayes Classifier (NB): Bayes Theorem is based on the probability that an event will occur given that another event has happened. Here, Naïve Bayes Classifier calculates the likelihood that the vector belongs to each class. NB Classifier is a simple model, fast, scalable that requires little data.

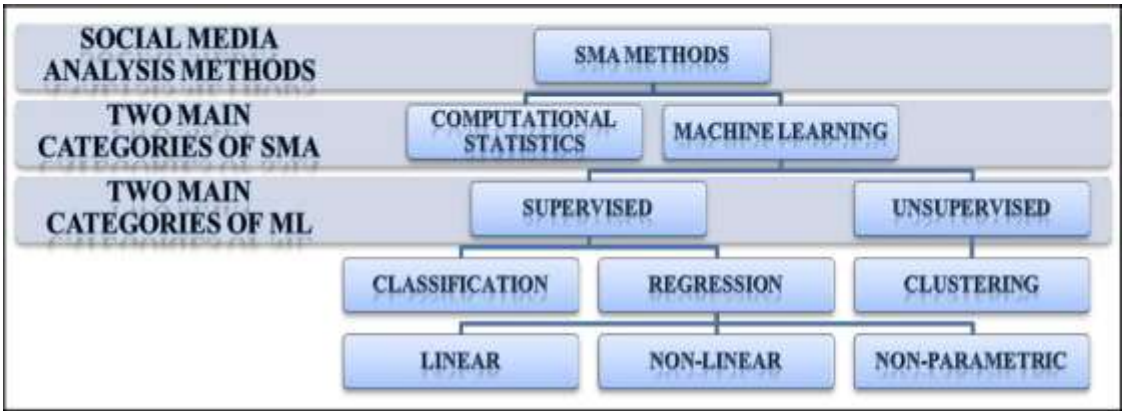

Figure 2: Methods of Social Media Analysis (SMA)

Regression: Regression is a method for analysis which uses various variables to find and analyze the relationship between them. It generates numerical output. Popular regression methods are linear regression, non-linear regression and non-parametric regression. Logistic regression is another popular regression method which is mostly used after linear. They are almost similar to each other but where Linear Regression focuses on predicting values, Logistic Regression focuses on classification tasks.

Support Vector Machine (SVM): It is basically a method to find a hyperplane in N-dimensional space to distinctly differentiate and eventually classify the data points. SVM has good accuracy and it uses less computational power as compared to many popular classification methods. SVM can be used for regression and classification both.

2b. Unsupervised Learning:

It is called unsupervised learning as here training data is available and by developing a predictive model based on input data, prediction is performed. Here the classes are not pre-labeled and not well identified.

Clustering: Clustering allows the use common attributes in different classifications to identify clusters in automatic manner. These classes are not label, unlike method of classification. This generates categorical output. This method is used when there are no certain clear labeled classes in the analysis process, so this process of clustering creates clusters with similar attributes. As the classes are not labeled, this comes under the category of unsupervised learning. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

VI. Tools for SMA

There is a wide range of several tools for SMA is readily available and popularly used by researchers, influencers and marketing agents all across the globe. Broadly, it can be classified into three types, Programming Based Tools, Application Based Tools and Platform Based Tools. This classification and their popular tools under each class have been briefly depicted in Figure 3. [4][10]

R is widely used for statistical programming, MATLAB is popularly used for numeric scientific programming, Mathematica is used for symbolic scientific programming, Python is very widely used tools used for prediction, Natural Language Processing (NLP), etc. There are many other popular tools which can be used for SMA are Apache UIMA, SAS, RapidMiner, Orange Data Analysis, Weka, Neo4j, etc. [5][11]

There are many several SMA tools which are domain/application oriented such as Keyhole used for customer analysis and it is widely used by influencers, AgoraPulse is a tool mostly used for choosing the best content a user must see based on their interests, BrandWatch is a tool mostly used for making wise and smart marketing decisions, BrandMentions is another powerful media marketing tool, RivalIQ is a SMA tool for monitoring the media presence and performance of your brand/company in the overall market, Mention is one of the most important tool which is popularly used by the analyst to monitor and count the number of mentions of a particular term over social media which give an insight about how popular that product/topic is trending, etc.

There are several platform specific SMA tools which use Instagram, Twitter, Facebook, Pintrest and other popular social media platforms. The popular ones are Snaplytics that can measure your brand/organization performance and presence over Snapchat and Instagram, Minister.io can track your Instagram audience engagements, Falcon.io is used to understand how viral your content is on Facebook, TweetReach is one of the popular tool which can let you understand the number of impressions and overall audience engagement using the tweets, Tailwind is used to analyze the posts of Pintrest and Instagram which gives you insight about the overall performance which is used further to amplify the reach and overall presence over Social Media.

VII. Discussion

As an explosion of data being generated by Social Media Networks, Social Media Analysis (SMA) helps us to use the same data and to better understands the consumers, customers, audience and their mindsets. With an exponential increase in the use of Social Media all over the globe, there are many Social Media Analysis tools are available which gives us internal insight of the audiences. There tools are basically belongs to Data Mining and Machine Learning categories due to their advantages over the several others. Now, most of the big organizations and brand companies use these tools to better understand their footprints all over the market and eventually plan their future marketing ideas and product launches. There are many issues in social media analysis and each issues has to be identified depending upon human interaction with social such as Community or Group Detection, Expert Finding, Link Prediction, Predicting Trust and Distrust among individuals, Behavior and Mood Analysis, Opinion Mining etc.

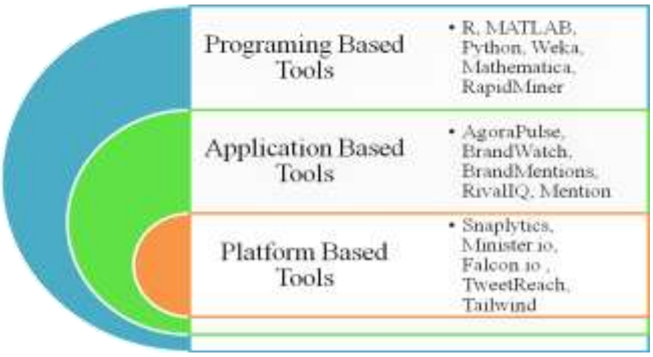
Programing Based Tools
• R, MATLAB, Python, Weka, Mathematica, RapidMiner
Application Based Tools
• AgoraPulse, BrandWatch, BrandMentions, RivalIQ, Mention
Platform Based Tools
• Snaplytics, Minister.io, Falcon.io , TweetReach, Tailwind

Figure 3: Tools of Social Media Analysis (SMA)

VIII. Conclusion

Social media remains the most effective means for brands to engage directly with their end customers, understand them, and what they want and eventually, Social Media Analysis has become a very popular, powerful domain which uses Big Data being generated by Social Media on daily basis. It is used to create useful info from raw data. This info can further analyzed to extract patterns and hidden information. It can be used for better marketing strategies, better product placement, better prediction on product consumption, better forecasting of future trends etc.

The SMA tools can also be used to develop new data mining and new machine learning algorithms for social networks which can directly create new applications in other areas of research like crime agencies, supermarkets, retailers, service providers etc. This is a very powerful domain for research and it is still in exploration mode and with time many researchers are studying and exploring it all across the globe.

References

[1]. Kumar, Vikas & Nanda, Pooja. (2019). Social Media to Social Media Analytics: Ethical Challenges. International Journal of Technoethics. 10. 57-70. 10.4018/IJT.2019070104.

[2]. Sebei, Hiba & Hadj Taieb, Mohamed Ali & Ben Aouicha, Mohamed. (2018). Review of social media analytics process and Big Data pipeline. Social Network Analysis and Mining. 8. 10.1007/s13278-018-0507-0.

[3]. Stieglitz, Stefan & Mirbabaie, Milad & Ross, Björn & Neuberger, Christoph. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management. 39. 156-168. 10.1016/j.ijinfomgt.2017.12.002.

[4]. Peng, Sancheng & Yu, Shui & Mueller, Peter. (2018). Social networking big data: Opportunities, solutions, and challenges. Future Generation Computer Systems. 86. 1456-1458. 10.1016/j.future.2018.05.040.

[5]. Aydin, Nursen. (2018). Social Network Analysis: Literature Review. 34. 10.5824/1309-1581.2018.4.005.x.

[6]. Elangovan D, Dr.Subedha V, Sathishkumar R, Ambeth kumar V D, A Survey: Data Mining Techniques for Social Media Analysis (2018). International Conference for Phoenixes on Emerging Current Trends in Engineering and Management (PECTEAM 2018), Advances in Engineering Research (AER), volume 142

[7]. Sponder, Marshall & Khan, Gohar. (2017). An Introduction to Social Media Analytics. 10.4324/9781315640914-8.

[8]. Gonçalves, Alex. (2017). The Future of Social Media Analytics. 10.1007/978-1-4842-3102-9_19.

[9]. Injadat, Mohammadnoor & Salo, Fadi & Nassif, Ali. (2016). Data Mining Techniques in Social Media: A Survey. Neurocomputing. 214. 10.1016/j.neucom.2016.06.045.

[10]. Chen, C. & You, Xinge & Tao, Dacheng. (2016). Big Learning in Social Media Analytics. Neurocomputing. 204. 10.1016/j.neucom.2016.02.069.

[11]. Fan, Weiguo & Gordon, Michael. (2014), The Power of Social Media Analytics. Communications of the ACM. 57. 74-81. 10.1145/2602574.