IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 18-23
www.iosrjournals.org

Social Media and Multimedia Data Analytics through Machine Learning

Vinod Desai [1] Dr. Dhanasekaran K[2], Aneel Narayanapur3, Vinayak Joshi[4],
[1](CSE, Angadi Institute Of Technology & Management, Belagavi, India)
[2](CSE, Jain College Of Engineering, Belagavi, India)
[3](CSE, Angadi Institute Of Technology & Management, Belagavi, India)
[4](CSE, Angadi Institute Of Technology & Management, Belagavi, India)

Abstract: *Machine learning is a subfield of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data. Both systems search through data to look for patterns. Spam filtering, face recognition, recommendation engines, when you have a large data set on which you'd like to perform predictive analysis or pattern recognition, machine learning is the way to go. This science, in which computers are trained to learn from, analyze, and act on data without being explicitly programmed, has surged in interest of late outside of its original cloister of academic and high-end programming circles. This rise in popularity is due not only to hardware growing cheaper and more powerful, but also the proliferation of free software that makes machine learning easier to implement both on single machines and at scale. The diversity of machine learning libraries means there's likely to be an option available regardless of what language or environment you prefer. This paper presents Data Analytic techniques which provide functionality for individual apps or whole frameworks, such as Hadoop.*

Keywords: *Big Data, Big Data Application, Text Analytics, Audio Analytics, Video Analytics.*

I. Introduction

The term big data is used to describe the growth and the availability of huge amount of structured and unstructured data. Big data are beyond the ability of commonly used software tools to create, manage, and process data within a suitable time. Big data is important because the more data we collect the more accurate result we get and able to optimize business processes. The Big data is very important for business and society purpose. The data came from everywhere like sensors that used to gather climate information, are available as posts or shared data on the social media sites, video movie audio etc. This collection of data is called Big Data. Now a days this big data is used in multiple ways to grow business and to know the world [1][2][15]. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data processing has a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Wal-Mart handles more than 1 million customer transaction every hour. Facebook handles 40 billion photos from its user base. Big data require some technology to efficiently process large quantities of data. It uses some technology like, data fusion and integration, genetic algorithms, machine learning, and signal processing, simulation, natural language processing, time series Analytics and visualization [12][13][16].

1.1. Big Data Features:

Big data contain many features such as a Volume, Velocity, Variety, Variability, Complexity, and Value.

II. Analytics In Machine Learning

Types of big data analytics are: Prescriptive: - This type of analytics help to decide what actions should be taken. It is very valuable but not used largely. It focuses on answer specific questions like, hospital management, diagnosis of cancer patients, diabetes patients that determine where to focus treatment. Predictive: - This type of analytics help to predict future or what might be happen. For example some companies use predictive analytics to take decision for sales, marketing, production, etc. Diagnostic: - In this type, look at past and analyze the situation what happen in past and why it happen and how we can overcome this situation. For example, weather prediction, customer behavioral analysis etc. Descriptive:-It describes what is happening

currently and prediction near future. For example, market analysis, contains behavioral analysis etc. By using appropriate analytics organization can increase sales, increase customer service, and can improve operations. Predictive Analytics allow organizations to make better and faster decisions [1][2][4][10].

2.1. Predictive Analytics

Predictive Analytics is a method through which we can extract information from existing data sets to predict future outcomes and trends and also determine patterns. It does not tell us what will happen in future. It forecasts what might happen in future with acceptable level of reliability. It also includes what if-then-else scenarios and risk assessment. Applications areas of Predictive Analytics are CRM (Customer Relationship Management), Clinical Decision Support, Collection Analytics, Cross Sell, Customer Retention, Direct marketing, Fraud detection, Portfolio, product or economy-level prediction, Risk management, Underwriting

2.2. Usage of Big Data Analytics in India through Machine Learning

From predicting ticket confirmations of trains to checking for water supply leakages and even for finding the perfect bride and groom, Big Data is being used in a number of creative ways in India. Following are few uses of Big Data Analytics in India in last few years [3][9].

a) Win elections (exit poll), (b)Finding a perfect match, (c) Detecting water leakages, (d) Gaining insights into shopping behavior, (e)Ensuring proper water supply,)f)Improve India's financial inclusion ratio (g)Improve product development, (g)Predict ticket confirmations for trains.

III. Social Media Analytics

The Social Media analytics is collecting information or data from the social media websites; blogs etc. and uses it in business purpose or decision making. Nowadays Social Media is the best platform for understanding the real-time customer choice or intentions and sentiments, using social media business advertising, product marketing easily. EBay.com uses two data warehouses at 7.5 peta bytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse, Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. Facebook handles 50 billion photos from its user base. As of August 2012, Google was handling roughly 100 billion searches per month.

3.1. Applications: a) Natural language processing, (b) Natural language understanding, (c) Optimization and meta heuristic, (d) Online advertising , (e) Recommender systems, (f) Robot locomotion, (g) Search engines , (h) Sentiment analysis (or opinion mining) ,(i) Sequence mining, (j) Software engineering

3.2. Challenges of social media analytics

a) Massive amounts of data require lots of storage space and processing power.
b) Shifting social media platforms.
c) Worldwide online accessibility provides more data in many languages.
d) Evolution of online language.

IV. Content Based Analytics

Content Based Analytics means whatever data that store in social media back-end site. For example Facebook users store their data, photos, and videos on Facebook storage. For this content they need big amount of storage but nowadays number of users increasing rapidly so, social networking sites like Facebook, twitter, WhatsApp need to increase their storage capacity day by day and that's the obstacle because they don't know how much of storage capacity they need to increase.

Content based predictive analytics recommender systems mostly match features (tagged keywords) among similar items and the user's profile to make recommendations. When a user purchases an item that has tagged features, items with features that match those of the original item will be recommended. The more features match the higher probability, then the user will like the recommendation. This degree of probability is called precision. User-based tagging, however, turns up other problems for a content-based filtering system (and collaborative filtering) like: Credibility, Scarcity, Inconsistency.

4. 1. Precision with constant feedback

One way to improve the precision of the system's recommendations is to ask customers for feedback whenever possible. Collecting customer feedback can be done in many different ways, through multiple

channels. Some companies ask the customer to rate an item or service after purchase. Other systems provide social-media-style links so customers like or dislike a product.

4.2. Measurement for effectiveness of system recommendations

The success of a system's recommendations depends on how well it meets two criteria: precision (think of it as a set of perfect matches, usually a small set) and recall (think of it as a set of possible matches usually a larger set). Issues in measurement for effectiveness:

a) Precision measures how accurate the system's recommendation was. Precision is difficult to measure because it can be subjective and hard to quantify.
b) Some recommendations may connect with the customer's interests but the customer may still not buy. The highest confidence that a recommendation is precise comes from clear evidence: The customer buys the item. Alternatively, the system can explicitly ask the user to rate its recommendations.
c) Recall measures the set of possible good recommendations your system comes up with. Think of recall as an inventory of possible recommendations, but not all of them are perfect recommendations. There is generally an inverse relationship to precision and recall. That is, as recall goes up, precision goes down, and vice versa.

The ideal system would have both high precision and high recall. But realistically, the best outcome is to strike a delicate balance between the two. Emphasizing precision or recall really depends on the problem you're trying to solve.

V. Text Analytics

Most of all information or data is available in textual form in databases. From these contexts, manual Analytics or effective extraction of important information are not possible. For that it is relevant to provide some automatic tools for analyzing large textual data. Text analytics or text mining refers to the process of deriving important information from text data. It is used to extract meaningful data from the text. It uses many ways like associations among entities, predictive rules, patterns, concepts, events etc. based on rules. Text analytic is widely used in government, research, and business needs. Data simply tells you what people did but text analytics tell you why. All information will be retrieved from unstructured or semi structured text data. From all textual data it will extract important information. After extracting information it will be categorized. And from these categorized information we can take decision for business.

5.1. Steps in Text Analytics system:

a) Text: In initial stage data is unstructured.
b) Text processing: All information will transfer in Semantic Syntactic text.
c) Text transformation: The important text will extract for future use.
d) Feature selection: The data is counted and display in the Statistics format.

Data mining: All data is classified and clustered.

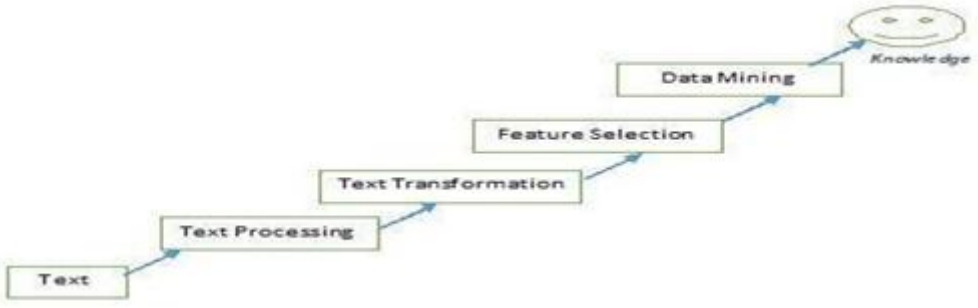

Fig 5.1: The steps in Text Analytics system

5.2. Text Analytics application areas: (a)Security application,(b) Marketing application, (c) Analyzing open, (d) Automatic process on emails and messages.

5.3. Distinct Aspects of Text in Social Media:

a) Time Sensitivity: An important feature of the social media services is their real-time nature. With the rapid growth of the content and communication styles, text is also changing. Because of the time sensitivity of the textual data the people's thoughts also changes from time to time.

b) Short Length: Successful processing of the short texts is essential for the text analytics method. As the messages are short, it makes people more efficient with their participation in social networking websites. Short messages are used in social media which consists of few phrases or sentences.
c) Unstructured Phrases: An important difference between the text in social media and traditional media is the difference in the quality of content. Different people posts different things according to their knowledge, ideas, and thoughts. When composing a message also have many new abbreviations and acronyms for e.g. How r u? Gr8, are actually not words but they are popular in social media.

5.4. Applying Text Analytics in Social Media:
a) Event Detection: It aims to monitor a data source and detect the occurrence of an event that is to be captured within that source. These data sources include images, videos, audios, text documents.
b) Collaborative Question Answering: As social networking websites has emerged, the collaborative question answering services have also emerged. It includes several expert people to answer the questions posted by the people. A large number of questions and answers are posted on the social networking websites.
c) Social Tagging: Tagging of the data has also increased to a great extent. For example when any particular user is looking or searching for a recent event like Bihar Election then the system will return the results that are tagged as Bihar or Election.

Textual data in social media provides lots of information and also the user-generated content provides diverse and unique information in forms of comments, posts and tags.

VI. Multi-Media Analytics

Audio analytics is the process of compressing data and packaging the data into single format called audio. Audio Analytics refers to the extraction of meaning and information from audio signals for Analysis. There are two ways to represent the audio Analytics: 1) Sound Representation 2) Raw Sound Files. Audio file format will store digital audio data on a system. There are three audio formats: Uncompressed audio format, Lossless compressed audio format, Lossy compressed audio format[11].

6.1. Application Area of Audio Analytics: The audio is the file format that used to transfer the data from one place to another. Audio analytics is used to check whether given audio data is available in proper format or in similar format that sender sends. The Application of audio Analytics are many: Surveillance application, Detection of Threats, Tele-monitoring System, Mobile Networking System.

Video is a major issue when considering big data. Videos and images contribute to 80 % of unstructured data. Now a days, CCTV cameras are the one form of digital information and surveillance. All these information is stored and processed for further use, but video contains lots of information and is generally large in size. For example YouTube has innumerable videos being uploaded every minute containing massive information. Not all video are important and viewed largely. This creates a situation where videos create a junk and hard-core contribution to big data problems. Apart from videos, surveillance cameras generate a lot of information in seconds. Even a small Digital camera capturing an image stores millions of pixel information in milli seconds. Video Data Analytics dimensions - Volume: Size of video being more, takes the network as well as the server time for processing. Low bandwidth connections create traffic on network as these videos deliver slowly. When stored on mass storage on secondary storage requires huge amount of space and takes more time for retrieving and processing. Variety: Videos consisting of various format and variety such as HD videos, Blu-ray copies etc. Velocity: It is speed of data. Nowadays, Digital cameras process and capture videos at a very high quality and high speed. Video editing makes it to grow in size as it contains other extra information about the videos. Videos grow in size faster as they are simply nothing but collection of images [7].

6.2. Application of video analytics: Useful in accident cases, Useful in schools, traffic police, business, security etc., Video Analytics for investigation (Video Search), Video analytics for Business Intelligence, Target and Scene Analytics, Direction Analytics, Remove the human equation through the automation.

VII. Results And Discussions

Before concluding this overview, we would like to discuss a few aspects and point out directions for further work. The hope is that readers will also have a better outlook of our quest. We welcome any feedback and suggestion. Test Automation Evaluation of a cluster system that is formed using new software and hardware is a challenging endeavor. A statistics-based approach is often practical and fruitful. Nevertheless, such an approach demands the formulation of a comprehensive test matrix, numerous test runs, careful analysis of results, and skillful data reduction and graphing to gain insights and convey understanding. Both the software and hardware should be sufficiently instrumented so as to be able to provide enough meaningful statistics. At the same time, such instrumentation must be light-weight so as not to impact the observed system performance.

To automate testing, a Python based test framework has been developed as the basis for a high-degree of test automation, in conjunction with an extensive use of ansible. For gathering system run-time statistics, we have been using collectd. For monitoring and visualization, we use graphite. Recently, we are introducing Grafana26 with graphite as its data source to enhance the visualization aspect. Zettar zx is capable of providing many run-time statistics which can be gathered using RESTful APIs. To observe its run-time statistics, both its built-in Web UI and/or a stand-alone CLI tool can be used. For casual statistics work, we use spreadsheet tools such as MS Excel and/or OpenOffice Calc.

7.1 Test Automation

Evaluation of a cluster system that is formed using new software and hardware is a challenging endeavor. A statistics-based approach is often practical and fruitful. Nevertheless, such an approach demands the formulation of a comprehensive test matrix, numerous test runs, careful analysis of results, and skillful data reduction and graphing to gain insights and convey understanding. Both the software and hardware should be sufficiently instrumented so as to be able to provide enough meaningful statistics. At the same time, such instrumentation must be light-weight so as not to impact the observed system performance. To automate testing, a Python based test framework has been developed as the basis for a high-degree of test automation, in conjunction with an extensive use of ansible. For gathering system run-time statistics, we have been using collectd. For monitoring and visualization, we use graphite. Recently, we are introducing Grafana26 with graphite as its data source to enhance the visualization aspect. Zettar zx is capable of providing many run-time statistics which can be gathered using restful APIs. To observe its run-time statistics, both its built-in Web UI and/or a stand-alone CLI tool can be used. For casual statistics work, we use spreadsheet tools such as MS Excel and/or Open Office Calc.

7.2 Encryption approach, cipher suite choice, and system energy-consumption

The fact that encryption increases CPU utilization should be evident. Zettar zx uses the Botan cry to library implemented in C++, and as a result has the flexibility of selecting both a desired encryption approach (TLS only or TLS + PFS) and cipher suite. A few obvious questions are: which encryption should be the default choice? Other than Botan's default cipher suite for each, what is the impact to data transfer performance of other supported cipher suites? For each combination of encryption approach and cipher suite, what is the impact to CPU utilization and thus energy consumption of the data transfer system? What's a good way to measure, record, and analyze such consumption rates? To the authors, there seems to be a dearth of literature for such highly practical engineering issues. We intend to address this shortage of useful references as much as we can in 2015 and beyond.

7.3 Hardware-assisted encryption and hash computation

The basic system will be upgrade to use the new Has well micro architecture featured in the 4th generation Intel® Core™ Processor family that implements several new instructions designed to improve cryptographic processing performance. We need to work with Intel to integrate this into our research.

7.4 Impact of compilers

Zettar zx is implemented in the C++ programming language, so we need a compiler to support the aforementioned Haswell instruction set. It should be of strong interest to compare the executables generated using different compilers, e.g. GCC (system default), Clang, and Intel C++ compiler.

7.5 Impact of WAN latency

The basic system so far have both clusters mounted on the same rack and connected using short patch cables, therefore it doesn't simulate data transfers over long distance. Since Zettar zx uses TCP for data transfers, thus the impact of the WAN (delays etc) on performance must be evaluated. For example, the software uses multiple distributed and randomized TCP streams to minimize the impact of TCP's two basic default behavior: slow start and collision avoidance, Nevertheless, the effectiveness of the approach must be evaluated over high-latency WAN environment – a major effort of 2015.

VIII. Conclusion And Future Work

Although regression is an important problem in data analysis, it has not been dealt with extensively by the ML community. In this paper we presented a system capable of learning regression rules. The system integrates the task of developing a regression model from data, with the technique of searching for logical conditions that enable a better fitting error by the model. Although regression trees also follow a similar strategy, $\mathbf{R^2}$ uses a more powerful descriptive language. Piecewise regression models like the ones built

by $\mathbf{R^2}$ have advantages of achieving better prediction accuracy when compared to one-model approaches. These advantages come at the cost of more specific models. $\mathbf{R^2}$ implements a flexible compromise between model generality and correctness. Our system compares reasonably to other ML regression algorithms and even outperforms them on some data sets. However, due to the size of the data sets many differences are not statistically significant.

In future we plan to extend our comparisons and weigh both accuracy and comprehensibility. These comparisons should include systems from other non-symbolic fields. We believe that this will show the advantage of $\mathbf{R^2}$ over other sub-symbolic methods. We also intend to explore techniques that enable our system to deal with large scale domains. We think that methods of sampling and iterative regression modelling will help to overcome these problems.

References

[1]. Aha,D. Kibler,D. (1991): Instance-Based Learning Algorithms. In *Machine Learning, vol. 6 - 1*. Kluwer Academic Publishers.

[2]. Breiman, L. , Friedman,J.H., Olshen,R.A. & Stone,C.J. (1984): *Classification and Regression Trees*, Wadsworth Int. Group, Belmont, California, USA.

[3]. Clark, P., Niblett, T. : Induction in noisy domains, in *Proc. of the 2th European Working Session on Learning* , Bratko,I. and Lavrac,N. (eds.), Sigma Press, Wilmslow, 1987.

[4]. Dillon,W., Goldstein,M. (1984) : *Multivariate Analysis methods and applications*. John Wiley & Sons.

[5]. Friedman,J. (1991): Mutivariate Adaptative Regression Splines. In *Annals of Statistics* , 19:1.

[6]. Friedman,J. Stuetzle,W. (1981): Projection Pursuit Regression. In *J. American Statistics Association* 76.

[7]. Karalic, A.. (1991): The Bayesian Approach to Tree-Structured Regression. In *Proceedings of ITI-91* , Cavtat, Croatia, 1991.

[8]. Karalic, A..(1992): Employing Linear Regression in Regression Tree Leaves. In *Proceedings of ECAI-92* , Wiley & Sons, 1992.

[9]. McClelland,J. Rumelhart,D. (1988): *Explorations in Parallel Distributed Processing*. Cambridge, Ma. : MIT Press.

[10]. Michalski, R.S. , Mozetic, I. , Hong, J., Lavrac, N. : The multi-purpose incremental learning system AQ15 and its testing application to three medical domains, in *Proceedings of AAAI-86*, 1986.

[11]. Michie,D., Spiegelhalter,D.J., Taylor,C. (1994) : *Machine Learning, Neural and Statistical Classification*. Ellis Horwood series in Artificial Intelligence. Ellis Horwood.

[12]. Quinlan, J.R. (1992): Learning with Continuos Classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*. Singapore: World Scientific, 1992.

[13]. Quinlan, J.R. (1993): Combining Instance-Based and Model-Based Learning. In *Proceedings of 10th IML*, Utgoff,P. (ed.). Morgan Kaufmann Publishers.

[14]. Rissanen,J. (1983) : A universal prior for integers and estimation by minimum description length. In Annals of Statistics 11, 2.

[15]. Torgo,L. (1993a) : Controlled Redundancy in Incremental Rule Learning. In *Proceedings of ECML-93*, Brazdil,P. (ed.). Lecture Notes in Artificial Intelligence - 667. Springer-Verlag.

[16]. Torgo,L. (1993b) : Rule Combination in Inductive Learning. In *Proceedings of ECML-93*, Brazdil,P. (ed.). Lecture Notes in Artificial Intelligence - 667. Springer-Verlag.

[17]. Torgo,L. (1995) : Applying Propositional Learning to Time Series Prediction. In *ECML-95 workshop on Statistics, Machine Learning and Knowledge Discovery in Databases*. Kodratoff, Y. et al. (eds.).

[18]. Urbancic,T., Bratko,I. (1994): Reconstructing Human Skill with Machine Learning. In *Proceedings of the European Conference in Artificial Intelligence (ECAI-94)*, Cohn, A.G. (ed.). John Wiley & Sons.

[19]. Weiss,S.M., Indurkhya,N. (1993): Rule-Based Regression. In *Proceedings of IJCAI-93*, Bajesy,R. (ed.). Morgan Kaufmann Publishers.