Review on Ontology based Rational Analysis in Automobile Sector via Tweeter using MapReduce and K-Means

Shelly[1], Gaurav Garg[2],
[1]*Perusing M-Tech, Department of CSE, AITM at Palwal, Haryana, India*
[2]*Assistant Professor, Department of CSE, AITM at Palwal, Haryana, India*
(E-mail: singh.shelly165@gmail.com)

***Abstract*—** Corpus Data alludes to an accumulation of tremendous datasets. Slant examination added to a famous research zone for twitter meant for opinion mining. The opinion mining is managed with the extraction neglects to give the profound outcome about the client's supposition, sentiments, behavior and opinions at the same time on specific context and highlights the space and structure of metaphysics which helps in getting the refined conclusion, investigation and directions therein based on recommendations and references. However, Ontology implies a formal, unequivocal particular of a common conceptualization for forming the context where the reference exists and mutates other conceptualizations alludes to a unique model for evaluation, assessments and estimation to the object and its references. Subsequently, under this scheme, we have utilized ontology to examine the tweets to build enlargement and effectiveness of assumptions which is acquired utilizing the K-Means and Map-Reduce framework. The work is done in the accompanying stages. In the principal arrange, the tweets are extricated from flume and put away in a vault of Hadoop ecosystem using Hadoop Distributed File System . At that point, sentences are extricated one by one. Sentences separated are improved by evacuating stop words and excess words termed as pre-processing. In the Second stage, the words left in the sentences are utilized for sense coordinating utilizing WordNet-an online semantic lexicon or dictionary. WordNet lexicon or dictionary is utilized to remove meaningless words or word-sets from tweets and features will be extracted of well defined dictionary words. In the Third stage, Ontology is being created by utilizing XML scripted unified code. Thereafter, the crawler is being planned to get insights regarding the automobile area. The information is put away in a content way. In the fourth stage, Mapping of information is done which incorporates mapping of ontology with the crawler information, together with philosophy approval. In the fifth stage, Analysis of tweets is finished utilizing philosophy by applying K-Means and Map Reduce framework and examination of automobile sector and improvised resolution and solution which may refer to the traits that other vehicle does not fall into this class and category for effective and accurate results with comparison.

Keywords*—Machine Learning, Ontology, Hadoop, Map-Reduce, Hodoop Distributed File System, K-Means.*

I. INTRODUCTION

Corpus Data is alluded to as a gathering of enormous datasets having an expansive amount of data. Corpus Data is delivered from fluctuated sources like interpersonal interaction locales including Face book, Twitter, and so forth, and in this manner, the information which is created is in raw, semi-organized, organized, semi-organized or unstructured configuration modes. Twitter is considered as a typical research territory for estimation investigation, behavior analysis, opinion mining, and sentiment analysis. For an assortment of spaces, it offers different points of interest based on context and the area of subject and object where the manifestation exists. The supposition examination or opinion mining is arranged without extraction of highlights neglects to give the profound outcome containing the client's assessment, behavior, sentiments, and opinions therein which even. highlights of the area which can be mined by building up an ontology that helps in acquiring the refined information for investigation and can provide rational analysis on the respective context or subject. Ontology is alluded to as a formal, unequivocal particular of a mutual conceptualization formed within the dynamic model of the subject which exists in the middle of manifestation and rational analysis. The connection between ontology is the idea that relations should be unequivocally characterized using the framework. Further, ontology models must be machine-intelligible and furthermore must catch the consensual learning that can be acknowledged by the entire systems to evaluate the best potential results. Ontology assumes a crucial job in sharing and reuse of learning. Data association, the user just need the basic understanding which can improve the basis and contextual model of ontology. In the regions were managing a tremendous measure of circulated and heterogeneous PC based data, similar to World Wide Net, Intranet data frameworks, or electronic business data, Ontology offers a noteworthy job to model the environment for rational analysis. The requirement for utilizing philosophy is as per the following. Right off the bat, to share the regular comprehension of the structure of learning among individuals and programming specialists. For instance, there are different distinctive locales containing restorative data or supporting medicinal web-based business

administrations. In the event that these locales have shared and distributed the comparatively hidden ontology of the different terms they all utilization, at that point the data from various medicinal destinations can be separated and amassed by PC operators. The users will utilize the totaled learning to answer and inquiries the info record to various applications. Furthermore, for empowering reuse of the area learning. For instance, models for a few totally extraordinary areas need to speak to the thought of your time. This representation incorporates the thoughts of your time interims, focuses in time, relative proportions of your time, etc. In the event that a solitary gathering of analysts grows such ontology in detail, others will just apply and reuse it for his/her areas. Besides, in the event that we wish to fabricate an enormous philosophy, one can coordinate it with different existing ontology depicting portions of the monstrous area of automobile domain for rational analysis using Map and Reduce using Hadoop and K-Means which is the machine learning algorithm. Below the explanation is depicted for the terms proposed in scheme for perusal and ready reference.

Twitter: Twitter is a miniaturized scale blogging device where clients select in to get and send very short content - or tweets - with others. Or then again, in layman's terms, it's an approach to share musings and thoughts in 280 characters or less.

Ontology: Ontology is the philosophical investigation of being all the more comprehensively, it contemplates ideas that straightforwardly identify with being, specifically, getting to be, presence, reality, just as the essential classifications of being and their relations. Customarily recorded as a piece of the real part of reasoning known as mysticism, metaphysics regularly manages questions concerning what substances exist or might be said to exist and how such elements might be gathered, related inside a chain of importance, and subdivided by similitude's and contrasts. The below figure depicts the generic reference to elaborate the nitty-gritty of ontology used in automobile domain.

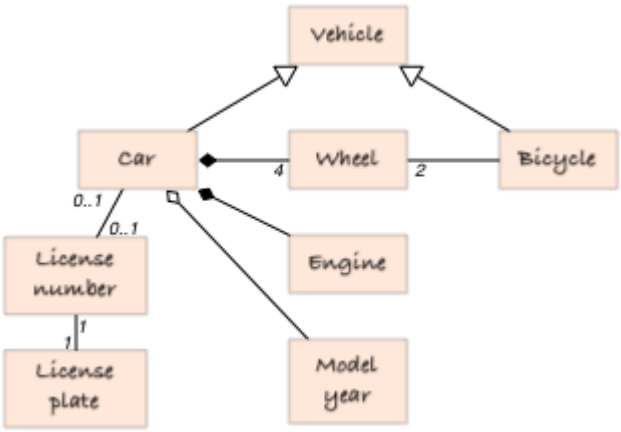

Figure 1: Ontology Example for Automobile Domain

Hadoop: Hadoop is an open source circulated preparing structure that oversees information handling and capacity for enormous information applications running in bunched frameworks. It is at the focal point of a developing environment of huge information advances that are essentially used to help progressed examination activities, including prescient investigation, information mining and AI applications. Hadoop can deal with different types of organized and unstructured information, giving clients greater adaptability for gathering, preparing and breaking down information than social databases and information stockrooms for mammoth storages. Figure 2 depicts the ecosystem comprising of various application and models which constitutes the ecosystem.

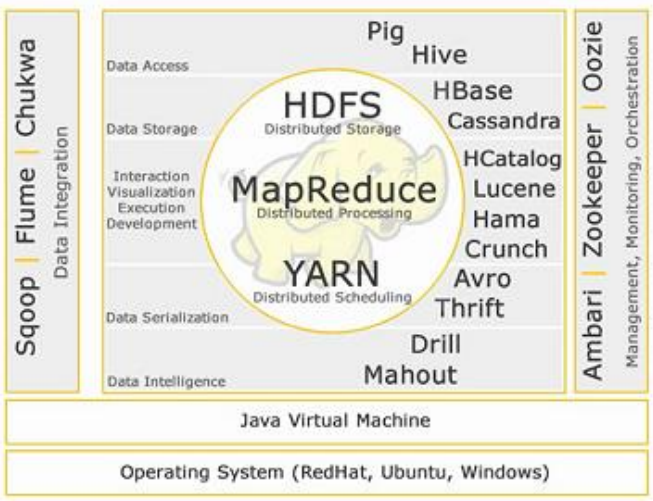

Figure 2: Hadoop Ecosystem

Hadoop Distributed File System: The Hadoop Distributed File System (HDFS) is the essential information stockpiling framework utilized by Hadoop applications. It utilizes a NameNode and DataNode design to execute a circulated record framework that gives superior access to information crosswise over very versatile Hadoop groups. Below Figure 3 depicts the scaffold of Hadoop Distributed File System.

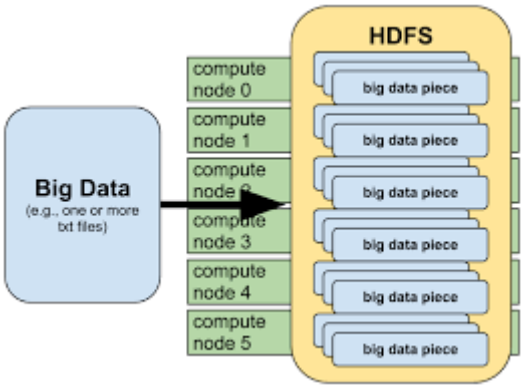

Figure 3: Hadoop Distributed File System

MapReduce: A MapReduce program is made out of a guide strategy (or technique), which performs separating and arranging, (for example, arranging understudies by the first name into lines, one line for each name), and a lessen technique, which plays out an outline task, (for example, including the number of understudies in each line, yielding name frequencies). The "MapReduce System" (additionally called "foundation" or "structure") organizes the handling by marshaling the dispersed servers, running the different

undertakings in parallel, dealing with all interchanges and information exchanges between the different pieces of the framework, and accommodating repetition and adaptation to internal failure. The model is a specialization of the part apply-join methodology for information examination. It is motivated by the guide and lessons works ordinarily utilized in practical programming, in spite of the fact that their motivation in the MapReduce structure isn't equivalent to in their unique structures. The key commitments of the MapReduce structure are not the real guide and decrease capacities (which, for instance, look like the 1995 Message Passing Interface standard's diminish and disperse tasks), however the adaptability and adaptation to non-critical failure accomplished for an assortment of uses by advancing the execution engine[citation needed]. All things considered, a solitary strung usage of MapReduce will normally not be quicker than a conventional (non-MapReduce) execution; any increases are generally just observed with multi-strung executions on multi-processor equipment. The utilization of this model is helpful just when the improved appropriated mix activity (which lessens arrange correspondence cost) and adaptation to non-critical failure highlights of the MapReduce system become an integral factor. Improving the correspondence cost is fundamental to a decent MapReduce calculation. The below figure 5 depicts the architecture of MapReduce Algorithm.

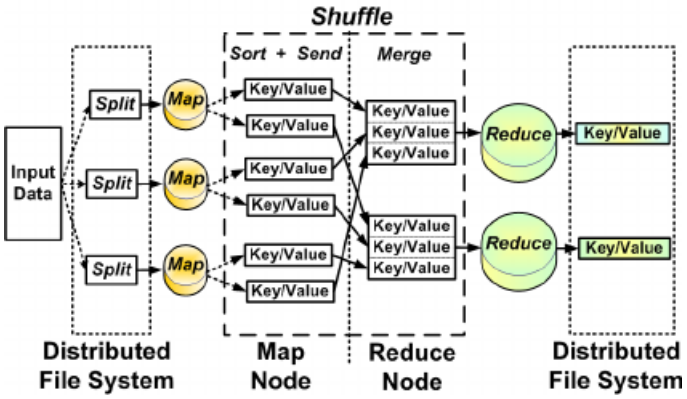

Figure 4: Workflow of MapReduce Algorithm.

K-means: K-means (MacQueen, 1967) is one of the least complex unsupervised learning calculations that tackle the notable clustering issue. The system pursues a basic and simple approach to group a given informational index through a specific number of clusters (accept k clusters) fixed from the earlier. The fundamental thought is to characterize k centroids, one for each group. These centroids ought to be put in a finesse route due to various area causes an alternate outcome. In this way, the better decision is to put them however much as could reasonably be expected far from one another. The subsequent stage is to take each direct having a place toward a given informational collection and partner it to the closest centroid. At the point when no point is pending, the initial step is finished and an early groupage is finished. Now, we have to re-ascertain k new centroids as barycenters of the clusters coming about because of the past advance. After we have these k new centroids, another coupling must be done between similar informational index focuses and the closest new centroid. A circle has been produced. Because of this circle, we may see that the k centroids change their area well ordered until no more changes are finished. As it were, centroids don't move any longer. At last, this calculation goes for limiting a goal work, for this situation, a squared mistake work. The target work is depicted as function below:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

II. LITERATURE SURVEY

Pratik Thakor, Dr.Sreela Sasi [1] Online networking gives a stage where clients share a plenitude of data on everything without exception. The data may comprise of clients' feelings, inputs, audits, and individual encounters. In this examination a novel Ontology-based Sentiment Analysis Process for Social Media content (OSAPS) with negative assumptions is displayed. The online networking content is consequently separated from the twitter messages. A metaphysics based procedure is intended to recover and break down the clients' tweet with negative assessments. This thought is shown with the recognizable proof of client disappointment of the conveyance administration issues of the United States Postal Service, Royal Mail of United Kingdom, and Canada post. The tweets identified with the conveyance administration incorporate deferral in conveyance, lost bundle/s or inappropriate client administrations at the workplace face to face or at call focuses. A mix of innovations for twitter extraction, information cleaning, abstract examination, metaphysics model structure, and assessment investigation are utilized. The outcomes from this investigation could be utilized by the organization to take remedial measures for the issues just as to create a mechanized online answer for the issues. A standard based classifier could be utilized for producing the robotized online answers.

Banu Yergesh,Gulmira Bekmanova, Altynbek Sharipbay,Manas Yergesh [2] Sentiment analysis one of the significant and fascinating undertakings with regards to common dialects and natural languages. Various assets and instruments have been created for slant examination of English, Turkish, Russian and different dialects. Sadly, there was no information and apparatuses accessible for notion examination in Kazakh. The Dictionary of Kazakh opinion words has been made amid the investigation. In this work, we depicted the standard based strategy utilizing a lexicon for feeling examination of writings in the Kazakh language, in light of the morphological principles and ontological model. Ontological model for principle extraction that decides notion was constructed. Our standard based technique accomplishes 83% precision for straightforward sentences. First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file "MSW_USltr_format".

Efstratios Kontopoulos,Christos Berberidis, Theologos Dergiades, Nick Bassiliades [3] The rise of Web 2.0 has definitely modified the manner in which clients see the Internet, by improving data sharing, joint effort, and interoperability. Miniaturized scale blogging is a standout amongst the most prevalent Web 2.0 applications and related

administrations, similar to Twitter, have developed into a viable method for imparting insights on practically all parts of regular day to day existence. Therefore, miniaturized scale blogging sites have since turned out to be rich information hotspots for assessment mining and estimation examination. Towards this heading, content based assumption classifiers regularly demonstrate wasteful, since tweets ordinarily don't comprise of agent and linguistically reliable words, because of the forced character limit. This paper proposes the organization of unique metaphysics based methods towards an increasingly proficient assessment investigation of Twitter posts. The curiosity of the proposed methodology is that posts are not just portrayed by a feeling score, similar to the case with AI based classifiers, yet rather, get a slant grade for each particular thought in the post. In general, our proposed engineering results in an increasingly point by point examination of post sentiments with respect to a particular theme.

Mustafa V. Nural, Michael E. Cotterell, Hao Peng, Rui Xie,Ping Ma, and John A. Miller [4] Prescient investigation in the enormous information time is taking on an ever progressively significant job. Issues identified with decision on demonstrating strategy, estimation methodology (or calculation) and proficient execution can exhibit critical difficulties. For instance, choice of suitable and ideal models for enormous information examination frequently requires cautious examination and extensive aptitude which may not generally be promptly accessible. In this paper, we propose to utilize semantic innovation to help information experts and information researchers in choosing fitting demonstrating procedures and building explicit models just as the reason for the strategies and models chose. To formally portray the displaying systems, models and results, we built up the Analytics Ontology that underpins inferencing for semi-robotized model choice. The SCALATION system, which at present backings more than thirty demonstrating procedures for prescient huge information investigation is utilized as a testbed for assessing the utilization of semantic innovation.

III. PROPOSED SCHEME

The proposed scheme will initiate the scenario from tweet extraction using flume which will act as an engine with connecter services using twitter API (Application Programming Interface). However, using the contextual query over the twitter the tweets will be extracted and will be preserved un-structurally in Hadoop Distributed File System, thereinafter using the Hive which is ORDBMS repository system under the umbrella of Hadoop Ecosystem for the schema model where the unstructured data will be transformed into structured and scheme related dataset on the next phase the data will be cleaned using stop word removal model based on the Porter Stemmer algorithm and will be forwarded for feature extraction using WordNet dictionary services subsequently the MapReduce will produce the count specific terms where the ontology will be formed using XML or OWL (Web Ontology Language). Consequently, using K-Means based on centroids the rational relation will be formed and results will be produced. The below diagram depicts the workflow of the proposed scheme for quick reference:

.

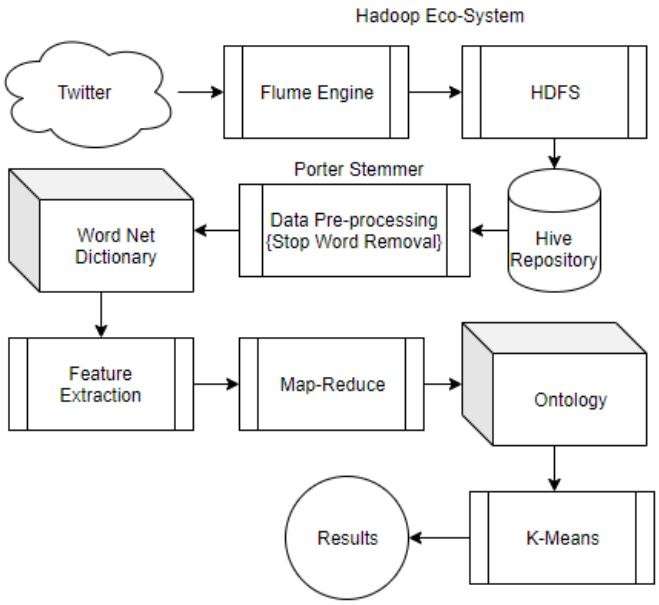

Figure 5: Proposed Workflow of Scheme comprising of Flume, HDFS, MapReduce, Ontology and K-Means

REFERENCES

[1] Pratik Thakor, Dr.Sreela Sasi Ontology-based Sentiment Analysis Process for Social Media Content, Procedia Computer Science, Volume 53, 2015, Pages 199-207.

[2] Banu Yergesh,Gulmira Bekmanova,Altynbek Sharipbay,Manas Yergesh, Ontology-Based Sentiment Analysis of Kazakh Sentences, International Conference on Computational Science and Its Applications ICCSA 2017: Computational Science and Its Applications – ICCSA 2017 pp 669-67

[3] Efstratios Kontopoulos,Christos Berberidis, Theologos Dergiades, Nick Bassiliades , Ontology-based sentiment analysis of twitter posts, Expert Systems with Applications Volume 40, Issue 10, August 2013, Pages 4065-4074, https://www.sciencedirect.com/science/article/pii/S0957417413000043?via%3Dihub#

[4] Mustafa V. Nural, Michael E. Cotterell, Hao Peng, Rui Xie,Ping Ma, and John A. Miller, Automated Predictive Big Data Analytics Using Ontology Based Semantics,Int J Big Data. Author manuscript; available in PMC 2018 Apr 13. Published in final edited form as: Int J Big Data. 2015 Oct; 2(2): 43–56.

[5] Ali Ghobadi, Maseud Rahgozar, "An Ontology-based Semantic Extraction Approach for B2C eCommerce", The International Arab Journal of Information T echnology, Vol. 8, No. 2, April 2011

[6] Zhongwu Zhai, Bing Liu, Hua Xu, Peifa Jia, "Clustering Product Features for Opinion Mining", Proceedings of ACM WSDM'II, February 9-12, 2011, Hong Kong, China.

[7] Erik Cambri, Robert Speer, Catherine Havasi, Amir Hussain, "SenticNet: A Publicly Available Semantic Resource for Opinion Mining", Commonsense Knowledge: Association for the Advancement of Artificial Intelligence, Fall Symposium, 2010.

[8] Ali Harb, Michel Plantie, Gerard Dray, Mathieu Roche, Franyois Trousset, Pascal Poncelet, "Web Opinion Minin g: How to extract opinions from blogs?", International Conference on Soft Computing as Transdisciplinary Science and Technology, 2008.

[9] Dongjoo Lee, Ok-Ran Jeong, Sang-goo Lee, "Opinion Mining of Customer Feedback Data on the Web".

[10] Shitanshu Verma, Pushpak Bhattacharyya, "Incorporating Semantic Knowledge for Sentiment Analysis", Proceedings of ICON-2008: 6th International Conference on Natural Language Processing, 2008.

[11] Manoj Manuja, Deepak Garg, "Semantic Web Mining of Unstructured Data: Challenges and Opportunities", International Journal of Engineering (UE), Volume 5, Issue 3, 2011.

[12] Larissa A. de Freitas, Renata Vieira, "Ontology-based Feature Level Opinion Mining for Portuguese Reviews", International Word Wide Web Conference Committee (IW3C2), 2013.

[13] Li Chen, Luole Qi, "Social opinion mining for supporting buyers' complex decision making: exploratory user study and algorithm

[14] Maciej Dabrowski, Thomas Acton, Przemyslaw Jarzebowski, Sean O'Riain, "Improving customer decisions using product reviews CROM", 2008.

[15] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.

[16] Xi-Quan Yang, Na Sun, Tie-Li Sun, Xue-Ya Cao, Xiao-Juan Zheng, "The application of Latent Semantic Indexing and Ontology in text classification", International Journal of Innovative Computing, Information and Control, Volume 5, No. 12(A), pp. 4491-4499, Dec 2009.

[17] Ayaz Ahmed Shariff, K, Mohammed Ali Hussai, Sambath Kumar, "Leveraging Unstructured Data into Intelligent Information-Analysis & Evaluation", International Conference on Information and Network Technology, Singapore, Vol.4, 2011.

[18] Harish Jadhao, Dr. Jagannath Aghav, Anil Vegiraju, "Semantic Tool for Analysing Unstructured Data", International Journal of Scientific & Engineering Research, Volume 3, Issue 8, August 2012.