ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

Ontology Based Rational Analysis in Automobile Sector via Tweeter using Map Reduce and K-Means

Shelly[1], Gaurav Garg[2]

P.G. Student, Department of CSE, AITM at Palwal, Haryana, India[1]

Assistant Professor, Department of CSE, AITM at Palwal, Haryana, India[2]

ABSTRACT: To propose a system or scheme for analyzing user behaviour by creating ontology on a particularly on automobile domain using tweets which are crowd-sourcing platform vide Machine Learning and Big Data Techniques. The ontology-based analysis helps in providing accurate and efficient results. It provides the overall score to the tweet in terms of opinion, domain and features rather than just Positive, Negative or Neutral. With the help of domain-specific ontology, we can find the hidden relationship and can have a comparison of which vehicle is best over other and what are the other attributes that other automobile company brand does not fall into this category thus resulting for better options.

KEYWORDS: Machine Learning, Twitter, Big Data, Ontology, Map-Reduce, K-Means.

I. INTRODUCTION

It is often the case that multiple applications need access to the same information. This type of information could be a material database, a specification of a part to be manufactured, or terrain characteristics of an environment that an autonomous vehicle must traverse. Instead of requiring an application to encode this information in its own internal format, ontology can provide this information in a neutral format that these different applications can reference. By not duplicating this information is numerous different applications, the ontology can allow the information to be represented only once, providing only a single source when information needs to be updated and ensuring that there is consistency between the information in different applications. In addition, the ontology provides a common set of vocabulary that all of the applications that access the ontology can reference. This will ensure that when information exchange between the applications needs to occur, mappings between concepts can be easily done based on the vocabulary in the ontology.

Scoring of the tweets can be done in two ways, either qualitative or quantitative. Qualitative is nothing but identifying whether positive, negative or neutral and quantitative is overall scoring of the tweets[2]. Ontology enables the addition of such terms to the knowledge base along with all the relevant features. This will speed up the process of retrieving relevant judgments based on the user's query. Ontology is defined as an explicit conceptualization of terms and their relationship to a domain [17]. It is now widely recognized that constructing a domain model or ontology is an important step in the development of knowledge based systems. Ontology in the field of computer science has been defined formally as "the specification of a conceptualization" by Tom Gruber. The concept of ontology was introduced by the Greek philosopher Aristotle. Wikipedia describes ontology as "the philosophical study of the nature of being, existence or reality in general, as well as of the basic categories of being and their relations". Ideally, each and every entity known to man would be represented by a URI (Uniform Resource Identifier) for unique identification. All known relationships to all other entities are recorded for each entity. This would form an all-encompassing ontology which is the ultimate fantasy of every computer scientist and also a very unrealistic idea. Hence, ontology are created such that

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

they are limited to contain entities only from a particular domain. Formation of these domain-specific ontology would normally require assistance of a domain expert. Examples of domain-specific ontology include WordNet which is a thesaurus in the form of ontology.

II. RELATED WORK

Ms. Swaminarayan Priya R. et al [1] have focussed on SPARQL i.e. a W3C standard, which is a built in querying tool available with Protégé. Though it supports for RDF and RDF graphs, it still works with OWL ontology. Their paper presented the purpose of using SPARQL on the OWL ontology to verify whether it is capable enough to query OWL ontology. Their results so obtained were a proof of their support.

Ms. Swaminarayan Priya R. et al [2] have developed a sample ontology which was tested by executing semantic queries against it. All the nodes as well as each instance metadata could be searched through the SPARQL query. According to them, the main advantages of using ontology are reusability and semantically searchable. Also, in the era of Semantic Web, the ontologies have become a powerful tool for knowledge sharing and it also supports the semantic interoperability among heterogeneous distributed systems. Ontologies and agent technologies are essential part of the semantic web, and their combined use will make possible the sharing of heterogeneous, autonomous knowledge sources in a capable, adaptable and extensible manner. In the multi-agent system, Ontology is used to assist the interactions among different agents and improve the quality of the service provided by each agent.

Pak A and Paroubek P [3] have focused on twitter which is considered as a corpus to collect the positive, negative and objective texts, the linguistic analysis is performed on the collected corpora and also a sentiment classification system is built where the features can be extracted based on the dataset and it is used to train the classifier whereas n-grams, bigrams and trigrams are examined as a feature and n-grams outperformed the other two. Naïve Bayes classifier is used as a classification system on N-gram and on part-of-speech, in future multilingual corpus data can be collected and the obtained data can be used for building multilingual sentiment classifier.

Mukherjeet S et al. [4] has proposed a multistage system used for analyzing the sentiment in tweet was proposed which uses 2 class and 4 class classification. Twisent addressed the problems like spam's, structural, Entity specification and Pragmatics. Twisent shows the accuracy improvement in 2-class as 14% and in 4-class setting 8.94% when compared with C-Fell-It, a rule based system similar to Twisent. Although the accuracy is achieved Twisent uses generic lexicon and so failed to capture the sarcasm or implicit sentiment. The focused on e a novel HL-SOT approach to labeling a product's attributes and their associated sentiments in product reviews by a Hierarchical Learning (HL) process with a defined Sentiment Ontology Tree (SOT). The empirical analysis against a human labeled data set demonstrates promising and reasonable performance of the proposed HL-SOT approach. The paper is mainly on sentiment analysis on reviews of one product, the proposed HLSOT approach is easily generalized to labeling a mix of reviews of more than one products.

R. Baracho, G. Silva, and L. Ferreira [5] has aimed to create a process of sentiment analysis based on ontologies in the automobile domain and then to develop a prototype. The process aims at making a social media analysis, identifying feelings and opinions about brands and vehicle parts. The method that guided the development process involves the construction of ontologies and a dictionary of terms that reflect the structure of the vocabulary domain. The proposed process is capable of generating information that answers questions such as: "In the opinion of the customer, which car is better: Corsa or Palio? Which one is more beautiful? Which engine is stronger?" To answer these questions by comparison, one can show a general view reflected on different social networks, indicating, for example, that for a given vehicle, a certain percentage of responses are considered positive, while for others, the percentage is considered negative

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

III. PROPOSED ALGORITHM

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: Therefore, the main idea of this approach is to classify sentences to a hierarchical ontology which captures the theme of the sentence and then do the comparative analysis of the automobile combining tweets and ontology. Our approach uses ontological knowledge to a greater extent since it is used to form a mapping between the text and the ontology. The proposed approach involves Ontology which is a formal specification of a shared conceptualization which helps in adding meaning and finds hidden relationship between the words and gives relevant and accurate summarization. Huge amount of work has been done in this area and many algorithms have been proposed. But there is still need of improvement to increase augmentation and efficiency of the sentiment analysis results obtained for the tweets. In this dissertation, a novel approach for analysis of tweets using ontology is presented.

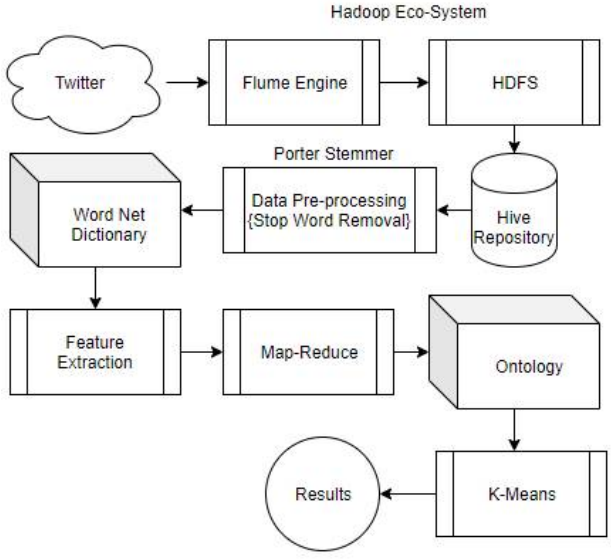

Figure1: Proposed Workflow of Scheme comprising of Flume, HDFS, MapReduce, Ontology and K-Means

IV. PSEUDO CODE

The Algorithm's pseudo-code:-

Step 1 – Map Reduce

```
map(String key, String value):
  // key: document name
  // value: document contents for each word w in value:
```

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

```
        EmitIntermediate(w, "1");

    reduce(String key, Iterator values):
      // key: a word
      // values: a list of counts int result = 0;
      for each v in values:
        result += ParseInt(v); Emit(AsString(result));
```

Step 2 : The instance layer of intermediate ontology}

= {OP1 , OP2, …., OPn*n} =

{(subject1,object1),…, (subjectn*n, objectn*n)}

CorrespondingConceptn } LD: Linked Data database Maxtime: maximum time preferred to search for RDF pages in Linked Data Database

Output: An Enriched Ontology Named O

Pseudo-Code:

1. for(int k=0;k<n*n; k++)

2. {

3. att=FindPredicate (LD, A[i][“subject”],

A[i][“Object”])

4. if (att != NULL)

5. add the

Assertional_knowledge”(A[i][“subject”],

att , A[i][“Object”])” to Ontology O

6. add the rule”(corresponding Concept Of(A[i][“subject”]), att , corresponding Concept Of (A[i][“Object”]))” to

Ontology O }

Step 3 : K-Means Algorithm

1 Generate K centroids C1, C2, ..,Ck by randomly choosing K Texts from D Repeat until there is no change in cluster between two consecutive iterations

```
2 for each Text di in D
        for j = 1 to K
      Sim(Cj, di) = Find cosine similarity between di and Cj
    end for
     Assign di to cluster j for which Sim(Cj, di) is maximum
    end for
3 Update centroid for each cluster
    end loop
4 end K-Means
```

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

V. SIMULATION RESULTS

In our approach, tweets are extracted from twitter and stored in a repository based on hadoop ecosystem. Then sentences are extracted one by one. Sentences extracted are simplified by removing stop words and redundant words. Stop words and redundant words are removed and keywords are extracted from each sentence. The words left in the sentences are used for sense matching using WordNet-an online semantic dictionary. WordNet dictionary is used to extract features from tweets. Ontology is being generated by using java customized code. Crawler is being designed next to get the details about the automobile domain. The data is stored in text manner. Mapping of data is done which includes mapping of ontology with the crawler data, together with ontology validation. Analysis of tweets is done using ontology by applying K-Means algorithm and comparison of automobile is done which one is better and what all are the attributes that other automobile does not fall into this category..

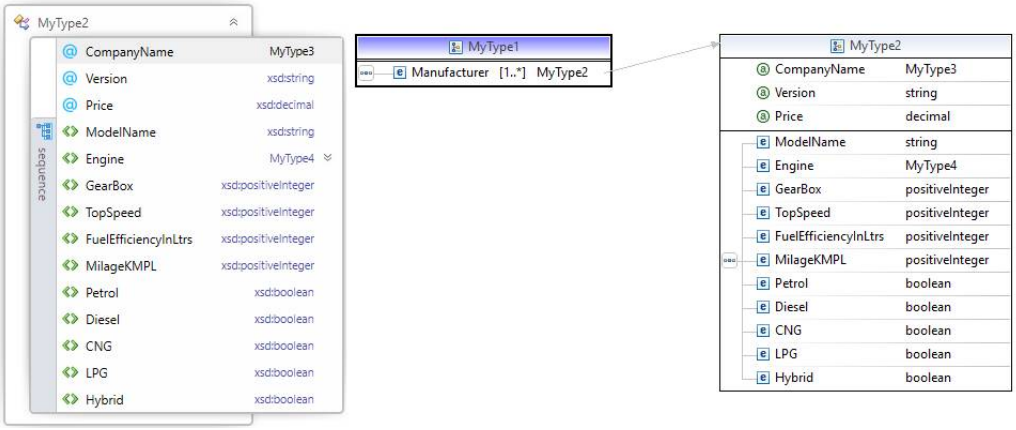

Figure 2: Ontology Description for Rational Analysis

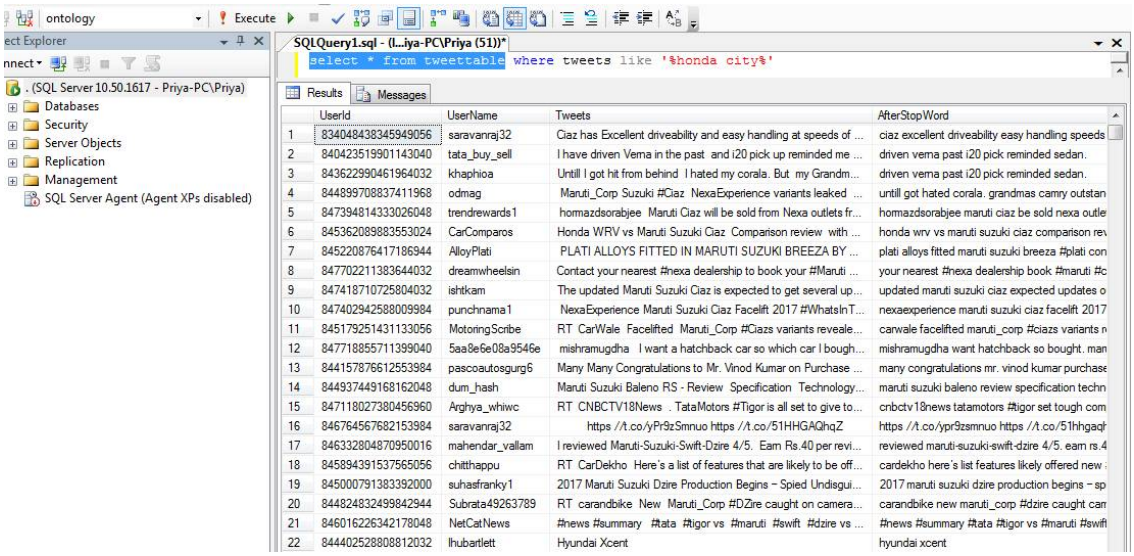

Figure 3: Tweets Fetched And Stored In Tweet Repository

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

Company Name	Version	Price	Model Name	Capacity	Horse Power	Torque	Gear Box	Top Speed	Fuel Efficiency in Lts.	Milage in KMPL	Petrol	Diesel
294	Maruti	Zxi	913000.0000	Ciaz	1373 cc	91 bhp @ 6000 RPM	130 Nm @ 4000 RPM	5	NA	43 Ltr	20.7 kmpl	yes

Error = java.lang.NullPointerException

User ID	User Name	User Tweets	Behavior Analysis	Behavior Analysis
1135515762418900992	MouthShut_com	maruti suzuki ciaz ambassador todays 5/5 #review @maruti_corp anunaylakra00789 best-car-in-this-budget https //t.co/mmfsmoa0ut	Pos :1 Neg :3 Neu :4	User Neutral
1135471836903292929	BBTheorist	maruti suzuki ciaz ambassador todays 5/5 #review @maruti_corp anunaylakra00789 best-car-in-this-budget https //t.co/mmfsmoa0ut	Pos :1 Neg :3 Neu :4	User Neutral

Figure 4. Rational Analysis using Ontology with Twitter Data

VI. CONCLUSION AND FUTURE WORK

In our proposed scheme, we have concluded that ontology used to analyses the tweets to increase augmentation and efficiency of techniques which is obtained using the K-Means algorithm is working properly. The ontology used is the specification of shared conceptualization and it helps in finding hidden relationships and adds meaning to our knowledge. The main components of ontology-based analysis system are tweeted extraction from Twitter4J/Flume, Tweet Repository, keyword extraction, sense matching, and ontology, Crawling automobile data, mapping ontology and crawler data and analysis. Our approach has some limitations are as follows:-

Negation Handling: Negation plays a vital role in altering the polarity of the associated adjective and hence the polarity of the text. Negation words in general, includes not, neither, nor. One possible solution to handle negations is to reverse the polarity of the adjective happening after a negating word, for instance, the restaurant is good (should be classified "positive") The restaurant is not good (should be classified "negative") However, this solution fails to entertain the cases like " No wonder the restaurant is good" and "Not only the food was yummy, the interior and service was also excellent". The use of pure language processing techniques or pure use of mathematical models fail to tackle negations completely.

Language Problem: People, organizations and software systems must communicate between and among themselves. However, due to different needs and background context, there can be widely varying viewpoints and assumptions regarding what is essentially the same subject matter Each uses different jargon. The consequent lack of a shared understanding leads to poor communication within and between these people and their organizations In the context of building an IT system, this lack of a shared understanding leads to difficulties in identifying requirements and thus in the defining of speciation of the system.

Hashtag Segmentation: For handling hashtags, we looked for the existence of the positive or negative words9 in the hashtag. But there can be some cases where it may not work correctly. For example, “#thisisnotgood”, in this hashtag if we consider the presence of positive and negative words, then this hashtag is tagged positive (“good”). We fail to capture the presence and effect of “not” which is making this hashtag as negative. We propose to devise and use some logic to segment the hashtags to get correct constituent words.

In the future, work can be carried forward in developing a fully automated ontology. Also, hashtag, navigation handling, etc can be worked upon.

In our proposed scheme, we have concluded that ontology used to analyses the tweets to increase augmentation and efficiency of techniques which is obtained using K-Means algorithm is working properly. Ontology used is specification of shared conceptualization and it helps in finding hidden relationships and adds meaning to our

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 7, July 2019

knowledge. The main components of ontology based analysis system are tweets extraction from Twitter4J/Flume, Tweet Repository, keyword extraction, sense matching and ontology, Crawling automobile data, mapping ontology and crawler data and analysis. Our approach has some limitations are as follows:-

Negation Handling: Negation plays a vital role in altering the polarity of the associated adjective and hence the polarity of the text. Negation words in general, includes not, neither, nor. One possible solution to handle negations is to reverse the polarity of the adjective happening after a negating word, for instance, the restaurant is good (should be classified "positive") The restaurant is not good (should be classified "negative") However, this solution fails to entertain the cases like " No wonder the restaurant is good" and "Not only the food was yummy, the interior and service was also excellent". The use of pure language processing techniques or pure use of mathematical models fail to tackle negations completely.

Language Problem: People, organizations and software systems must communicate between and among them selves. However, due to different needs and background context, there can be widely varying viewpoints and assumptions regarding what is essentially the same subject matter Each uses different jargon. The consequent lack of a shared understanding leads to poor communication within and between these people and their organisations In the context of building an IT system, this lack of a shared understanding leads to difficulties in identifying requirements and thus in the defining of a speciation of the system.

Hash tag Segmentation: For handling hashtags, we looked for the existence of the positive or negative words9 in the hashtag. But there can be some cases where it may not work correctly. For example, “#thisisnotgood”, in this hashtag if we consider the presence of positive and negative words, then this hashtag is tagged positive (“good”). We fail to capture the presence and effect of “not” which is making this hash- tag as negative. We propose to devise and use some logic to segment the hashtags to get correct constituent words.

In future, work can be carried forward in developing a fully automated ontology. Also, hash tag, navigation handling, etc can be worked upon.

REFERENCES

1. Ms. Swaminarayan Priya R. et al, “A Comprehensive study of Query Languages for Semantic Web and retrieval of data from University Ontology Using SPARQL” , International Journal of Information and Computing Technology” ISSN: 0976 – 5999, Volume 2, Issue 1, November 2011.
2. Ms. Swaminarayan Priya R. et al, “Knowledge Representation of Published Articles in Semantic Web using Upper Ontology”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 8, August 2012.
3. Pak A, Paroubek P., “Twitter as a corpus for sentiment analysis and opinion mining”, Proceedings of the 7th international conference language resources and evaluation (LREC ’10); 2010.
4. Mukherjeet S, Malu A, Balamurali AR, Bhattacharyya P, “TwiSent: A multistage system for analyzing sentiment in twitter”, 21st ACM Conference on Information and Knowledge Management; 2012.
5. Miller, G., “WordNet: An on-line lexical database”, International Journal of Lexicography, 3(4), 1991.
6. Zol S, Mulay P., “Analyzing Sentiments for Generating Opinions (ASGO) - A New Approach”, Indian Journal of Science and Technology, 8(S4):206–11, February 2015
7. Agarwal A, Xie B, Vovsha I, Rambow O, “Passonneau R. Sentiment analysis of twitter data” Proceedings of the ACL 2011 workshop on languages in social media, p. 30–8, 2011
8. Volk, M. “Using the Web as Corpus for Linguistic Research”. In: Catcher of the Meaning ,2002
9. C. Schlenoff, S. Balakirsky, T. Barbera, C. Scrapper, J. Ajot, E. Hui, and M. Paredes, “The NIST Road Network Database: Version 1.0”, National Institute of Standards and Technology (NIST) Internal Report 7136,2004.
10. C. Schlenoff, M. Gruninger, F. Tissot, J. Valois, J. Lubell, and J. Lee, “The Process Specification Language (PSL) Overview and Version 1.0 Specification”, in NISTIR 6459, National Institute of Standards and Technology 2000.
11. C. Schlenoff, S. Balakirsky, M. Uschold, R. Provine, and S. Smith, “Using Ontologies to Aid in Navigation Planning in Autonomous Vehicles”, Knowledge Engineering Review, vol. 18, no. 3, pp. 243-255, 2004.
12. Amrita Kaur and Neelam Duhan, “A Survey on Sentiment Analysis and Opinion Mining”, International Journal of Innovations & Advancement in Computer Science, ISSN 2347 – 8616, Volume 4, Special Issue May 2015