International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Classifying Mined Online Discussion Data for Reflective Thinking Based on Ontology

Leji Elza Abraham[1], Merin Mary Philip[2]

[1]*M.Tech Student, Department of CSE & Belivers Chruch Caarmel Engineering College, Kerala- 689711, India.*
[2]*Asst. Professor, Department of CSE & Belivers Chruch Caarmel Engineering College, Kerla-689711, India.*

Abstract – *The discussion of online text data giving a ray on reflective thinking's. The manual coding will be challenged because the growth of text data. To process the unstructured text data, it is essential to integrate the inductive content analysis, and data mining technique. An inductive content analysis was implemented on the online discussed data and categories the reflective thinking's. Based on result of inductive content analysis, a single-label text classification algorithm are implemented to classify the sample data. Then applied the trained classification models on the large scale and undetected online discussion text data. An ontology which is applied to the text data to giving explicit specification of conceptualization.*

***Key Words*: Online text data, Reflective thinking's, Inductive content analysis, Single-Label Naïve Bayes Classifier, Ontology**

1. INTRODUCTION

Data mining is the exercise of finding large data sets patterns involving methods at the intersection of machine learning, statistic and a database system. Data mining is a collaborative sub-field of statistic and computer science and the goal of the data mining is to extract information from a data set and remodel the information into the all-inclusive or detailed structure for further use. It having six major tasks and they are anomaly detection, association rule learning, clustering, regression, classification and, summarization.

Online discussion data is the data which is shares in the online for various activities or purpose. Day by day, most of the participants of the online communities where share their experiences, opinions, advice and social support. The shared data is without any such constraints. The online discussion is intermixed with a clutter of disagreement, flame wars and trolling and so on. The online discussion data is shared with by different peoples from different countries and each of them have their own views about a topic. These different views on a particular topic that is shared each other because of them in a particular topic for a person will increase and it helps to change their attitude towards the topic. Most of the online discussion data being available as in the format of portable document format. The portable document format which is used to storing a document on the computer is in file format.

Ontology is a systematic account of existences and it is borrowed from philosophy. It's an explicit specification of conceptualization. Based on computer science it is a formal representation of the knowledge by a set of concepts within a domain and the relationship between the concepts. It is mainly used in artificial intelligences, semantics web, system engineering, software engineering, biomedical informatics, library science and so on. It is widely used for the representation of knowledge.

For manually analyze the text content, inductive content analysis was used. The purpose of the research is to find reflective thinking's in mined data and a detailed classification of mined data done by the basics of ontology.

Qingtang Liu and Si Zhang [1] was find the reflective thinking's of teachers'. To process the large-scale unstructured text data, it is needed to integrate the inductive content analysis and, educational data mining technique. Based on result of inductive content analysis, a single-label text classification algorithm are implemented to classify the sample data. Then applied the trained classification models on the large scale and undetected online discussion text data. Here the ontology concept was not applied and also the visualization of the reflective thinking's is quiet difficult.

Atiya Kazi and Prof. D. T. Kurian [5] found the problem of assessing a given ontology for a particular criterion of an application and also determine which ontology is suite more for current application domain. But by using different ontologies its makes confusion.

2. METHODOLOGY

2.1 Research Design

In order to understand reflective thinking's and detailed knowledge about a topic from the online discussed mined data, this study went through six main phases (in Fig-1).

Phase 1: Online discussed data is retrieved by using a tool called 'Unpdf' and place the data into the data storage.

Phase 2: From the data storage any one of the random sample is selected for pre-processing.

Phase 3: In pre-processing, the removal of stop words, special symbols, and, punctuations.

Phase 4: Classifying the data based on the single-label Naïve Bayes text classifier.

Phase 5: It is applied to the large scale data and places on the reflective thinking's categories.

Phase 6: Ontology is applied to the text data to getting deeper classification about a topic.

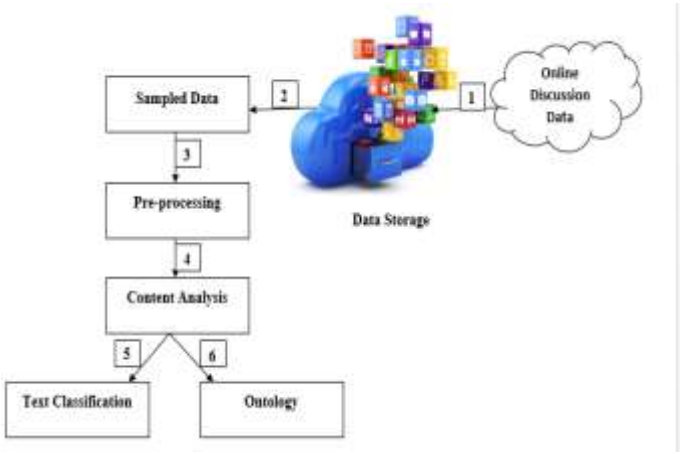

Fig -1: Research Design

2.1 Data Collection and Pre-processing

For collecting the online discussed data, a tool is used as 'Unpdf'. By using the tool, can fetch a large amount of online discussed data and be stored in the data storage one of the collected data used for data pre-processing. The online discussed data which having some special symbol to squeeze out certain meaning. For example, representing reflection in themselves, emoticons are used. After pre-processing of data, training of the classifier will takes place. On pre-processing

1) Some special symbols, punctuation's are removed.

2) Common stop words also removed such as 'the', 'a', 'in' and so on.

2.2 Automatic Coding Scheme

There are large number of coding scheme which is used to classified mined online discussed data. Van Manen says that there are three levels of reflective. The first level of reflection is concern with technical rationality and the second level of reflection is concern with technical analysis. The highest level of reflection is concerned with the technical critique. But the Vali summarized into six different types of reflection. They are "Technical-Description, Technical-Analysis, Technical- Critique, Personalistic-Description, Personalistics- Analysis and, Personalistic-Critique".

Technical-Description is concern with instructional, managerial, or contextual aspects of a topic by providing descriptive information of an action. Technical-Analysis is concern with the instructional, managerial, or contextual aspects of a topic by providing rationale and logic of an action. Technical-Critique is concern with instructional, managerial, or contextual aspects of a topic by providing explanations and evaluation of an action.

Personalistic-Description is concern with beliefs or professional development by providing descriptive information of an action. Personalistic-Analysis is concern with beliefs or professional development by providing rationale and logic of an action. Personalistic-Critique is concern with beliefs or professional development by providing explanation and evaluation of an action.

As the conclusion that each post or the data only be labeled with one category, so this was a single-label classification.

2.3 Single-Label Naïve Bayes Classifier

Single-Label Naïve Bayes classifier is a classifier which having strong independence assumptions between the features with Bayes' theorem. Naïve Bayes' classifiers, is a family of classifier that are based on the popular Bayes probability theorem. The Bayes' theorem says that how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from the conditional probability of the outcome of each cause and knowledge of the probability of each causes. The Bayes' theorem is as follows:

Posterior probability= (conditional probability . prior probability)/evidence

For example,

$$p(A|B) = (p(B|A)\, p(A))/(p(B)) \quad (1)$$

where, p(A|B) is posterior probability of A given the evidence.
p (B|A) is likelihood of evidence B if the hypothesis A is true.
p (A) is prior probability.
p (B) is prior probability that the evidence itself is true.

Naïve Bayes' classifiers, is a family of classifiers that are based on the popular Bayes probability theorem. Naïve Bayes' classifier is a simple classifier but efficient. In the text classification based on Single-Label Naïve Bayes classifier "tokenization, stop words, stemming and lemmatization" are performed. In tokenization the general process of breaking down a text focus into individual elements that serve as input for various natural language process algorithm. The stop words are words that are particularly common in a text focus and thus considered as rather un-informative. For example so, or, the, and, thus and so on. The stemming is the exercise of transforming a word into its root from aim of the stemming into obtain canonical forms of word.

There are many classifier which is used for classifying the data such as Support Vector Machine (SVM), Decision Tree (DT) and so on. But the Single-Label Naïve Bayes classifier to classify posts based on the categories developed on the inductive content analysis process. The

basic procedure of the single-label Naive Bayes classifier was described as below:

Suppose there are X total number of labeled documents in a training document set and Y of them were in the category c. The category can be set as six, C= {c_1, c_2, c_3, c_4, c_5, c_6}. The prior probability of category c is

$$p(c) = \frac{X}{Y} \quad (2)$$

Suppose there are N total number of words in the training document set, W={w_1,w_2,…, w_n}. If a word w_n present in one category c for $s_{wn,c}$ times, then based on the maximum likelihood estimation, the prior probability of word w_n in category c is

$$p(w_n|c) = \frac{Swn,c}{\sum_{n=1}^{N} Swn,c} \quad (3)$$

In the testing document set, there are K words in a document d_j that is Wd_j ={w_{j1}, w_{j2},…,w_{jk}} and it is a subset of W. Based on the Bayes' theorem, the posterior probability that a document d_j belongs to a category c is

$$p(c|d_j) = \frac{P(dj|c).\ P(c)}{p(dj)} \quad (4)$$

Repeated this procedure for each category. Then, assigned one category with the largest probability to document dj. If all the posterior probabilities that document dj belonged to each category is 0, then document dj is assigned to "others".

The advantage by using the Single-Label Naïve Bayes classifier is efficient, it need only some number of training data to estimate the parameter necessary for classification. It is relatively robust, easy to implement, fast and accurate. One of the drawback by using Single-Label Naïve Bayes classifier is nonlinear classification and the strong violations of the independence assumptions problem can leads to poor performances.

2.4 Applying Ontology

Ontology is a systematic account of existences and it is borrowed from philosophy. It's an explicit specification of conceptualization. It is mainly used in artificial intelligences, semantics web, system engineering, software engineering, biomedical informatics, library science and so on. It is widely used for representation of the knowledge.

Ontology is a concrete form of a conceptualization of community's knowledge in a specific domain. For modelling the knowledge, some requirements should be taken into consideration such as

1) Heterogeneity: the ontologies must handle various data sources.
2) The size: the ontology is overlapped rapidly due to different evolutions.
3) Reasoning: it leads to inferring some new knowledge and to promote the personalization process.
4) Integration of data and reutilization.

Mainly, there are four classification of ontology i,e statics, dynamic, intentional and social. Static ontology describes things that exist, their attributes and relationship. Dynamic ontology explains the world in terms of states, state transitions and processes. Intentional ontology represents the world of agents, things believe in, want, prove or disprove and argue about. Social ontology describes social settings, permanent organizational structures or shifting networks of alliances and independencies.

Here to the mined data implement a large-scale text classification based on the trained classification model. Where the ontology is applied to the text classified data to get more knowledge about the mined data. The advantages of using ontology is having essential relationship between concepts built into them, they enable automated reasoning about data. Next, is the ontology work and reason with the concepts and relationships in ways that are close to the way humans perceive interlinked concepts. It also provide a more coherent and easy navigation as users move from one concept to another in the ontology structure.

3. CLASSIFICATION RESULTS

When the Single-Label Naïve Bayes classifier was applied for the first time on the dataset, the performance of the classifier was very uneven. There are some methods to improve the performances of the classifier. First, the most popular term weighting approach, the Term frequency and inverse document frequency (TF-IDE) which is used to give appropriate weight to the terms which increase the performances. Another method is the smooth method which is also used to improve the classifier performances. Most of the probable words comes in each category are given below table.

Category	Most Probable Words
Technical-Description	compare, part, operation, layer, present, stimulate,
Technical-Analysis	harmful, against, promote, traditional, abstract,
Technical-Critique	situation, task, assignment, in good time, exploration
Personalistic -Description	think, consider, understand, reflection, process
Personalistic -Analysis	help, must, essential, necessity, constant,
Personalistic-Critique	collaborative learning, discuss, seminar, workshop, positive factors,

Table -1: Most probable words comes in each category.

By applying ontology we can obtain deeper classification about a topic. Some of them are given below:

Topics	Classification
Intellectual Capital	Structural Capital, Human Capital, Relational Capital
Knowledge Management	Content Management, Collaboration, Search
Business Intelligence	Big Data, Data Mining, Text Mining, Data Warehousing, Predictive analysis
Validation & Verification	Deployment, Design Tool, Testing, Rule Management

Table -2: Deeper classification on a topic.

3. CONCLUSION

To classify the large scale unstructured text data it is needed to integrate both the qualitative content analysis method such as inductive content analysis and data mining techniques. The inductive content analysis revealed the reflective thinking included Technical-Description, Technical-Analysis, Technical-Critique, Personalistic-Description, Personalistic-Analysis, Personalistic-Critique. As the result of the inductive content analysis, a single-label text classification algorithm are implemented to classify the sample data and the trained classification model is applied on a large scale and unexplored online discussion text data set. The ontology is a systematic accounts of existence and it is borrowed from philosophy. It's an explicit specification of conceptualization. The ontology are applied to the text classified data to get more knowledge about the mined online discussion data.

Classifying mined online discussion data for reflective thinking based on ontology provided a method for analyzing large-scale unstructured text data and could overcome the limitations of manual content analysis and pure machine learning method. It also help to giving a new way view to the people by giving more information about each of the topic which is present in the mined data.

REFERENCES

[1] Qingtang Liu and Si Zhang, "Mining Online Discussion Data for Understanding Teachers' Reflective Thinking," IEEE Transaction on Learning Technologies, Volume 11, June 2018.

[2] Mohammad Khanbabaei, Reza Radfar "Developing an integrated framework for using data mining techniques and ontology concepts for process improvement", *Universal Access Inf. Soc*., Volume. 16, pp. 1–11, 2017

[3] F. G. K. Yilmaz, and H. Keser, "The Impact of Reflective Thinking Activities in e-learning: A Critical Review of the empirical Resaerch,"Comput. Edu., Volume 95, pp. 163-173, 2016.

[4] M. Liu, R. A. Calvo, A. Pardo, and A. Martin, "Measuring and visualizing students' behavioural engagement in writing activities," IEEE Transaction on Learning Technologies, Volume 8, Issue 2, pp. 215-224, June 2015.

[5] Atiya Kazi and Prof. D. T. Kurian, "An Ontology Based Approach to Data Mining," Internal National Journal for Engineering and Research, Volume 2, Issue 4, 2014.

[6] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences," IEEE Transaction on Learning Technologies, Volume 7, Issuse 3, pp. 246-259, September 2014.

[7] Diana Man "Ontology in Computer Science," Didactica Mathematica, Volume 31, Issuse 2, pp. 43-46, September 2013.

[8] C. D. Epp and S. Bull, "Uncertainity Representation in Visualizations of Learning Analytics for Learners: Current Approaches and Oppurnities," IEEE Transaction on Learning Technologies, Volume 8, Issue 3, pp. 242-260, 2013.