USER INTERESTED DATA CLASSIFICATION USING ONTOLOGY & STATISTICAL METHOD FROM RSS FEEDS

Vrushali M. Parmar[1], Ishan K. Rajani[2], Nilesh D. Solanki[3]

[1] Student, Department of Information & Technology, SOCET, Ahmedabad, INDIA
[2]A/Prof., Department of Computer Engineering, SOCET, Ahmedabad, INDIA
[3] A/Prof., Department of Information & Technology, SOCET, Ahmedabad, INDIA

ABSTRACT

The Classification techniques are extensively used in data mining to classify data among various classes. The Classification techniques are being used in various industries to simply identify the type and group to which a particular data belongs. This paper proposed approach have need of Domain ontology based classification of RSS Feeds into a number of dynamic define news categories and to use naïve bayes algorithm, to find the probability. In our experiments, we used the English language DBPEDIA converted to an RDF ontology to categorize a corpus of current RSS news into selection of topics of interest and to classified news category and used the IPTC news industry and DBPEDIA Ontology.

Keyword : - *Ontology, Naïve Bayes Algorithm, Text Classification*

1. INTRODUCATION

Text categorization is a task of assigning one or more predefined categories to the analyzed document, based on its content. In text classification the People categorize text documents based on their general knowledge and their interest that determines which facts are treated as more significant [4]. Although reading a news document we can detain most important actors, facts and places, they connecting them into a one consistent event. Ontology is an explicit specification of a conceptualization [5].

The term is rented from philosophy where Ontology is a systematic account of Existence of the Theory. In artificial intelligence systems, what "exists" is that which can be represented to the ontology. The Ontology is a Conceptualization is a theoretical and the simplify view of the world that desire to keep up a correspondence to for some specific purpose. The Ontology is the knowledge based system, or knowledge level agent is faithful to some conceptualization and explicitly or implicitly [1]. The Classification is a supervised learning technique and first the classifier is trained with a number of examples for the documents whose category is already known and then the classification algorithm is applied on other documents whose category is unknown to find category.

There is a large range of classifiers in including k Nearest Neighbor, Decision Tree and Neural Networks, and Bayesian classifiers follow probabilistic approaches that make physically powerful assumptions on how data are generated, and propose a model that embodies those assumptions [3]. In Classification technique of new documents is performed in accordance with the Bayes rule and selecting the most likely category for each document. Naïve Bayes algorithm or classifier is one of the simplest, efficient, and effective algorithms for being used with data mining and machine learning purposes. Naïve bayes is grand performance is surprising and due to its assumption of conditional independence which rarely happens in real world applications given a set of documents to classify, and being those documents represented as a series of features, Naïve Bayes classifier assumes "ingenuously" that those features are self-governing of one another[28]. Although being so simple, and the Naïve Bayes classifier performs surprisingly well compared to other classifiers that are much more sophisticated, and is to be assumption of self-determination.

The Rest of this Paper is structured in the following way: We divide our research paper into five sections. First section delivers about our background and objective in this research. Our related works can be seen in second section. For third section, we define about our proposed method. Fourth section is used to define our experiment dataset and experiment scenario. Last section is used for providing our research conclusion and our further research.

2. RELATED WORK

Text classification is used in many application such as sentiment analysis and emotion identification from twitter , spam filtering in and Other researches in text classification for Indonesian Language are developed in and Rachmania et.al. in provides classification for Indonesian news document and Laksana et.al. in provides classification for Indonesian tweets. Some research that uses Naive Bayes as classifier can be found [2]. By looking at this, we also try to use Naive Bayes as classifier algorithm in our research.

In this Research algorithm because Naive bayes is widely used in text classification and have some advantages, that are simple to implement and have a good performance in text classification. To Research analysis that automatic text classification can be expert by relying on the semantic similarity between the information included in a document and a suitable fragment of the ontology [4].

In this argument is based on the assumption that entities in the works in the document text along with relationships among them can determine the document's categorization, and that the entities classified into the same or similar domains in the ontology are semantically closely related to each other. In order to be able to achieve meaningful results, that require that the ontology (i) cover the categorization domain for particular area, (ii) include a rich instance base of named entities and meaningful relationships in the middle of them, (iii) proper labels for named entities that enable their identification in categorized documents, and (iv) have the entities classified according to a class taxonomy included in the ontology [4].

Naive Bayesian classifiers are efficient in terms of time, and CPU usage and memory. They can carry out well even with small training sets and are computationally less extensive. They take less training time than other methods [9].

There are several works done in the recent past on text classification. Li et al. is proposed to text classification technique using positive and unlabeled data[13]. For, This list is helpful when the probabilistic classifier calculates the probability of each word being annotated to each category[8]. The list of words is then used to generate a table, containing the words which are extracted from the input document. The probabilistic classifier is needed to be trained with a set of well categorized training dataset [8].

3. PROPOSED METHOD

In research consider those concepts which are part of news Ontology. To design the Naive Bayes algorithm in classification and use the International press telecommunication Council (IPTC) Standard of news Industry. In news there are so many category and user they not interested in news of all category. Auto filtering based on user requirement category is required, we are using RSS feeds of various English News paper website.

We are classifying this RSS using IPTC and DBPEDIA Ontology. We are using Naïve Bayes algorithm for this purpose. We present our proposed method in this section. Our proposed method is divided into two parts. First part is Architecture of proposed work. Second part presents about RSS classification that we us in this research. Architecture system for our work can be looked at figure 1.

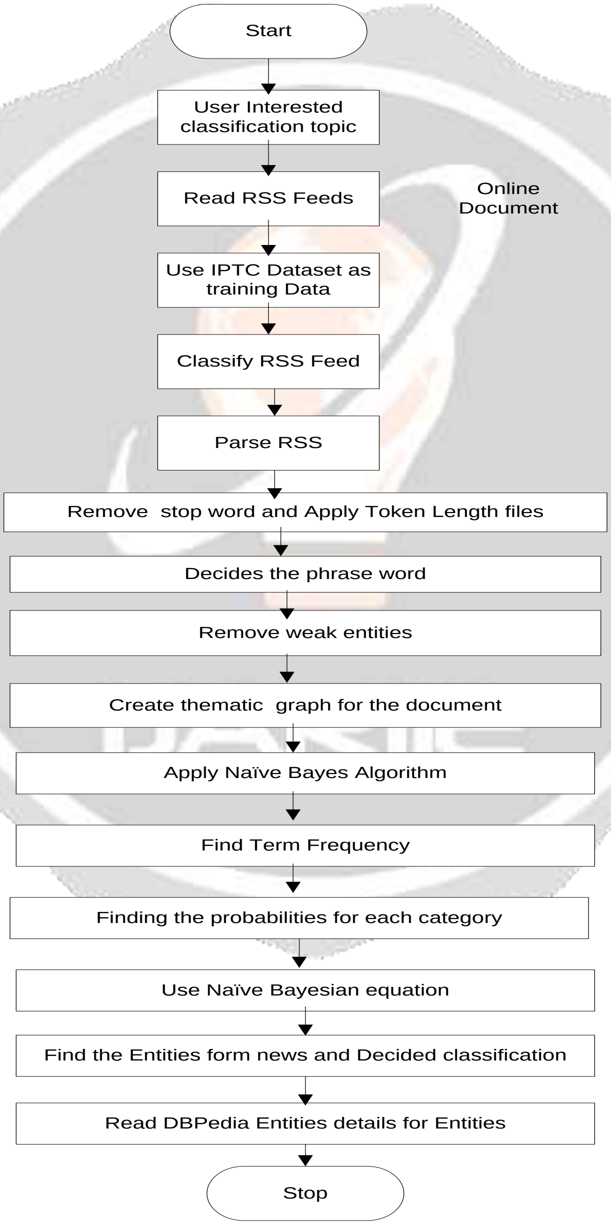
Start
User Interested classification topic
Read RSS Feeds
Online Document
Use IPTC Dataset as training Data
Classify RSS Feed
Parse RSS
Remove stop word and Apply Token Length files
Decides the phrase word
Remove weak entities
Create thematic graph for the document
Apply Naïve Bayes Algorithm
Find Term Frequency
Finding the probabilities for each category
Use Naïve Bayesian equation
Find the Entities form news and Decided classification
Read DBPedia Entities details for Entities
Stop

Fig 1: Architecture of Proposed Work

A. Naive Bayes Text Classification

Naïve Bayesian classifiers assume that there are no dependencies amongst attributes. This assumption is called class conditional independence. Naïve bayes is made to simplify the computations involved and, hence is called "naive". This classifier is also called idiot Bayes, simple Bayes, or independent Bayes.
The advantages of Naïve Bayes are:

- It uses a very intuitive technique. Bayes classifiers is unlike neural networks, and naïve bayes do not have several free parameters that must be set. This really simplifies the design process.
- While the naïve bayes classifier returns probabilities, it is simpler to apply these results to a wide variety of tasks than if an subjective scale as used.
- Naïve Bayes does not require large amounts of data before learning can begin.
- Naive Bayes classifiers are computationally fast when making decisions.

Naive Bayes algorithm is broadly use as text classification algorithm. This classifier uses bayes theorem with some naive assumption about class limited independence for each feature. Naive Bayes classifier uses probability to find output target class with maximum posterior probability. We can define naive bayes classifier model in 1. We can compute P(*ci*|D) using Bayes Theorem in 2 as follows.

$$\mathcal{Y} = \underset{ci \in C}{argmax}\, P(c|i|D)$$

Where:

- D : document
- *ci* : target class *ci* in C
- C : all target class in dataset
- P(*ci*|D) : posterior probability class *ci* if given document D

$$P(c|D) = \frac{P(D|c)P(c)}{P(D)}$$

- Usually some research in naive bayes classifier is dropping denominator so this equation can define.

$$P(c|D) = P(D|c)\, P(c)$$

Where:

- D : document
- c : target class
- P(D|c) : posterior probability document D if given target class c
- P(c) : prior probability for target class c in dataset And for computing prior probability P(c),

$$P(c) = \frac{Nc}{N}$$

Where:

- *Nc* : total target class c in dataset
- *N* : total instance in dataset

$$P(D|c) = \prod_{Wi \in D} P(wi|c)$$

Where:

- D : document D
- c : target class
- *11wi* : word in document D
- P(*wi*|c) : posterior probability word *wi* in D if given class c

$$P(wi|c) = \frac{Count(wi, c) + 1}{Count(c) + |V|}$$

Where:

- |V |: Vocabulary size
- count(*wi*,c) : total term *wi* in class c
- count(c) : total term in class c

B. International Press Telecommunication Council

The International Press Telecommunications Council (IPTC) it is based on London, United Kingdom, is a consortium of the world's major news agencies and the other news providers and news industry vendors and acts as the universal standards body of the news media. IPTC aims at simplifying the allotment of information.

IPTC news to achieve these technical standards is developed to improve the management and exchange of information between content providers and the intermediaries and consumers. IPTC is committed to open standards and makes all standards without restraint available to its members and the wider community. This news agency is a consortium of the world's major news agencies, news publishers and news industry vendors. This taxonomy contains three layers as well as given, namely the Subject, the Subject Matter and the Subject Detail. This taxonomy automatically classiness' the articles based on their content. However, it is not very septic concerning economical news. The company that is using this taxonomy is Your News, which is a part of MD Info.

4. EXPERIMENTS

In our experiments, we used select link of RSS news feed from the English news website. News Items in selected feed are classified into different categories. Results show the classified news items based on the category of users choice. Our dataset are IPTC and used DBPEDIA in this experiment. In this used ENGLISH News website and classified the category. This proposed method can provide a robust system for text classification for specifying domain specified and given the higher accuracy.

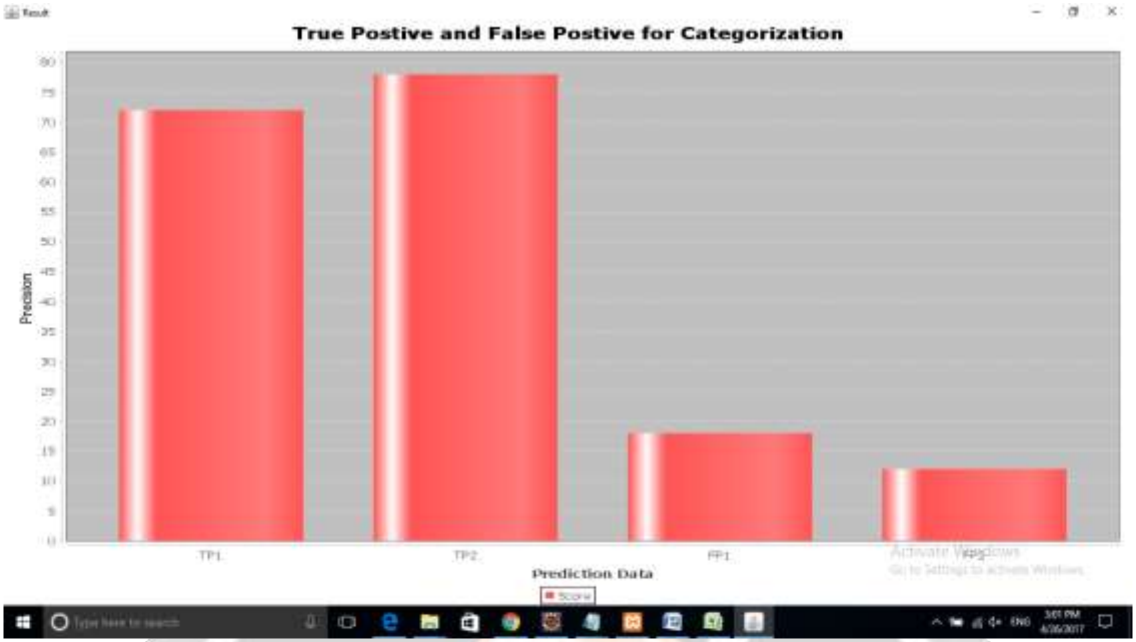

Fig-2 Result Analysis

Figure shows the precision and recall measured using these two approaches. Precision=True Positive/(True Positive + False positive) and Recall= True Positive/(True Positive + False Negative)

5. CONCLUSIONS

After implementation we presented a novel approach to text categorization, relying only on the ontological knowledge. Categories of interest can be defined as context projections or their combinations. Our experiments proved the applicability of ontologies for automatic text categorization and demonstrated a significant value of knowledge represented in DBPEDIA and IPTC when applied to this problem and used the naïve bayes algorithm.

In Future, different languages documents or the news data can be classified successfully in multiple classes with the help of ontology. Machine learning can be used for self learning of ontology. Self learning ontology must be implemented for better accuracy.

6. REFERENCES

[1] http://queksiewkhoon.tripod.com. (n.d.). Retrieved 4 18, 2017, from queksiewkhoon.tripod.com: http://queksiewkhoon.tripod.com/ontology_01.pdf

[2] Joan Santoso∗†, E. M. (2015). Large Scale Text Classification using Map Reduce and Naive Bayes Algorithm for Domain Specified Ontology Building. IEEE , 428-432.

[3] Marcos Mouriño-García, R. P.-R.-R.-C. (2016). Bag-of-Concepts Document Representation for Bayesian Text Classification. IEEE , 281-288.

[4] Mehdi Allahyari1, Krys J. Kochut, and Maciej Janik,” Ontology-based text Classification into Dynamically Defined Topics”, International Conference on Semantic Computing, IEEE 2014,978-1-4799-4003-5/14,Pages:-273-278

[5] Maha Maalej and Achraf Mtibaa,” Enriching User Model Ontology for Handicraft domain by FOAF”,IEEE 2015, 978-1-4799-8679-8/15.

[6] Marcos Mouriño-García, R. P.-R.-R.-C. (2016). Bag-of-Concepts Document Representation for Bayesian Text Classification. IEEE , 282-288.

[7] Shikha Agarwal, A. S. (2012). Classification of RSS Feed News Items using Ontology. IEEE , 251-255.

[8] Shweta Joshi, B. N. (2011). Categorizing the Document using Multi Class Classification in Data Mining. IEEE , 251-255.

[9] Amna Rahman, U. Q. (2016). A Bayesian Classifiers based Combination Model forAutomatic Text Classification. IEEE , 63-67.

[10] Prof.Sumitra Pundlik, Prachi Kasbekar, Gajanan Gaikwad, Prasad Dasare Akshay Gawade and Purushottam Pundlik,” Multiclass Classification and Class based Sentiment Analysis for Hindi Language” Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India, 978-1-5090-2029-4/16,Pages:-512-518.

[11] Prof.Sumitra Pundlik, Prachi Kasbekar, Gajanan Gaikwad, Prasad Dasare Akshay Gawade and Purushottam Pundlik,” Multiclass Classification and Class based Sentiment Analysis for Hindi Language” Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India, 978-1-5090-2029-4/16,Pages:-512-518.

[12] Guang Yang, Jin-Kun Tian, Yun-Hua Liu, Zhong-Yi Lin, Lei Wang,and Yu-Xin Chang,” Chinese Text Classification Based On Improved Domain Ontology Graph Model- DOG”, 2nd International Conference on Computer Science and Network Technology, IEEE 2012, 978-1-4673-2964-4/12/,Pages:-1331-1335.

[13] Sadia Zaman Mishu, S. M. (2016). Performance Analysis of Supervised Machine Learning Algorithms for Text Classification. IEEE , 409-413.

[14] Domain Ontologies in Software Engineering:. (n.d.). Retrieved November 12, 2016, from www.ics.uci.edu: http://www.ics.uci.edu/~andre/informatics223s2009/musen.pdf

[15] Hoehndorf, R. (n.d.). Ontogenesis. Retrieved November 13, 2016, from http://ontogenesis.knowledgeblog.org: http://ontogenesis.knowledgeblog.org/740

[16] janez.brank@ijs.si, m. d. (n.d.). A SURVEY OF ONTOLOGY EVALUATION TECHNIQUES. Retrieved November 12, 2016, from http://ai.ia.agh.edu.pl/: http://ai.ia.agh.edu.pl/wiki/_media/pl:miw:2009:brankevaluationsikdd2005.pdf

[17] Monika Yadav, M. P. (n.d.). International Journal of Advanced research in Computer Science and Software Engineering. Retrieved November 10, 2016, from www.ijarcsse.com: https://www.ijarcsse.com/docs/papers/Volume_3/3_March2013/V3I3-0335.pdf

[18] ontology. (n.d.). Retrieved November 12, 2016, from http://whatis.techtarget.com/: http://whatis.techtarget.com/definition/ontology

[19] Process ontology. (n.d.). Retrieved November 12, 2016, from en.wikipedia.org: https://en.wikipedia.org/wiki/Process_ontology

[20] Specification of Conceptualization. (n.d.). Retrieved November 11, 2016, from http://www.obitko.com/: http://www.obitko.com/tutorials/ontologies-semantic-web/specification-of-conceptualization.html

[21] V. Maniraj, D. S. (n.d.). Ontology Languages – A Review. Retrieved November 15, 2016, from http://www.ijcte.org/: http://www.ijcte.org/papers/257-G750.pdf

[22] Lecture #4: HITS Algorithm - Hubs and Authorities on the Internet. (n.d.). Retrieved November 15, 2016, from http://www.math.cornell.edu/: http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

[23] (n.d.). Retrieved 3 11, 2017, from wikipedia.org: https://en.wikipedia.org/wiki/Ontology_(information_science)

[24] (n.d.). Retrieved 3 11, 2017, from wikipedia.org: https://en.wikipedia.org/wiki/RSS

[25] (n.d.). Retrieved 3 12, 2017, from wikipedia.org: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf

[26] https://en.wikipedia.org/wiki/Ontology_(information_science). (2017, 3 11). Retrieved from wikipedia.org: https://en.wikipedia.org/wiki/Ontology_(information_science

[27] BIBLIOGRAPHY \l 1033 (n.d.). Retrieved 3 19, 2017, from http://rss.softwaregarden.com/aboutrss.html

[28] Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In Machine Learning: ECML-98, pages 4-15.

IJARIIE