Ensemble Text Mining using NLP and AI Techniques

M. Aruna Safali[1], Dr. Ch. Suneetha[2]
[1]*Research Scholar, Dept of Computer Science & Engg., Acharya Nagarjuna University,*
[2]*Associate Professor, RVR & JC College of Engineering*

Abstract - The other name of text mining is text analytics. In the present world everyday huge data is generated everywhere such as social networking, airlines, spam mails and finding the interesting and important data from various sources. Text mining reads the data, analyze the data based on the topic present in that data. Natural Language processing (NLP) is the sub domain of the text mining. Many researchers have been done to solve the ambiguity problem the work is still immature. In this paper, the new ensemble approach which is merged with the features of text mining approach, NLP technique and artificial intelligence. This shows the result based on the documents mining with document or article belongs to which topic such as political, sports, technology and various domains. The dataset utilized for this project is 29-documents and 65-documents synthetic datasets.

Keywords - Artificial Intelligence, NLP, text mining.

I. INTRODUCTION

All kind of colleges, associations, and business ventures are using to store data in various databases. An enormous measure of content is streaming over the web as advanced libraries, archives, and other textual data, for example, online journals, web-based social networking system and messages [1]. It is demanding task to decide proper examples and patterns to retrieve important information from this huge volume of information [2]. It is very tedious for the existing data mining techniques to handle the textual data.

Text mining retrieves the interesting and various remarkable patterns to define the textual data [3]. A few content mining systems like outline, grouping, bunching and so forth, can be connected to remove information. Content mining manages regular dialect content which is put away in the semi-organized and unstructured organization [4]. Content mining procedures are persistently connected in industry, the scholarly world, web applications, web and different fields [5]. Application territories like web indexes, client relationship administration framework, channel messages, item proposal investigation, extortion discovery, and online networking examination utilize content digging for assessment mining, include extraction, supposition, prescient, and drift examination [6].

Extracting the precious data from various documents, HTML files and articles belongs to various domains such as sports, political, social networking etc. To identify the patterns in every article and categorize the articles based on the data present. In this paper, text mining algorithm is utilized to extract the information; NLP is used to extract the huge interesting data from the given articles. The AI algorithm Naive Bayes is used for feature extraction. The results are shown in three phases matching articles, feature extraction, and show the labels.

This paper is organized in different sections. Previous work is discussed in Section II. In Section III, existing techniques of text mining are explained. Section IV presents the proposed system. In section V, Results. Section VI concludes the outcomes.

II. LITERATURE SURVEY

This section II explains the previous works based on text mining, NLP and AI.

S.H. Liao *et al.* [5] that get-together, separating, pre-preparing, content change, highlight extraction, design determination, and assessment steps are a piece of content mining process. What's more, extraordinary broadly utilized content mining methods, i.e., grouping; arrangement, choice tree order, and their application in assorted fields are reviewed.

N. Zhong et al. [8] featured the issues in content mining applications and systems. They talked about that managing unstructured content is troublesome when contrasted with organized or unthinkable information utilizing conventional mining instruments and systems. They have demonstrated the uses of content mining process in bioinformatics, business knowledge and national security framework. NLP and ERT have lessened the issues that happen amid content mining process. In any case, there exist issues which require consideration.

A. Henriksson et al. [9] investigated MEDLINE biomedical database by incorporating a structure for named element acknowledgment, characterization of content, theory age and testing, relationship and equivalent word extraction, separate truncations. This new system takes out pointless subtle elements and concentrates important data.

B. Laxman and D. Sujatha [10] examined the content utilizing content mining designs and demonstrated term based methodologies can't investigate equivalent words and polysemy appropriately. Also, a model was intended for determination of examples regarding allocating weight as indicated by their conveyance. This approach upgrades the productivity of content mining process. C. P. Chen and C.-Y. Zhang [11] introduced a wrongdoing identification

framework utilizing content mining instruments and connection disclosure calculation was intended to correspond the term with contraction.

R. Rajendra and V. Saransh [12] displayed a best down and base up approach for online content mining process. To join the comparable content records, they apply k-mean bunching procedure for base up dividing. To discover the closeness inside the record TF-IDF (Term Frequency-Inverse Document Frequency) calculation has been utilized to discover data with respect to particular subjects.

K. Sumathy and M. Chidambaram [13] gave a diagram of uses, instruments and an issue emerges to mine the content. They talked about that records might be organized, semi organized or unstructured and separating helpful data is a tedious errand. They displayed a bland structure for idea based mining which can be imagined as content refinement and learning refining steps. The transitional type of substance portrayal mining relies upon particular area.

P. J. Joby and J. Korra [14] introduced imaginative and effective example revelation strategies. They utilized the example advancing and finding methods to upgrade the adequacy of finding pertinent and fitting data. They performed BM25 and vector bolster machine construct sifting in light of switch corpus volume-1 and content recovery meeting information to gauge the adequacy of the recommended method.

Z. Wen et al. [15] performed different examinations of grouping utilizing multi-word includes on the content. They proposed a hand-created technique to separate multi-word highlights from the informational collection. To characterize and separate multi-word content they partition content into direct and nonlinear polynomial frame in help of vector machine that enhance the viability of the extricated information.

III. EXISTING SYSTEM

Back propagation Network algorithm denoted as BPN. Various predictions are merged in this algorithm. The sigmoid function is the mostly the hidden model which is called as output function. The process of BPN produced in two phases here. Forward signal propagation happens in first phase network. Error objects occur in all other input units in the second phase.

The algorithm as follows:

Phase 1:
1. Find the primary handling highlight vector in the layer quickly over the current layer.
2. Set the present information aggregate to zero.
3. Calculate the result of the principal input association weight and the yield from the transmitting highlight vector.
4. Add that item to the aggregate.
5. Restate steps 3 and 4 for each info association.
6. Process the yield an incentive for this unit by applying for the yield work f (x)=1/(1+e-x), where x = input adds up to.
7. Restate steps 2 through 6 for each component vector in this layer.
8. Restate steps 1 through 7 for each layer in the system.

Once a output value has been ascertained for each unit in the system, the qualities processed for the classes in the yielding layer are contrasted with the coveted output choice, component by component. At every class in the yield, a mistake values is ascertained. These blunder terms are sustained back to every other unit in the system.

Phase 2:
1. Find the main handling unit in the layer promptly underneath the output layer.
2. Set the present blunder aggregate to zero.
3. Register the result of the principal yield association weight and the blunder gave by the unit in the upper layer.
4. Add that item to the total mistake.
5. Output steps 3 and 4 for each yield association.
6. Duplicate the combined blunder by o (1-o), where o is the yield estimation of the shrouded layer unit created amid the nourish forward activity.
7. Output steps 2 through 6 for every unit of this layer.
8. Output steps 1 through 7 for each layer.
9. Find the main preparing unit over the information layer.
10. Figure the weight change an incentive for the principal input association with this unit by including a small amount of the total mistake at this unit to the information incentive to this unit.
11. Adjust the weight change term by adding a force term equivalent to a small amount of the weight change an incentive from the past cycle.
12. Spare the new weight change an incentive as the old weight change an incentive for this association.
13. Change the association by including the new association weight change an incentive for this association.
14. Repeat steps 10 through 13 for each information associated with this unit.
15. Repeat steps 10 through 14 for every unit in this layer.
16. Repeat steps 10 through 15 for each layer in the system.

There are sure viewpoints worth saying in BPN. The principal thing is that BPN is great at speculation. Extraneous information will be overlooked. The second thing is that if the output work is sigmoidal, at that point we need to scale the output values. On account of the sigmoid capacity, the system yields can never achieve 0 or 1. Therefore use values such as 0.1 and 0.9 to represent the smallest and largest output values.

IV. PROPOSED ALGORITHM

The proposed algorithm is implemented with three phrases.

1.) Training with NLP.
2.) Pre-processing with text mining

3.) Finding the commonalities in the documents in a corpus and grouping them into predefined labels based on the topical themes exhibited by documents.

Naive Bayes is a very simple classification algorithm that makes some strong assumptions about the independence of each input variable.

In a classification problem, our proposition (p) may be the label to assign for a new data occurrence (o).

The selecting the most probable proposition given the data that are using in this paper. Bayes' Theorem provides a way that we can calculate the probability of a proposition given our prior knowledge.

Bayes' Theorem is stated as:

$$P(\text{p}|\text{o}) = (P(\text{o}|\text{p}) * P(\text{p})) / (\text{o})$$

Results:

Id	Text Category	ArticlesMatchingCount
1	Tech	32
2	Politics	8
3	Business	4
4	Sports	6
5	Entertainment	4
6	Health	11

Fig.1: Article Matching Count

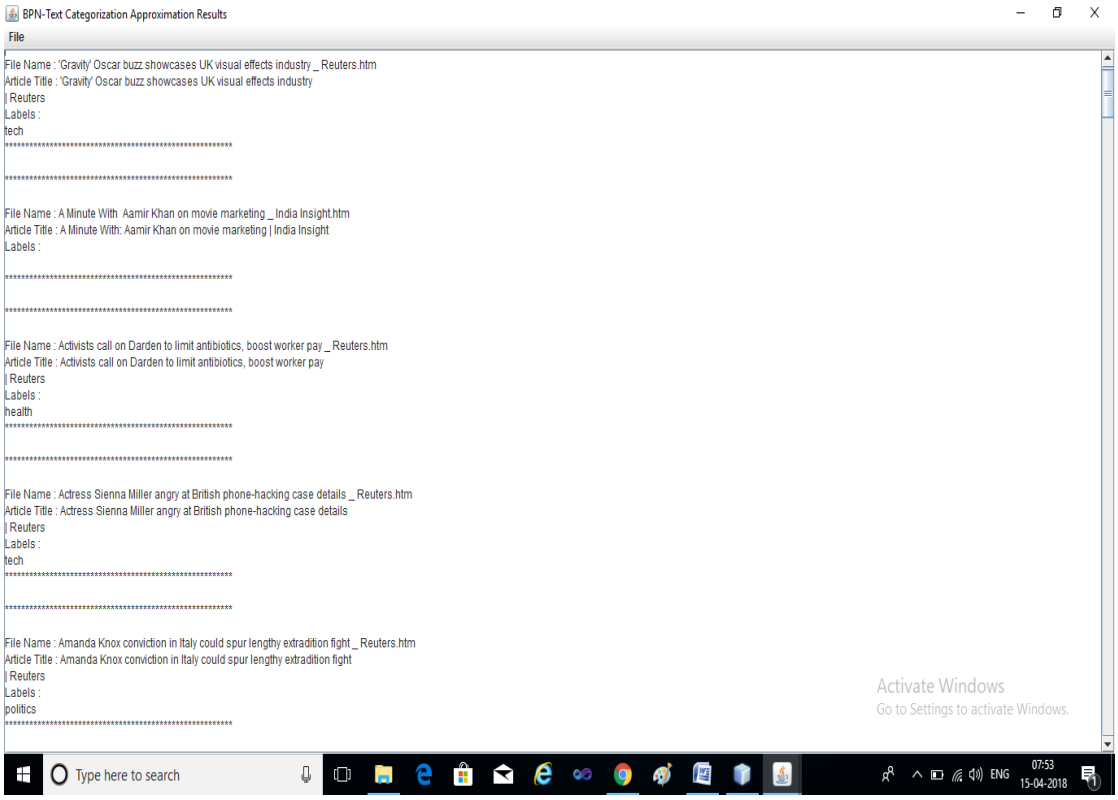

Fig: 2: Proposed Computation Results

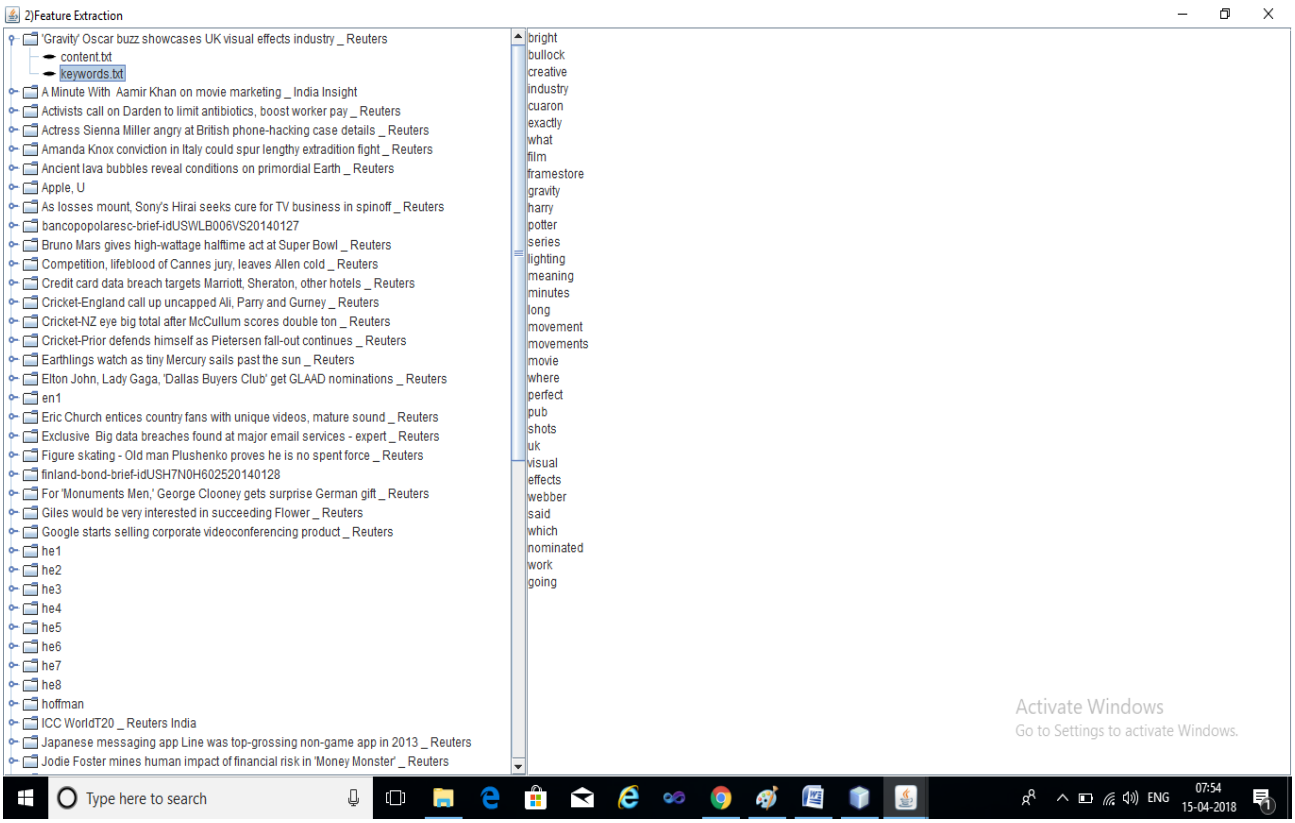

Fig.3: Feature Extraction

V. CONCLUSION

The proposed system new ensemble approach which is merged with the features of text mining approach, NLP technique, and artificial intelligence. To identify the patterns in every article and categorize the articles based on the data present. In this paper, text mining algorithm is utilized to extract the information; NLP is used to extract the huge interesting data from the given articles. The AI algorithm Naive Bayes is used for feature extraction. The results are shown in three phases matching articles, feature extraction, and show the labels.

VI. REFERENCES

[1]. R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

[2]. N. Padhy, D. Mishra, R. Panigrahi., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

[3]. W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

[4]. S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.

[5]. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[6]. W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.

[7]. G. King, P. Lam, and M. Roberts, "Computer-assisted keyword and document set discovery from unstructured text," Copy at http://j. mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper, vol. 456, 2014

[8]. N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012.

[9]. A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and ˇ M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.

[10]. B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013.

[11]. C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," Information Sciences, vol. 275, pp. 314–347, 2014.

[12]. R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," International Journal of Computer Applications, pp. 159–171, 2013.

[13]. K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications, tools and issues-an overview," International Journal of Computer Applications, vol. 80, no. 4, 2013.

[14]. P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on. IEEE, 2015, pp. 634–638.

[15]. Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan, vol. 51, 2007, p. 45.