Document Classification using Various Classification Algorithms: A Survey

Bipanjyot Kaur , Gourav Bathla

Department of Computer Science & Engineering

Chandigarh University, Mohali, Punjab, India

Abstract–Text classification is used to classify the document of similar types . Text classification can be also performed under supervision i.e. it is an supervised leaning technique Text classification is a process in which documents are sorted spontaneously into different classes using predefined set. The main issue is that large scale of information lacks organization which makes it difficult to manage. Text classification is identified as one of the key methods used for recognizing such types of digital information. Text classification have various applications such as in information retrieval, natural language processing, automatic indexing, text filtering, image processing, etc. Text classification is also used to process the big data and it can also be used to predict the class labels for newly added data. Text classification is also being used in academic and industries to classify the unstructured data. There are various types of the text classification approaches such as decision tree, SVM, Naïve Bayes etc. In this survey paper, we have analysed the various text classification techniques such as decision tree, SVM, Naïve Bayes etc. These techniques have their individual set of advantages which make them suitable in almost all classification jobs. In this paper we have also analysed evaluation parameters such as F-measure, G-measure and accuracy used in various research works.

.***Keywords* – *Text classification, supervised learning methods, Neural Network, Naïve Bayes and Complex algorithms.***

*****

I. INTRODUCTION

In today's era, there is large amount of information available in library science, computer science etc. sso the management of the information is done by text classification .As there is the tremendous increase in information or documents online text classification becomes essential for the management and standardization of the texture data. The amount of data is obtainable on the internet but its recovery is tedious. Unorganized information requires being altered from the result. There are several webcasting services which define the customized documents to the end-user based on the client. In this system consumers to create their interest profile which is difficult task and it is the manual procedure. According to this manual profile, creation is time consuming and tedious process [1].

Text mining studies relations, the designs and rules from the texture data. Features data that is extracted from the key attributes form novel facts or hypotheses to be explored further. Text classification is dissimilar from what are familiar with in web-search [2]. In the analysis, the consumer is normally looking for something that is called and has been written by someone else. Basically , text classification is also an important task in natural language processing. The collection of categories is given known as controlled vocabulary. The text classification refers to the act of assigning the class to the various documents. Text classification is normally used to separate the fake emails from the genuine emails, categories , also assigns the classes to the lager collection of documents .IT also used to handle training and also to help internet search-engines. A major drawback of text-classification is high dimensionality of the feature space [3]. The Naïve characteristics space consists of 100s and 1000s of terms refer as features for even a newly sized text classification [7]. This review also focuses on the several methods and also the applications of text-classification.

Several techniques included in text classification are: Naïve Bayes classifier, Expectation mining, SVM, Artificial Neural Network, concept mining, decision trees, etc. Text Classification is implanted in different fields that includes Spam Filtering to discern E-mails spam messages, Language Identification determines the text language, Data Mining determines the opinion and nature of writer as per the relevance to topic, health related classifications uses media for public health surveillance, etc.

II. LITERATURE SURVEY

[4] presented an approach using closest neighbouring algorithm with cosine analogy to classify research papers and patents published in several fields and stored in different conferences and journals database. Experimented results proves that user get better outcomes by traversing research paper or patent in specific category. The primary advantage of presented technique is that search area become compact and waiting time for query's solution has reduced. They have calculated the threshold depending upon similarity of terms of query, patent and research paper. Threshold calculation was not numerical value based. Hence the presented technique categorize more precisely than existing approach.**[5]** examined that social media posts can analyse the personal intelligence. Primary base of human behaviour is personality. Personality tests elaborates the individual's persona that influence the relations and priorities. User reveal their opinions on social media. The text classification was exploited to predict the character and nature on the basis of their comments. Indonesian and English language were used for this test. Naïve Bayes, SVM and K-Nearest

Neighbour are executed methods for classification. Naïve Bayes performed better than other techniques. The research work uses MyPersonality dataset. In this dataset used to classify the personality based-on an online ques **[6]** traversed internet for huge data to gather knowledge. It consists of huge unstructured data like text, image and video. Challenging issue is organization of big data and gather useful knowledge that could be used in bright computer system. Ontology covers the big area of topic. To construct an ontology with specific domain, big dataset on web was used and arranging with particular domain before the completion of organization. Naïve Bayes classifier was implemented with Map reduce model to organize big dataset. Plant and animal domain articles from encyclopaedia available online were used to experiment. Proposed technique yielded robust system with high accuracy to classify data into domain specified ontology. In this research work, datsets use plant and animal domain animals article in online encyclopedia and Wikipedia as dataset. **[7]** presented a Bayesian classification technique for text categorization using class-specific characteristics. Unlike regular approaches of text categorization proposed method chosen a particular feature subset in every class. Applying such class-dependent characteristics for classification, a Baggenstoss's PDF Projection Theorem was followed to recreate PDFs from class-specific PDFs and construct a Bayes classification rule. The importance of suggested approach is that feature selection criteria, like: MD (Maximum Discrimination), IG (Information Gain) are included easily. Evaluated the performance on several actual benchmark data set and compared with feature selection approaches.The experiments , they tested approach for texture classification on binary real time benchmarks : 20-Reuters and 20-Newgroups.**[8]** proposed a BI-LSTM (Bi-directional long short term memory) network to inscribe the short text classification with 2 settings. The short-text classification is required in applications of text mining, especially health care applications in short texts mean linguistic ambiguity bound semantic expression due to which traditional approaches fails to capture actual semantics of limited words. In health care domains, the text includes infrequent words, in which due to lack of training data embedding learning is not easy. DNN (Deep neural network) is potential to boost the performance as per their strength of representation capacity. Initially, a common attention mechanism was adopted to guide network training with domain knowledge in dictionary. Secondly, direct cases when knowledge dictionary is unavailable. They presented a multi-task model to learn domain knowledge dictionary and performing text classification task in parallel. They applied suggested technique to existing healthcare system and exclusively available ATIS dataset to get better results. **[9]** surveyed the process of text classification and existing algorithms. Large amount of data is stored as e-documents. Text mining is a technique of extracting data from these documents. Classifying text documents in specific number of pre-defined classes is Text classification. Its application consists of email routing, spam filtering, language identification, sentiment analysis, etc.**[10]** introduced a fuzzy logic based technique to solve text classification. Data inserted in proposed model are extracted from twitter's message. Social media offers plenty of data to study human behaviour. Hurricane Sandy 2012 was used to extract information and classifying text. It's beneficial to analyse the relation between human influenced events and social media. Several fuzzy rules are designed and de-fuzzification methods were combined to get desired results. Suggested technique was compared to popular search method as per rate and quantity correctness. Results shows that proposed technique is suitable for classification of twitter messages. The experimental uses the twitter review using social media. **[11]** proposed a technique which uses the connection between lexical things and labels before finishing Latent Dirichlet Allocation (LDA) theme display. They modified parameters of SVM (Support Vector Machine) to locate optimized values by K-crease cross approval. It's an awesome test that comprehending high-measurement and content sparsity issues in short content arrangement. Also, utilizing piece SVM as classifier, we effectively arrange named short Chinese content reports. Contrasting and other two regular techniques k-Nearest Neighbour and Decision Tree of short content arrangement, the exploratory outcomes demonstrate that our strategy outflanks them on order exactness, accuracy, review and F-measure. **[12]** suggested an improved KNN text classification algorithm depending on Simhash and average Hamming gap of adjacent texts as an item that solves the problems generated by data imbalance and the large computational overhead in the traditional KNN text classification algorithms. Experimented results explained that proposed algorithm performs with higher precision and recall and better F1 value. The experiment uses the mailing dataset like electronic and internet mail data.

In below table 1 defined that the previous performance parameter in text classification like accuracy, precision, recall, F-measure and G-measure and similarity etc.

Table 1: PerformanceParameters in Related Work

Research Work	PARAMETERS				
	Accuracy	Precision	Recall	F-measure	G-measure
[4]	✓	X	X	X	X

ISSN: 2454-4248
150 – 155

[5]	✓	✓	✓	✓	X
[6]	✓	X	X	X	X
[7]	✓	✓	✓	✓	✓
[8]	✓	X	X	X	X
[9]	X	X	X	X	X
[10]	✓	X	X	X	X
[11]	✓	✓	✓	✓	X
[12]	X	✓	✓	✓	X

III. TEXT CLASSIFICATION

In an active research field of text classification where the documents are classified with un-supervised, supervised and semi-supervised knowledge.

It refers to resolving the issue to identify documents based on their content into a definite number of predefined classes. To classify a novel document is the main aim of classification.

It plays a significant role in wide variety of fields such as information-retrieve, web-pages classification and several more[13]. Text Classification process are explained as follows:

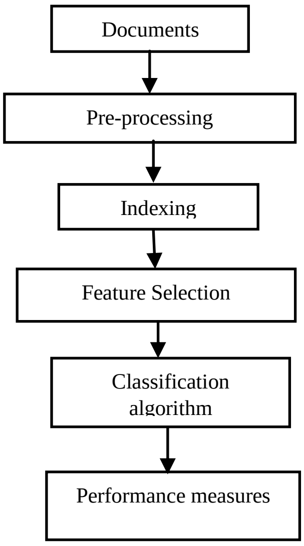

Fig1 :document classification

- ***Document Collection:*** The initial step is to collect different formats of documents such as html, web content, .pdf, .doc, etc.
- ***Pre-processing:*** It consists of 2 steps. First step is tokenization presents text document into word format and in second step text represented by number of features to conflate the token to their root format, e.g. computing to compute
- ***Indexing:*** it's used to minimize the complexity of document and model them easy to handle.
- ***Feature Selection:*** used to construct vector space to enhance scalability, accuracy and efficiency of classifier. The main concept of FS is to choose subsets from original document.
- ***Classification:*** automatically classifies the document into pre-set classes. A document is classified by 3 methods: supervised, semi-supervised and unsupervised. Several approaches are also used in recent times.
- ***Performance Evaluations:***Text classification evaluation is regulated experimentally instead of analytically. Classifier evaluation focussed to evaluate effectiveness i.e. capable to proceed with accurate classification decision. [15].

The applications of the text classification are as follows:

- Automatic indexing
- Document Organization [18]
- Text Filtering and Word-sense Dis-ambiguation
- Hierarchical web-page classification [14]
- Spam Filtration
- Target marketing[18]

The text classification types are defined as:-

- ***Multilabel and Single Label Text Classification[23].***

In this case, where only single class is assigned to the input texture is refer to as single label classification , Whereas, the case in which more than single class are assigned to the input texture is refer to as the multilabel classification.

- ***Document and Category Text Classification.***

Offered a document, the classifier selects every category to which the document belongs. This is specified as a document classification. Like, the classifier which finds all the documents that must be recorded under category text classification.

- ***Hard-Soft Text Classification.***

Hard classification is based on a binary decision completed by automated classification system on each document category couples, while soft-classification defines a ranking the input documents/ ordered the outcome categories is known as ranking classification.

IV. TEXT CLASSIFICATION APPROACHES

Recently, several methods are accessible for classification of text documents such as SVM (Support Vector Machine), Naïve Bayes and Decision Tree etc[16].

4.1 Decision tree

Decision tree is basically heirarchical in structure .It consist of the root node, internal and leaf nodes. Decision Tree method is used[21][26] re-construct the manual classification of the training documents .It creates well-known true and false queries that is in the form of heirarchy, where the nodes indicates about the query and leaves they indicate the class of the document. To classify a novel document, after tree being created set is as the root of the tree and execute query until it spread its main root. Decision tree is utilized because it's outsput is easy to understand. It can also be easily understood by the consumers who are not familiar with details of model. The tree design created by the model gives the consumer with a combined-view of the classification logic and tree techniques is called as over-fitting.

Decision tree shows the better performance with the huge no of records Classifiers represented as feature vectors. Decision trees are used to predict its label .Decision tree is also used in the various research works[13][16][21][26]

.

4.2 K-NN (K Nearest Neighbour):

K-NN compare the classify occurrences of the k-nearest documents [25]. It instance based algorithm. K-Nearest Neighbour is an easy method that stores all available phases and classifies a novel phases based on a similarity measures.

KNN can be implemented in the forms of phases . In the training step: A model is built form the training instances. The classification method searches relationships b/w identification and targets. The relationships are explained in a model wise. In testing step: Model test on a test sample whose group labels are called, but not used for training the model. Usage step: Utilize the model for classification on novel data whose group labels are un-known. KNN is also used in the various research works[12][17][23].

4.3 Naïve Bayes:

Naive Bayes combines the probabilities of characteristics to estimate the group probabilities of a document following to Bayesian formula. Classifiers are completely applied in document classification. Two different normal models in common use of the NAÏVE BAYES. [22]

Single model reveals that a document be defined by a feature-vector of binary instances indicating the words which are defined and undefined in the document. This is called as binary Naïve Bayes Classifiers.

Secondly, the model identifies that a document can be defined by the set of word occurrences from the word. This is called as multi-nomial Naïve Bayes algorithm[20].

The specific illustration y,y = (Y_1, Y_2,…… Yn), where y_i are the characteristics of the illustration. The odd-ratio between negative and positive classes is

$$\text{AR} = \frac{\prod Ph\ (Yi\ |Cl{=}Cli\ )Ph\ (Cl{=}ci)}{\prod(Ph(Yi{=}yi|Cl{=}Co)Ph\ \ (Cl{=}Co)}$$

Where Cl defines the positive group and co-defines the negative group. The classical Naïve Bayes classifier use the provisional group probabilities of characteristics defined and not defined in a document. The class probabilities of characteristics defined in a document, thus mainly optimizing calculations while at the similar interval infrequently hurting the performance. It is called by the Asymmetric Naïve Bayes classifiers.

$$\text{AR}_{\text{anb}} =\left(\frac{\prod Ph\ (Yi{=}1|Cl{=}ci)Ph\ (Cl{=}Ci)}{\prod(Ph\ (Yi{=}1|Cl{=}c0)Ph\ (Cl{=}C0)} \right)^{1/|0|}$$

Where A = { i|Yi = 1} is the sub-script set of all of characteristics defiend in the illustration.

In this algorithm implement the three ways:

- ***Guassian:***This is used for classification purposed and it defines that characteristics prefer ND (Normal Distribution).
- ***Multi-model:*** In this methods used for a discrete counts.
- ***Bernoulli Methods:*** Binomial Model is valuable if the feature-vectors are binary form (one and zeros). Application will be text classification with bag-of-words model, then the ones and zeroes are word defines in the document.

Naive Bayes is used in various research work[7][19][20][22]

4.4 Support Vector Machine(SVM):

Support Vector machine algorithm described [24] that the SVM classifier is a classifier with maximum margin . SVM it can be used for the classification and regression. It's well defined that the Support Vector Machine classifiers could simply outperform all the conventional classifiers on the text classification jobs. SVM is used in various research works[1][2][11] [17][23][24].

4.5 Artificial neural networks (ANN):

It is a computing system consists of simple yet highly inter-connected elements in which information processed by their fresh energetic state response.ANN also tries to simulate the human learning process. In ANN information transmission occurs by the neurons .It is the output of one neuron is given as input to the other neuron The concept of ANN is inspired from biological network of artificial neurons designed for particular task performance. ANN performs several tasks like clustering, pattern recognition, classification, etc [27][28].

So, we have analysed the various approaches for the text classification. In above section we described that the various techniques of classification in machine learning and artificial intelligence such as Naïve Bayes, SVM and ANN.

The text classification algorithm main advantages i.e Support Vector Machine used to optimize the dimension of the information. Artificial Neural Network main advantages have the ability to study and structure non linear and difficult relationships, which is really significant since in real life several of the relationships between input and outputs are non linear as-well-as difficult .Naïve Bayes algorithm benefits are easy to implement and it can compute the performance parameters by using minimal amount of the training data

CONCLUSION

One of the important application of data mining is text classification. We have surveyed the Text Classification and its process and several approaches to text classification. Decision tree is heuristic approach where distribution is not required and good for several category variables. Naïve Bayes is good for several category variables and compute multiplication of independent distributions. ANNs are highly convex, inherently and do not need to get trained using straightforward techniques In this survey paper we have analysed and compared various classification techniques such as SVM, ANN, Naïve Bayes, decision Tree etc. We have also analysed the performance evaluation parameters used in several research works like F-measure, G-measure and accuracy. It is concluded from this survey that there is need to analyse text classification with maximum evaluation metrices which was not used in many research works.

References

[1] Erlin,et al., "Text message classification of collaborative learning skills in online discussion using support vector machine," In *Computer, Control, Informatics and Its Applications (IC3INA), 2013 International Conference on*, pp. 295-300. IEEE, 2013.

[2] T. Joachims, "Transductive inference for text classification using support vector machines," In *ICML*,vol. 99, pp. 200-209. 1999.

[3] D.Xue& F. Li. "Research of Text Classification Model based on Random Forests." In *Computational Intelligence & Communication Technology (CICT), IEEE International Conference on*, pp. 173-176. IEEE, 2015.

[4] B. Gourav& R. Jindal, "Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold," International Journal of Computer Applications, vol. 33, no. 5. 2011.

[5] B.P.Yudha, and R. Sarrno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *In Data and Software Engineering (ICoDSE)*, in proceedings od International Conference on, pp. 170-174. IEEE, 2015.

[6] J. Santoso, E. M. Yuniarno, et al., "Large Scale Text Classification Using Map Reduce and Naive Bayes Algorithm for Domain Specified Ontology Building." *In Intelligent Human-Machine Systems and Cybernetics (IHMSC*), in proceedings of the 7th International Conference on, vol. 1, pp. 428-432. IEEE,2015.

[7] B.Tang, H. He, et al., "A Bayesian classification approach using class-specific features for text categorization." IEEE Transactions on Knowledge and Data Engineering 28, pp: 1602-1606,no. 6, 2016.

[8] S. Cao, B. Qian, et al.," Knowledge Guided Short-Text Classification for Healthcare Applications", 2017 IEEE International Conference on Data Mining (ICDM) vol. 2, no. 6,pp: 234-289. 2017.

[9] V. K. Vijayan, K. R. Bindu, et al., "A comprehensive study of text classification algorithms." IEEE Advances in Computing, Communications and Informatics (ICACCI),, vol 12, no. 1 pp: 42-53. 2017.

[10] K.. Y. Wu, M. Zhou, et al., "A fuzzy logic-based text classification method for social media data," Systems, Man, and Cybernetics (SMC), IEEE International Conference on, vol.13,no.3 pp:23-32. 2017.

[11] X. Wang, J. Wang, et al., "Labelled LDA-Kernel SVM: A Short Chinese Text Supervised Classification Based on Sina Weibo." In 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 428-432. IEEE, 2017.

[12] J. Liu, T. Jin, et al., "An improved KNN text classification algorithm based on Simhash." In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2017 IEEE 16th International Conference on, pp. 92-95. IEEE, 2017.

[13] M.Somvanshi, P. Chavan. "A review of machine learning techniques using decision tree and support vector machine," In *Computing Communication Control and automation (ICCUBEA), International Conference on*, pp. 1-7. IEEE,2016.

[14] W.B. Michael, *Survey of Text Mining: Clustering, Classification and Retrieval",* "Automatic Discovery of Similar Words.", *Springer Verlag, New York* 200 (2004): 25-43. 2004.

[15] V. Korde, and C. N. Mahender. "Text classification and classifiers: A survey."*International Journal of Artificial Intelligence & Application,vol.* 3, no. 2 pp: 85. 2012

[16] S. Yasubumi, K. Misue, et al., "Text classification and keyword extraction by learning decision trees," In *Artificial Intelligence for Applications, in Proceedings of the 9thConference on*, pp. 466. IEEE,,1993.

[17] V. Gupta, and G. S. Lehal,"A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence* ", no. 1 pp: 60-76. 2009.

[18] C.C. Aggarwal, C. X. Zhai. "A survey of text classification algorithms," *Mining text data* , pp: 163-222. 2012.

[19] A.McCallum, K. Nigam "A comparison of event models for naive Bayes text classification." In *AAAI-98 workshop on learning for text classification*,vol. 752, pp. 41-48. 1998.

[20] A. Y. Ng, and MI. Jordan.,"On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes",In *Advances in neural information processing systems*, pp. 841-848,2002.

[21] H. Schmid,"Probabilistic part-ofispeech tagging using decision trees", In *New methods in language processing*, pp. 154,2013.

[22] Kohilavani, S., T. Mala, and T. V. Geetha. "Automatic tamil content generation," In *Intelligent Agent & Multi-Agent Systems, 2009, in proceedings of International Conference on*, pp. 1-6. IEEE, 2009.

[23] R.Jindal, R. Malhotra, et al., "Techniques for text classification: Literature review and current trends." *Webology* 12, no. 2 pp: 1-28, 2015.

[24] Cortes, C.; Vapnik, V. "Support-vector networks". *Machine Learning*.vol. **20** (3):pp:273–297. doi:10.1007/BF00994018,1995.

[25] N. S. Altman "An introduction to kernel and nearest-neighbour non parametric regression,". *The American Statistician*,vol.46 (3),pp:175–185,1992.

[26] B. R. Patel, and K. K. Rana.," A survey on decision tree algorithm for classification", *International Journal of Engineering Development and Research*,vol.2(1),2014.

[27] G. F. Hepner, T. Logan, et al., " Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification", *Photogrammetric Engineering and Remote Sensing*, vol.*56*(4), pp469-473, 1990.

[28] E. Byvatov , et al," Comparison of support vector machine and artificial neural network systems for drug/nondrug classification", *Journal of chemical information and computer sciences*,vol. *43*(6),pp. 1882-1889, 2003.