Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

A Curated and Comparative Study on Semi-Supervised Learning Techniques for Text Classification

Priyanka N [1] and Sumukh B K [2]
[1][2] Undergraduate Student, Department of Information Science and Engineering,
Global Academy of Technology, Bangalore.

***Abstract*:- For classifying a text from a document involves several types of learning techniques supervised, unsupervised together with Semi-supervised. For improved accuracy, supervised learning necessitates many labelled documents; yet labelled documents are generally hard and steep to gather, whereas unlabeled documents may be collected quickly. This paper compares learning algorithms like expectation-maximizations, Naive- bayes classifier, and different types of vector methods to enhance the accuracy of the training documents. From our study it was found that SFE (Supervised Frequency Estimation) consistency produces better conditional likelihood values and shows lower computational cost than combining expectation-maximization and Naive Bayes.**

I. INTRODUCTION:

Semi-supervised learning is a machine learning approach in which a little quantity of labelled data is combined with a large amount of unlabeled data to train a model. Semi-supervised approach is an intermediate to supervised learning and un-supervised learning. Historically semi-supervised learning has gained a lot of popularity, with it being used in a variety of applications as in self-training of object detection models [1] medical image segmentation [2] speech quality assessment [3] content classification [4] text classification [5]. When unlabeled data is combined with a modest bit of labelled data, learning accuracy can be significantly improved. Additionally, obtaining a complete set of labeled data is quite a complex and costly task, thus semi-supervised learning makes the training an easier process.

Text classification is a machine learning approach for categorizing open-ended text into a collection of predetermined categories. Text categorization is a fundamental problem in natural language processing with several plea such as sentiment analysis, topic labeling, junk detection in mail, and fixed detection. Semi-supervised learning attempts to make use of labelled as well as unlabeled documents to improve the text classification, labeled documents can be defined as dataset which has complete set of attributes and values. Unlabeled documents consist of a complete set of attributes, but some attributes do not contain values. SSL makes use of labeled documents to build an algorithm and this algorithm can be used in sorting the text from unknown quantity in documents. Semi-Supervised Learning can be used in various fields such as classifying web pages, names in a text etc. Text Classification organizes and structures any kind of texts from the documents. For example, when a user creates a review system in which reviews of people all over the world reaches a user in those hundreds of reviews, the system must be able to differentiate good review, bad review and average review using text classification. The system does this part and categories the input reviews.

The problem that develops during the implementation of semi supervised learning is that they mainly rely on assumptions which might be true or false. The semi-supervised learning algorithm and makes the following assumes the following assumptions.

1. Continuity assumptions: The method expects that points that are near together will possess the related output tally.
2. Cluster assumption: The statistics may be divided into groups, with points in the same cluster having a greater likelihood of having the same output label.
3. Manifold assumption: The input is approximated to be on a manifold with a actually limits dimension than the input; this predictions allows for usage of distance and frequency specified on a manifold.

This paper is standardized as follows. In section 2, literature survey we have surveyed through the different research papers their idea and their methods. In section 3, using the comparison table we have indicated the different results obtained from the literature survey.

II. LITERATURE SURVEY

Semi-Supervised Text Classification from Unlabeled Documents Using Class Associated Words

Class related words are words that indicate the topic of classes and give previous classification knowledge for training a classifier, like attributes that represent the whole column in the datasets that functions as the information's leader. A technique expected-maximize with naive bayes approach has been created for organize papers from unknown texts using subject relevant phrases. Class related terms are adapted to impose grouping restrictions during the learning process, limiting document categorization to matching class labels and improving classification accuracy. For example, if we have 10,000 papers, we may divide them into copyright and patent cases. If we, do it manually, it would take longer, but semi-supervised learning provides a superior option. We can improve accuracy by increasing the number of labeled documents because finding labeled datasets is challenging and time-consuming. Nigam realized that the unlabeled include information on the joint probability, which is useful for classification accuracy, and created an approach that combines expectation-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

maximization with the naive bayes classifier. This work begins with a generative model approach in which a mixed model, such as a union EM and a NB, is used, which requires a set of training sets for each class. To begin, train a labeled dataset and assign probable class labels to every unknown value using the classifier acquired from training the labeled dataset, then train a classifier that combines the labeled and unlabeled datasets using expectation-maximization iteration. The next phase is to collect class-related terms that have a lot of significance and indicate the relative class subject. Then there is defining the bounding between classes and class-based words, which includes mapping such as one-one mapping, which is difficult to get since one class controls numerous subjects and is referred to many classes related words to solve this difficulty. To map between class and class related terms, multi-one mapping is utilized. After mapping, utilize probable weights to documents based on class-based terms for each document, then repeat the process for an accurate prediction using Nigam's approach [6]. The last stage explains document categorization by referring to the likely weights of class-related terms. This divides the two ways into two categories. The first involves commonly choosing class labels to documents using the maxi membership degree order. The second step is to define a threshold, after which only papers with a maximum degree larger than the threshold will be tagged.

Semi-Supervised Text Classification Using Enhanced KNN Algorithm

K nearest neighbor is a term used to describe a person's closest algorithm and the fundamental machine learning algorithms are based on the supervised learning approach. KNN thinks that the unused data and existing instances are comparable and keeps the new base in the division that most closely resembles the existing categories. The K nearest Neighbor algorithm observes all input data and creates new data points based on parallel. This implies that fresh data may be quickly sorted into a suitable category using the K-nearest neighbor algorithm. Although the K-nearest neighbor technique may be used for the two regression and classification, it is more often employed in classification problems. KNN is a non-parametric algorithm, meaning it makes no assumptions about the data KNN simply stores the dataset during the training phase, and when it receives new data, it classifies it into a category that is very similar to the new data. For example, if we have a picture of a cat or dog, the KNN algorithm uses the features of the new data and compares them to features that have been previously trained, and it places it in either the cat or dog category. They have considered many vector methods before going into the prediction in this paper, which will enhance the KNN and improve accuracy. The methods include binary vector, frequency vector, normalization, length normalized binary vector with uncommon words, length normalized binary vector with every words, min-max normalization, length normalized frequency vector with uncommon words, length normalized frequency vector with every words, root mean-square frequency normalization, length normalized frequency vector with every words, root mean-square frequency normalization, length normalized frequency vector with all words. By examining various results, improvements in KNN and implementation may be suggested. K-nearest neighbor classifiers work by analyzing a given test tuple to training tuples that are like it, or by learning by analogy. The goal of this work is to categorize the data using semi-supervised classification with various similarity metrics and vector generation approaches. All the similarity metrics produced nearly identical findings, allowing the end users to pick between the several parallel measures.

Semi-Supervised Learning for text categorizing for Text Classification by Layer Partitioning

Semi-supervised algorithms, according to numerous studies, rely mostly on steadiness which particularly forces the framework to generate good accurate assumptions on the provided input. In image classification, appearances of image that can be expressed by dense vectors in a continuous space are utilized as inputs. Each token in the input text is characterized as a one-hot vector, resulting in a rare high-dimensional space, like text categorization. This work advocated perturbing all word by adding adverbial reports to the word submerging to prevent continuous noises to discrete inputs. The noisy output from a perturbation function on a sentence should nonetheless reflect a legitimate sentence with equivalent meaning. Many strategies, such as layer partitioning neural networks, perturbation textual input consistency restrictions, and progressively freezing, are used in semi-supervised learning utilizing layer partitioning. We have seen a semi-supervised framework for different text input in this research. We can observe the competitive outcomes obtained by merging the two models on various text categorization. Furthermore, without LM precise tweaking, this framework provides higher performance.

Large Scale Text Classification Using Semi-Supervised Multinomial Naïve Bayes

In this case, the multinomial nave bayes employs a learning approach known as frequency estimate, which calculates suitable frequencies from data to estimate word likelihood. Frequency estimations provide a good prediction performance and are timesaving. First, we can represent the text document in various forms that are suitable for classification, such as d=w1, w2..., wi, c, where wi represents the variable and c represents the class label in this often arranged in bag of words approach. This approach states that a document is frequently stored using the sparse condition, in which only words other than zero are stored. The multinomial naive bayes algorithm is a probable learning approach used in natural language processing. The naive bayes classifier is a consumption of variety algorithms with one common principle: all features. A feature's existence or absence has no bearing on the existence or vacancy of another feature. Using the Bayes theorem, the program estimates the tag of a text, alike email or a newspaper. It assesses the likelihood of each tag for a given sample and returns the tag with the highest likelihood. We may see multiple combinations of algorithm expectation-maximization in this work, which maximize the log

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

likelihood, proving that the p(w) information from unlabeled texts is used. Semi-supervised frequency estimate that combines the word frequency derived through unsupervised learning with the supervised learning's class prediction for that word. The performance of the semi-supervised frequency estimate is better than the combination of EM and Nave bayes when the two techniques are compared. SFE enhances the accuracy of multinomial nave bayes in a consistent and substantial way, as well as producing superior log likelihood. The final study and findings, however, reveal that Semi-supervised Frequency Estimation has a significantly lower computing cost than the combination of expectation-maximization and nave bayes.

Semi-Supervised Text Classification Using Unsupervised Topic Information

Unsupervised systems attempt to construct models by exploiting dependencies and similarities in unlabeled training materials. The first hypothesis model is built to categorize text documents using information obtained from of a particular document is calculated after they utilized a limited number of labeled documents to increase the number of trainings to increase the amount of training data. Many stages are involved in data augmentation. It entails obtaining a list of keywords from the entire training data. The previously generated list of keywords is used to train the naive bayes classifier using labeled dataset, and the classifier is then used to predict the categories of the unlabeled half of the training data. All of this demonstrates that an acceptable performance may be achieved with a semi-supervised technique and a small amount of training data.

A New SVM Method for Short Text Classification Based on Semi-Supervised Learning.

Short Text are the text forms which are commonly used in text fields, microblogging, short commentary etc. As the bulk of the data grows higher with each passing day, brief sentences are becoming increasingly significant in big data. A semi-supervised learning technique and a support vector machine can be used to accomplish this (SVM), which searches for significant data in the sort text and decreases the data's size. We need to examine the obtained data because there is a lot of unnecessary information in the brief, which diminishes the value of text classification. Data dictionary information Z to fuzzy which is invalid but still exists it matches the words from the short text and remove the unwanted information so here, we gain apply the Semi supervised learning to label the unlabeled data. The SVM classification model is used to train the sort text, which is already prearranged, and a semi-supervised learning algorithm is used to analyze the similarity between the samples on each iterative training set. This is done until all of the samples in the training set have been labelled completely. Comparison of the KNN algorithm and the algorithm which is proposed in this paper and showed the results, when compared to the KNN approach, the examination of the algorithm provided in this research shows that it is more accurate, therefore employing the SVM classifier and semi-supervised learning has produced better results.

an unsupervised model, while the second hypothesis model is built to generate a system whose performance comparable to state-of-the-art approaches while employing the least number of instances feasible. Using an algorithm to get a list of keywords, and then training a Bayesian classifier using these keywords as features to supplement the data. This is a study of the word size that is required to train various classifiers using either labeled data or labeled data along with augmented data. The latent Dirichlet allocation approach is a generative probabilistic model built of probabilistic mixtures that describe distributions over words that is employed here. The LDA model posits that a probabilistic process samples words from a large lexicon to build a collection of texts. To enhance the number of features collected with discriminative or supervised models, these words compose each subject. The Nave Bayes model for data augmentation LDA generates a list of keywords based on the topic distribution on the entire dataset. The likelihood

Text Classification Based on Semi-supervised learning.

This paper presents the solution and experimental findings of semi-supervised learning techniques application and development of the SVM algorithm. They started by creating tagged data and then improved it with unlabeled data. comparison of the accuracy of the classification and enhances the classification quality. As we all know that Machine Learning is a method widely used for problem-solving of recognition and classification, to get a good and quality model is incredibly challenging because we will have to get many data which is exceedingly rare and awfully expensive to overcome this constraint, a new technique is used that is called semi-supervised machine learning. They are attempting to construct a features model that incorporates the typical components of training documents, which will then be used to access documents into various classes, using a simple example of text categorization. A self-training training algorithm is used which makes use of the small amount of labelled data this algorithm is used in semi-supervised learning.

Semi-supervised Fuzzy Learning in Text Categorization.

This study proposed semi-supervised learning based on the Fuzzy C-means algorithm, which improved the effectiveness of the classifier, allowing it to accept a high number of unlabeled samples and a small amount of labelled data. Fuzzy C-means (FCM) is a clustering approach that permits points to be in several clusters. Dunn created this technique in 1973, and Bezdek enhanced it in 1981, resulting in a new method known as the supervised Fuzzy C-means text classifier (SFCTC) This algorithm is a text classification procedure based on text samples. The stages begin with filtering all stop words and conclude with the creation of a pure classification model using the training data, detail steps in [14]. Hence the algorithms like SFCTC, KNN and Naive Bayes analyses the text classification in the same means, in [14] comparing all the three algorithms, The accuracy of Naive Bayes was kept low, with only 40% accuracy. With the labelled samples, KNN fluctuated and did not increase. SFCTC, on the other hand, obtained an accuracy of 80

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

percent, which is the highest of the three. They used the Must-link and Cannot-link mechanisms to introduce semi-supervised strategies in the text classification sector in this article.

Keyword-Based Semi-Supervised Text Classification.

Proposes natural learning that is semi-supervised. The performance of the semi-supervised approach given here is comparable to a supervised classifier. Classifying the underrepresented class makes a significant difference in [15], also discussed is the semi-supervised classification process, which maintains an acceptable balance between supervised and unsupervised learning on one end of the spectrum and unsupervised learning on the other. also described three steps in the semi-supervised classification methodology, namely

1- Preparing the Inputs
2- Building a Dictionary
3- Evaluating Performance

They measured the efficacy of our semi-supervised technique using Precision, Recall, and F1 score; these methodologies are used to categories accounts receivable debates for many organizations. Table 1 in [15] shows the dispute categories and number of root causes. [15] Before arriving at the procedure, the data is subjected to four forms of pre-processing, these pre-processing decreases vocabulary from 4799 unique words down to 304 by removing English stop words, by applying Lancaster stemming, removing extremely rare words, and removing extremely familiar words. Results say that most of the semi-supervised classifier works only when a small amount of labelled data is available so here in this paper, they made use of the small amount of data, samples, or keywords rather than making use of unwanted and huge data.

Semi Supervised Learning Based Text Classification Model for Multi Label Paradigm

Automatic text categorization represents multiple label text classification domains. In 2006 Liu, Jīn and Yan presented a limited non-negative matrix factorization-based multi-label classification method. The use of semi-supervised learning in multiple-label text categorization improves the classifier's decision-making ability. In this paper, they have even formulated multi-label classification data using semi-supervised learning so that classifier can handle both labelled and unlabeled data. This report also included a detailed graphic of the proposed categorization model. Multi-label learning is well-versed in issue translation and algorithm adaptation techniques, and there are a few popular algorithms presented [16]. There are some approaches for multi-label learning which are supervised in nature. They developed the Proposed Classifier model to increase the accuracy of the multi-label classification process because this classifier is based on a semi-supervised learning technique, they used both labelled and unlabeled texts for training. [15] mentioned the performance measures by Precision, Recall, F1 score after the performance check they have conducted the experiments on four text-based datasets namely Enron, Slashdot, Bibtex and Reuters in [16] table 1 is shown the results of the experiment. To evaluate the performance of the classifier model using a Semi-supervised learning approach comparison of a few results of the supervised algorithm like C4.5, Adaboost, ML-kNN, BP-MLL, SVM-HF these are algorithm adaption method and BAkEL, MetaLabeler, CC, PS and EPS are problem transformation method.

SL No.	Author/Citation	Methodology used	Result
1	HAN Hong_qil, , WANG Xue-feng ,ZHU Dong-Hua [7]	Expectation-Maximization and naïve Bayes	Classification without constraints has an accuracy of 86 percent, whereas classification with constraints has an accuracy of 91 percent. Both techniques provide a high level of categorization precision. Clearly, the accuracy of employing classification constraints is substantially greater than learning a classifier for the first time using random articles.
2	Mohammad Abdul wajeeb, T. Adi Lakshmi [8]	Different vector generation techniques, K-Nearest Neighbor method	According to the results of the experiment, the square root of the mean results are superior to other vectors, as proven by the author.
3	Alexander Hanbo Li, Abhinav Sethy [9]	layer partitioning neural networks, perturbation textual input consistency restrictions	The author has proposed the semi-supervised framework especially for discrete text images, also attains better performance for short texts like personal memory retrieval without Language model fine-tuning.
4	Jiang Su, Jelber Sayyad-Shirbad, Stan Matwin [10]	Combination of expectation-maximization naïve bayes, Semi-supervised frequency estimation	Using Expectation-Maximization reduces Multinomial Naive Bayes's area under Curve by 6% in the supplied 512 labeled documents and enhances MNB's AUC by 2% in the given labeled documents. Semi-supervised frequency estimation significantly improves accuracy when compared to unsupervised frequency estimation. For supplied 512 labeled datasets, the Nave Bayes method beats expectation-maximization with an average accuracy gain of 8% for 64 labeled documents and a 10% increment for 512 labeled documents.
5	Ruben Dorado, Sylvie Ratte [11]	latent Dirichlet allocation approach	Latent Dirichit allocation and naïve bayes approach shows that an accuracy of 80% can be achieved with 3% of 600 examples datasets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

6	Chunyong Yin, Hui Zhang, Jin Wang, Jun Xiang, [12]	SVM categorizing to reduce the data size, semi-supervised learning algorithm, KNN algorithm	Semi-supervised learning algorithm shows the more accuracy when compared to KNN algorithm
7	Vo Duy Thanh, Pham Minh Tuan, Vo Trung Hung, Doan Van Ban [13]	Semi- supervised learning and SVM, self-training algorithm	When the size of labelled and unlabeled data is compared with the change in accuracy on learning technique and labelled data, both situations demonstrate an increase in the size of labelled and unlabeled data.
8	Xin Pan, Suli Zhang [14]	Fuzzy C-means algorithm for improving the effectiveness of the classifier	When Fuzzy c-means is enhanced Semi-supervised Fuzzy c-means Text classifier (SFCTC) is found. Comparing SFTC, KNN and Naïve Bayes results to certain percent of accuracies where SFCTC shows the highest accuracy compared to KNN and Naïve Bayes.
9	Karl Severin, Swapna S. Gokhale, Aldo Dagnino [15]	Precision, Recall and F1 Score are the techniques for performance measure,	Performance measure techniques are used to categorize the accounts, every semi-supervised learning works when there is small amount of labeled data available, so the small amount of data is made use
10	Shweta C. Dharmadhikari, Maya Ingle, Parag Kulkarni [16]	Multi-label Classification approach	When using semi-supervised learning in this multi-label text classification, the classifier's decision-making capacity improves.

CONCLUSION

This paper unveils a review on semi-supervised classification technique based on previous knowledge of class relevant terms. For categorization learning, the training set does not need to be supplied ahead of time. Class related words are selected for classification that can represent the subjected to terms related to class values and play a significant part in the algorithm. They must be chosen with the help of a user who has prior understanding of the subjects covered in class.

[10] The author shows that using the expectation-maximization algorithm for semi-supervised learning will not improve the area under the curve of naive bayes but using the semi-supervised method called semi-supervised frequency estimate (SFE) on different datasets SFE significantly and consistently improves the area under the curve and accuracy of naive bayes, as well as producing better conditional log likelihood values than the expectation-maximization algorithm.

Furthermore, our study and actual findings reveal that frequency estimate has a lower computing cost than EM+NB, making it the superior alternative when dealing with big unlabeled datasets.

REFERENCES

[1] C.Rosenberg, M. Hebert and H. Schneiderman, “Semi-Supervised Self-training of Object Detection Models,” 2005 Seventh IEEE Workshops on Applications of computer Vision (WACV/MOTION’05) – Volume 1, 2005, pp. 29-36, Doi: 10.1109/ACVMOT.2005.107.

[2] K. Wang, B. Zhan, Y. Luo, J. Zhou, X. Wu and Y. Wang, “multi-tasking Curriculum Learning For semi-supervised medical Image Segmentation,” 2021 IEE 18th International Symposium on Biomedical Imaging.

[3] J. Serra, J. Pons and S. pascual, “SESQA: Semi-Supervised learning for speech quality assessment,” ICASSP 2021 – 2021 IEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp.381-385, doi:10.1109/ICASSP39728.2021.9414052. (ISBI), 2021, pp.925-928, Doi: 10.1109/ISBI48211.2021.9433851.

[4] Stoica, A. S., heras, S., Palanca, J., Julian, V., & Mihaescu, M. c. (2021). Classification of educational videos by using a semi-supervised learning method on transcripts and keywords. Neurocomputing. Doi: 10.1016/i.neucom.2020.11.07

[5] C. Liu, W. Hsaio, C. Lee, T. Chang, and T. Kua, “Semi-supervised text Classification with Universum learning,” in IEEE transactions on Cybernetics, vol. 46, no. 2, pp. 462-473, Feb. 2016, Doi: 10.1109/TCYB.2015.2015.2403573.

[6] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. "Text classification from labeled and unlabeled documents using EM", Machine learning, Vol. 39(2/3): ppl03-134, 2000.

[7] HAN Hong_qil, ZHU Dong-Hua, WANG Xue-feng, “Semi-supervised Text Classification from Unlabeled Documents Using Class Associated Words”, in International Conference on Computers and Industrial engineering, DOI: 10.1109/ICCIE.2009.5223918.

[8] Mohammad Abdul wajeeb, T. Adi Lakshmi, “Semi-Supervised text classification using Enhanced KNN”, DOI: 10.1109/WCIT.2011.6141232.

[9] Alexander Hanbo Li, Abhinav Sethy, “Semi-Supervised Learning for Text Classification for Text Classification by Layer Partitioning”, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), DOI: 10.1109/ICASSP40776.2020.9053565.

[10] Jiang Su, Jelber Sayyad-Shirbad, Stan Matwin, “Large Scale Text Classification Using Semi-Supervised Multinomial Naïve Bayes”, Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011,

[11] Ruben Dorado, Sylvie Ratte, “Semi-Supervised Text Classification Using Unsupervised Topic Information”, DOI: 10.7551/mitpress/9780262033589.003.0003.

[12] Chunyong Yin, Jun Xiang, Hui Zhang, Jin Wang, “A new SVM method for short text classification based on semi-supervised learning”, IEEE in 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), DOI: 10.1109/AITS.2015.34

[13] Vo Duy Thanh, Vo Trung Hung, Pham Minh Tuan, Doan Van Ban, “Text classification based on semi-supervised learning”, IEEE in 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), DIO: 10.1109/AITS.2015.34

[14] Xin Pan, Suli Zhang, “Semi-supervised Fuzzy learning in Text categorization”, in IEEE 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), DOI: 10.1109/FSKD.2011.6019630

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

[15] Karl Severin, Swapna S. Gokhale, Aldo Dagnino, "Key-word based semi-supervised Text classification", in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), DOI: 10.1109/COMPSAC.2019.00067

[16] Shweta C. Dharmadhikari, Maya Ingle, Parag Kulkarni, "Semi-supervised learning based text classification model for multi label paradigm", DOI: https://doi.org/10.1007/978-3-319-11629-7_26