Text Classification of Data Using Deep Learning Technique

Amandeep Kaur[1], Abhinash Singla[2]
[1, 2] Dept of CSE
[1, 2] BGIET, Sangrur

Abstract- *Text classification also known as text tagging or text categorization is the process of categorizing text into organized group. By using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content. Deep Neural Network (DNN) has been used for text classification. Literature survey on deep learning techniques has been presented*

Keywords- Deep Learning, Neural Network, Machine learning

I. INTRODUCTION

Various text classification techniques were initially identified through Wikipedia and other encyclopedias and corroborated with the content of various research articles. The major approaches were further arranged as a tree structure after analyzing the similarities and differences among these various approaches along with their respective algorithms. Generally, a classification technique could be divided into statistical and machine learning (ML) approaches. Statistical techniques purely satisfy the proclaimed hypotheses manually, therefore the need for algorithms is little, but ML techniques were specially invented for automation.

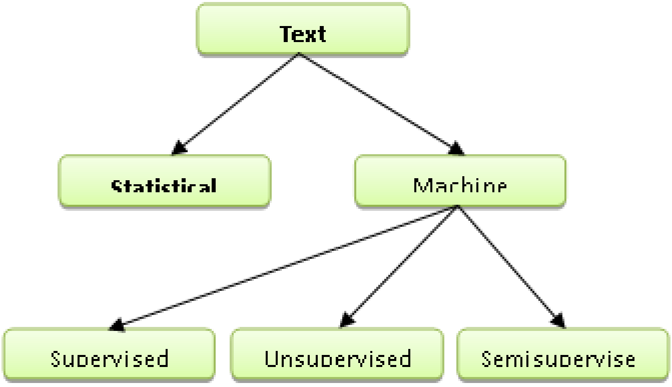

Fig 1: Shows text classification techniques

1.1 Statistical Approach

Statistical techniques are purely mathematical processes, and they act as the mathematical foundation for all other text classifiers. It works similar to a computer program, executing the given instructions without any ability of its own.

1.2 Machine Learning:

The increase in data volume, velocity, and variety called for automation in text processing techniques including text classification. In some situations, defining a set of logical rules using knowledge-engineering techniques and based on expert opinions to classify documents helps to automate the classification task. Text classification could be divided into three categories: supervised text classification, unsupervised text classification, and semi-supervised text classification based on the learning principle followed by the data model. In machine learning terminology, the classification problem comes under the supervised learning principle, where the system is trained and tested on the knowledge about classes before the actual classification process. Unsupervised learning occurs when labeled data is not accessible. The process is complicated and has performance issues.

i. *Supervised Learning:* Supervised learning is the most expensive and highly difficult of the three. It requires a human intervention while assigning labels to classes which is not possible in large datasets. Though the work flow mimics the techniques followed in AI processes, it is time consuming. It is also called inductive learning in ML.

ii. *Unsupervised learning:* Unsupervised learning is a type of ML algorithm where, inferences are drawn from the data by clustering data into different clusters without labeled responses i.e. expected outcomes. In other words, no training data is provided to the system. It appears complex initially, but when more data is fed into the model, the algorithm refines itself to efficiency.

iii. *Semi-supervised learning:* Semi-supervised learning is a combination of supervised and unsupervised learning techniques. This type of learning employs small amount of labeled data and large amount of unlabeled data for training. Some of the SSL techniques are such as self-training or self-teaching or bootstrapping, co-training, transductive SVMs, generative models and graph-based methods. The application of semi-supervised algorithms is highly useful in information filtering requirements.

II. DEEP LEARNING

A popular way to approach different tasks in NLP is to use a Long Short-Term Memory Recurrent Neural Network. The strength of the LSTM is that it can capture information in any part of the document. It also allows the model to account for the specific order of words which. Another interesting approach is to use a Convolution Neural Network (CNN) on a character level. Since CNNs are the current state-of-the-art in image recognition it has been suggested and shown that they can be successful in various NLP tasks. The idea of character level CNNs by using up to 29 convolution layers with promising results is also presented. A combination of RNN and CNN has been tried and shown to work well for text classification with few classes. They use a pre-trained language model and then show that it can be fine-tuned with small amounts of data to perform well on a range of different tasks. This is the technique that will be used in this thesis since it has shown to be very successful on other text classification tasks.

III. NEURAL NETWORK

Artificial neural networks (ANNs) work in the same way as human brain in arriving at a decision. Swarm intelligence and evolution algorithm are used to generalize a neural network model. It works on the virtue of learning and evolution with minimal or no human intervention. For data classification, competitive co-evolution algorithm based neural network model is suggested. Radial Basis Function is the ANN component as it employs faster learning algorithms. It has a compact network architecture that increases classification accuracy. Also, evolutionary algorithms have a tendency to perform well in dynamic environments by learning rules on the fly and highly adaptive of ‘fuzzy’ characteristics. Neural networks are also popular among cases where a hierarchical multi-label classification approach is required. This kind of classification is complex as each sample may belong to more than one class and predictions of one level is fed as inputs to next level to make a final decision. Also in a similar setup, linear regression could be used for feature selection in an ensemble boosted classifier [1]. Neural network forms the base of the ensemble with the help of composite stumps. The ANNs have good application value, development potential and it is also not necessary to train the individual binary classifiers for multi class problems therefore they form better base classifiers in an ensemble approach. Further, over fitting is taken care of by Adaboost and accuracy is maintained through ANNs [2].

IV. LITERATURE REVIEW

M. Amajd et al. (1989) analyzed the use of different neural networks for the text classification task. The accuracy of the studied text classifiers could be changed by a small number of previously classified texts. The convolution neural network could work better at the level of words, and does not require knowledge of the syntactic or semantic structure of the language. On the other hand, a recurrent neural network for the level of data representation in the form of a sequence can effectively classify the text. The presented results obtained for text corpora from two different sources show that using a vector data representation can also improve the accuracy of the classification [3].

JingjingCai et al. (2018) discussed Common text classification applications include spam identification, news text classification, information retrieval, emotion analysis, and intention judgment. Traditional text classifiers based on machine learning methods have defects such as data sparsity, dimension explosion and poor generalization ability, while classifiers based on deep learning network greatly improve these defects, avoid cumbersome feature extraction process, and have strong learning ability and higher prediction accuracy. Such as convolution neural network (CNN) and introduced the process of text classification and focuses on the deep learning model used in text classification [4].

P. Parvathi and T. S. Jyothis (2018) worked on identifying relevant text from text document and found it very crucial. Text classification is the task of relegating a document under a predefined category. There are several methods to identify which words in text documents are important to explain the category it is associated with. They proposed approach uses convolution neural network with deep learning. And the deep learning is used to predict the categories accurately and concluded that calculating the test's accuracy by F1 Score, we get an accuracy value which is approximately equal to 1 [5].

F. Wei et al. (2018) reported the preliminary studies in using deep learning in legal document review. Specifically, conducted experiments to compare deep learning results with results obtained using a SVM algorithm on the four datasets of real legal matters. Our results showed that CNN performed better with larger volume of training dataset and should be a fit method in the text classification in legal industry [6].

E. Phaisangittisagul et al. (2019) presented work on filtering of promotional advertising is an essential part to detect improper information before posting on the websites and social media. A model to classify promotional advertising is proposed to identify whether relevant promotion content for a specific business or service in order to meet precise customers' attention. The proposed algorithm in the work is based on deep learning is designed to handle promotional image and message in competition with the 2nd KU Data Science Boot Camp 2018. The performance is evaluated on the promotional advertising data provided by Wongnai. Finally, the accuracy of the proposed method could be achieved satisfactory performance of 82.95% in testing data [7].

V. CONCLUSION

In this article, we have addressed the importance of deep learning in text classification, different types of deep learning techniques has been discussed in brief. Further on, how the neural network plays an important role in deep learning has been discussed. An extensive literature review on deep learning techniques has been presented.

REFRENCES

[1] Hiew, B. Y., Tan, S. C., & Lim, W. S., Intra-specific competitive co-evolutionary artificial neural network for data classification. Neurocomputing, 185, 220–230, 2016.

[2] Nie, Q., Jin, L., Fei, S., & Ma, J., Neural network for multi-class classification by boosting composite stumps. Neurocomputing, 149, 949–956, 2015.

[3] M. Amajd, Z. Kaimuldenov and I. Voronkov, "Text classification with deep neural networks", Proc. CEUR Workshop, vol. 17, pp. 362-370, 1989.

[4] J. Cai, J. Li, W. Li and J. Wang, "Deep learning Model Used in Text Classification," 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, pp. 123-126, doi: 10.1109/ICCWAMTIP.2018.8632592, 2018.

[5] P. Parvathi and T. S. Jyothis, "Identifying Relevant Text from Text Document Using Deep Learning," 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, pp. 1-4, doi: 10.1109/ICCSDET.2018.8821192, 2018.

[6] F. Wei, H. Qin, S. Ye and H. Zhao, "Empirical Study of Deep Learning for Text Classification in Legal Document Review," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 3317-3320, doi: 10.1109/BigData.2018.8622157, 2018.

[7] E. Phaisangittisagul, Y. Koobkrabee, K. Wirojborisuth, T. Ratanasrimetha and S. Aummaro, "Target Advertising Classification using Combination of Deep Learning and Text model," 2019 10th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), Bangkok, Thailand,pp. 1-4, doi: 10.1109/ICTEmSys.2019.8695956, 2019.