National Research Journal of Information Technology & Information Science
Volume No: 13, (January) Year: 2026 (Special Issue)
PP: 640-644
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

CLASSIFICATION OF TEXT DATA USING A DEEP LEARNING LONG SHORT-TERM MEMORY NETWORK

Peerzada Moien Ahmad

Gulzarpora Awantipora, Jammu and Kashmir India

Amreen Kaur

Punia Colony Sangrur

Yadwinder singh

Guru Nanak Colony Sangrur

ABSTRACT

The presented study explores various signal classification techniques, systematically categorizing them into statistical and machine learning (ML) approaches. While statistical methods operate on predefined mathematical foundations requiring minimal algorithmic support, ML techniques were developed to address automation demands in signal processing. ML-based classification is further sub-divided into supervised, unsupervised, and semi-supervised learning models, each characterized by unique methodologies and challenges. Supervised learning, though effective, is resource-intensive due to its reliance on labeled data. Unsupervised techniques leverage clustering and self-organizing maps to overcome labeling constraints. Semi-supervised approaches combine the strengths of both, significantly enhancing classifier accuracy with minimal labeling costs. In this work, a Long Short-Term Memory (LSTM) neural network is proposed for text classification tasks, employing word embedding layers to capture semantic relationships. The developed LSTM model demonstrates 100% accuracy in classifying testing data, highlighting its efficacy for high-dimensional signal classification problems.

Keyword: Signal Classification Techniques, Machine Learning Approaches, Statistical Methods, Long Short-Term Memory (LSTM), Supervised and Unsupervised Learning

1. INTRODUCTION

Various signal 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 [6].

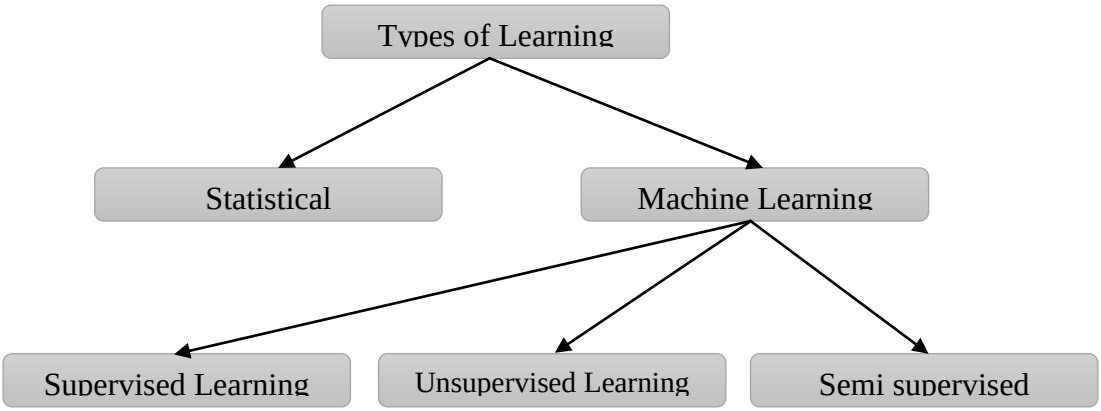

Figure 1 Shows different machine learning techniques

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

1.2 Statistical Approach

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

2. MACHINE LEARNING APPROACH

The increase in data volume, velocity, and variety called for automation in signal 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 classifica-tion, unsupervised text classification, and semi-supervised text classification based on the learning principle followed by the data model [7]. 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.

2.1 Supervised learning

Supervised learning is the most expensive and highly difficult of the three. The main reason behind this notion is that it requires a human intervention while assigning labels to classes which is not pos-sible 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 [8]. Supervised learning becomes expensive when different data distributions, different outputs and different feature spaces occur as in heterogeneous text corpora. One of the most widely used supervised methods is maximum likelihood estimation

2.2 Unsupervised learning

Unsupervised learning is a type of ML algorithm where, inferences are drawn from the data by clus-tering 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. Principal component analysis, clustering and self-organizing maps are frequently used in unsupervised learning. In many scenarios clustering is the same as unsupervised learning. Many times, expert knowledge required to label the samples is either non-existent or inadequate. In such case, self-organizing maps and correlation coefficient are used to cluster the documents and use it to label the documents for further classification [9]. It eliminates the curse of dimensionality and expert intervention as well. This kind of hybrid model is more suitable for high volume data.

1.4.3 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 train-ing. The labels are assigned by combining labeled and unlabeled instances, as unlabeled data mitigate the effect of insufficient labeled data on classifier accuracy. 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. Vector space models are mostly used in language processing problems to address natural language semantics that supposes words in similar contexts have similar meanings. Meaning values are calcu-lated according to the Helmholtz principle. This model is non-iterative but effective in augmenting the efficiency of classifier. The system can be combined with semantic kernels that smooth docu-ment term vectors using term to term semantic relations. Finding out more approaches to extract the information from the context of a class could be tried in the future. Traditional text classification approaches become null when there is no labeled data for a particular class of the dataset, for example, the labeled data is only available for positive samples and not for negative samples. A semi-supervised algorithm based on tolerance roughest and ensemble learning is recommended. The unavailable class is extracted approximately from the dataset and set as the labeled sample. The ensemble classifier iteratively builds the margin between positive and negative classes to further approximate negative data, since negative data is mixed with the positive data. Therefore, without the need for training samples, classi-fication is achieved through a hybrid approach. It eliminates the cost of hand labeling data, especially in

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

big data. The application of semi-supervised algorithms is highly useful in information filtering requirements. The role of semi-supervised algorithms in multi-label hierarchical classifi-cation is an area where there is still a need for more exploration. Self-training along with semi-supervised classifier is recommended for multi-label hierarchical classification.

3. PROPOSED WORK

A long short term memory (LSTM) network is a type of recurrent neural network (RNN) that have long term dependencies between time steps of sequential data. In the presented work, LSTM neural network has been used to learn and use long term dependencies fro classification of text data. To input is a text file applied to an LSTM network. First the text data has been converted into numeric sequences. This has been achieved by using a word encoding which maps documents to sequences of numeric indices. To achieve better results a word embedding layer has been included in the network. Word embeddings has been used to map words in a vocabulary to numeric vectors rather than scalar indices. These embeddings have captured semantic details of the words and will result in generation of words which have similar meanings to have similar vectors. This has also helped into modeling relationships between words through vector arithmetic.

To input the documents into an LSTM network, word encoding has been used to convert the documents into sequences of numeric indices. Then the documents have been padded and truncate to make them of same length. To pad and truncate the documents, first a target length has been chossen, and then truncate the documents that are longer than it and left-pad documents that are shorter than it. For best results, the target length should be short without discarding large amounts of data. To find a suitable target length, a histogram of the training document lengths shown in Fig 2 is simulated. As most of the training documents have fewer than 11 tokens. So, we have used 11 as the target length for truncation and padding.

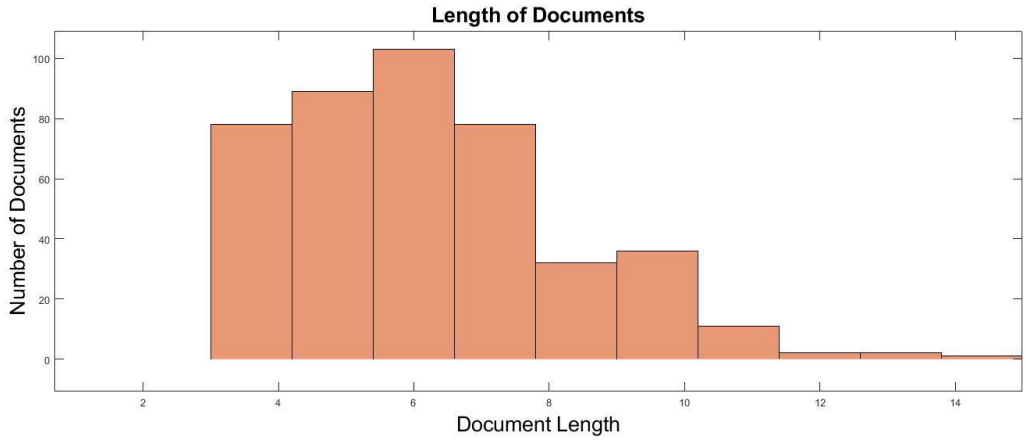

Figure 2 Histogram to Decide for Target Length

The final and most important step is to develop the LSTM network architecture. To input sequence data into the network we have used an input layer size of 1. The size of second layer of neurons has been taken as 70. This layer is a word embedding layer and has the same number of words as the word encoding. Then the number of hidden neurons has been set as 100. Lastly, to use the LSTM layer for a sequence to label classification problem the output mode has been set as “last”.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

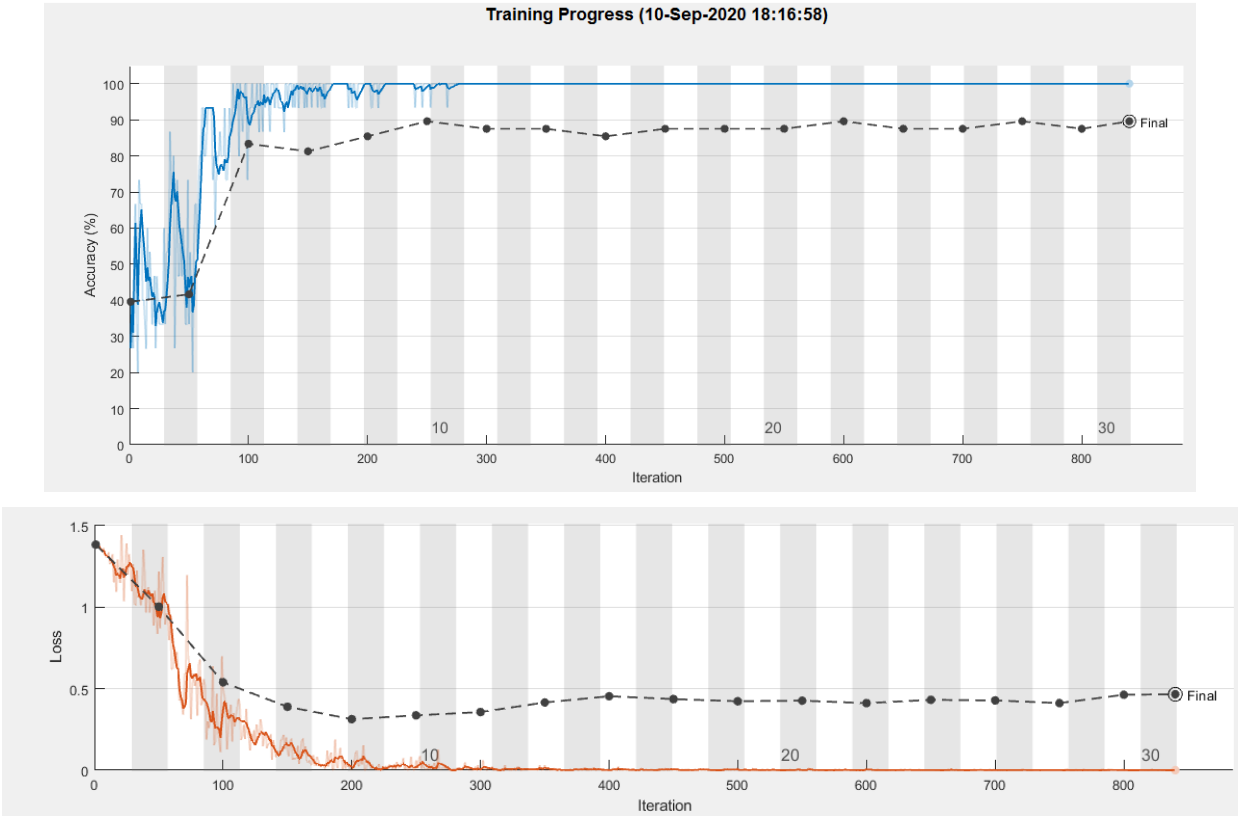

Figure 3 LSTM Network Training

To predict and testing the accuracy of the trained LSTM network, a test data has been used with the event type classified into three new reports. The string array containing the new reports has been created as shown:

"Coolant is pooling underneath sorter."

"Sorter blows fuses at start up."

"There are some very loud rattling sounds coming from the assembler."];

The text data has been preprocessed using the preprocessing steps as done during training the documents. The text data has been converted into sequences with the same options as done during the training sequences process. Then the new sequences have been classified using the trained LSTM network. The following classifications have been obtained from the testing data:

- Leak
- Electronic Failure
- Mechanical Failure

From the classification results it has been concluded that LSTM network has classified the tesing data with 100% accuracy.

This study explores various signal classification techniques, systematically categorizing them into statistical and machine learning (ML) approaches. While statistical methods operate on predefined mathematical foundations requiring minimal algorithmic support, ML techniques were developed to address automation demands in signal processing. ML-based classification is further sub-divided into supervised, unsupervised, and semi-supervised learning models, each characterized by unique methodologies and challenges. Supervised learning, though effective, is resource-intensive due to its reliance on labeled data. Unsupervised techniques leverage clustering and self-organizing maps to overcome labeling constraints. Semi-supervised approaches combine the strengths of both, significantly enhancing classifier accuracy with minimal labeling costs. In this work, a Long Short-Term Memory (LSTM) neural network is proposed for text classification tasks, employing word embedding

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

layers to capture semantic relationships. The developed LSTM model demonstrates 100% accuracy in classifying testing data, highlighting its efficacy for high-dimensional signal classification problems.

4. CONCLUSION

The study demonstrates that signal classification methods span diverse techniques, from traditional statistical approaches to advanced ML-based solutions. Statistical methods serve as a foundation but lack automation capabilities. ML approaches, leveraging supervised, unsupervised, and semi-supervised models, address automation challenges while enhancing classification accuracy. The proposed LSTM-based model, incorporating word embeddings and optimized processing techniques, effectively captures semantic details and dependencies in sequential data. Experimental results affirm its potential in achieving high accuracy, even with complex data sets. The findings underscore the significance of ML in advancing classification technologies and open avenues for further research in semi-supervised learning and multi-label hierarchical classification.

REFERENCES

1. S. Haykin, "Cognitive radio: brain-empowered wireless communications," IEEE Journal on Selected Areas in Communications, vol. 23, no. 2, pp. 201-220, Feb. 2005.
2. S. Subray, S. Tschimben and K. Gifford, "Towards Enhancing Spectrum Sensing: Signal Classification Using Autoencoders," IEEE Access, vol. 9, pp. 82288-82299, 2021.
3. A. Fehske, J. Gaeddert and J. H. Reed, "A new approach to signal classification using spectral correlation and neural networks," First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), Baltimore, MD, USA, pp. 144-150, 2005
4. J. Mitola, "Cognitive Radio Architecture Evolution," Proceedings of the IEEE, vol. 97, no. 4, pp. 626-641, April 2009.
5. X. Zhu, Y. Lin and Z. Dou, "Automatic recognition of communication signal modulation based on neural network," 2016 IEEE International Conference on Electronic Information and Communication Technology (ICEICT), Harbin, China, pp. 223-226, 2016.
6. S. Rajendran, W. Meert, D. Giustiniano, V. Lenders and S. Pollin, "Deep Learning Models for Wireless Signal Classification With Distributed Low-Cost Spectrum Sensors,"IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 3, pp. 433-445, Sept. 2018.
7. M. Bkassiny, S. K. Jayaweera and Y. Li, "Multidimensional Dirichlet Process-Based Non-Parametric Signal Classification for Autonomous Self-Learning Cognitive Radios," IEEE Transactions on Wireless Communications, vol. 12, no. 11, pp. 5413-5423, Nov. 2013.
8. M. Bkassiny, Y. Li and S. K. Jayaweera, "A Survey on Machine-Learning Techniques in Cognitive Radios," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136-1159, Third Quarter 2013.
9. A. Upadhye, P. Saravanan, S. S. Chandra and S. Gurugopinath, "A Survey on Machine Learning Algorithms for Applications in Cognitive Radio Networks," 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, pp. 01-06, 2021.
10. L. Yu, J. Chen and G. Ding, "Spectrum prediction via long short term memory," 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, pp. 643-647, 2017.