Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

Machine Learning for Natural Language Processing: Techniques and Evaluation

Piyush Raja[1], Dr. Santosh Kumar[2*], Digvijay Singh Yadav[3], Dr. Amit Kumar[4], Ram Krishna Kumar[5]

[1,3]Assistant Professor, Department of CSE, COER University, Roorkee, Uttarakhand, India
[2]Assistant Professor, PSIT College of Higher Education, Kanpur, India*
[4]Assistant Professor, Department of CS, Gaya College Gaya, Bihar, India
[5]Research Scholar, Department of CS, Magadh University, Bodh Gaya, Bihar, India

Corresponding Email: [2]yajisantosh.sk@gmail.com*

Received: 02 December 2022 **Accepted:** 28 February 2023 **Published:** 03 April 2023

Abstract: ***The purpose of this research study is to analyse the approaches that are used in machine learning for the solution of natural language processing (NLP) issues. Also, the assessment criteria that are used in the performance analysis of these models are looked at as part of this investigation. We investigate the challenges that come with using machine learning for natural language processing (NLP), including the issue of data bias and the need that these models have the capacity to be explained. Research is conducted into a variety of supervised and unsupervised learning approaches, including as neural networks, topic modelling, and clustering. Also, a number of evaluation metrics, such as accuracy, precision, recall, and F1 score, are discussed. As the research comes to a conclusion, it is highlighted how important it is to find solutions to problems like data bias and the inability to explain results. This is done in order to ensure that the results generated by machine learning models for natural language processing are accurate and unbiased.***

Keywords: NLP, CNN, RNN, Clustering, Topic Modelling, Machine Learning.

1. INTRODUCTION

Machine learning is becoming an increasingly essential technology among the numerous applications of natural language processing (NLP), which has many benefits. These applications include anything from study of feelings to translation of languages. In this research study, we will investigate the many approaches that are used in machine learning for natural language processing (NLP) problems, as well as the evaluation metrics that are utilised to evaluate the performance of these models. In addition, we will investigate the many evaluation metrics that are utilized to evaluate the performance of these models. We will also discuss the challenges that come with using machine learning for natural language

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

processing (NLP), such as the issue of data bias and the need for explain ability in the models that are used (Fig. 1).

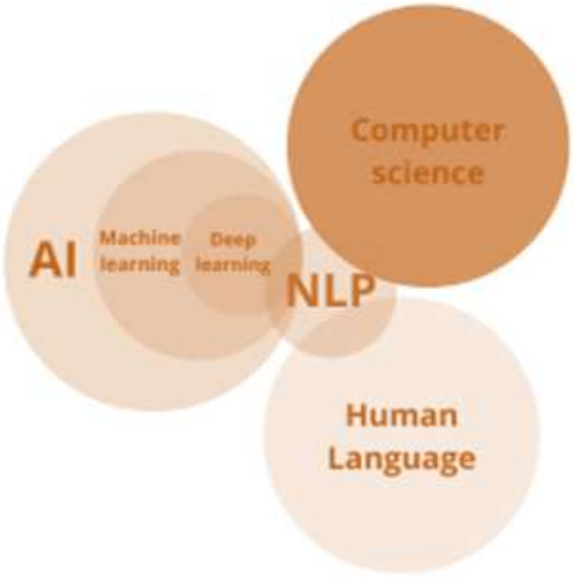

Figure 1: The comprehension of human language by robots is made possible through NLP.

Techniques for using machine learning in natural language processing
In machine learning for natural language processing, both supervised and unsupervised learning strategies may be used as learning methodologies. Unsupervised learning algorithms are able to learn from data that has not been labelled, in contrast to supervised learning algorithms, which need access to data that has been labelled in order to train a model.

A prominent kind of supervised learning that is used in natural language processing (NLP) is the application of neural networks. Some examples of neural networks that are employed in NLP include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are often employed for text classification tasks like sentiment analysis, whilst RNNs are utilised for sequence-to-sequence tasks including language translation. Another prominent strategy for natural language processing is known as unsupervised learning. This kind of learning involves clustering as well as topic modelling (NLP). Although clustering algorithms can group documents that are similar together, topic modelling can uncover the underlying topics in a collection of documents. [C]lustering techniques can also group documents that are similar. Both methods are capable of doing analysis on substantial volumes of text.

CNN
Convolutional neural networks, also known as CNNs, are a type of neural network that are frequently used in computer vision applications such as the categorization of images, the identification of objects, and the segmentation of images. CNNs are also referred to by their

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

acronym, which stands for "convolutional neural networks." On the other side, CNNs have also been successfully used in natural language processing (NLP) applications, such as text classification, sentiment analysis, and named entity recognition, amongst others.

CNNs are able to accomplish their aim of automatically finding hierarchical representations of the data that they are provided with because they submit the input to a series of convolutional filters in succession. The filters, which are typically rather small in size, slide over the input while carrying out element-wise multiplications and summations in order to extract local features. The goal of this process is to find patterns that are unique to the area. The created feature maps are then routed via a non-linear activation function such as the rectified linear unit (ReLU), which is a typical approach, in order to generate non-linearity. This is done so that the non-linearity may be formed.

Convolutional neural networks, often known as CNNs, are used in natural language processing. These networks may be trained to acquire sentence representations by thinking of phrases as strings of words. Word embedding, such as Word2Vec or GloVe, are used, as is standard procedure, in order to convert the words into dense vector representations. This process is called "word embedding." After that, the sequence of word vectors that was formed is fed into a convolutional neural network, also known as a CNN. Inside the CNN, convolutional filters slide along the sequence to extract local properties.

One of the numerous advantages of adopting CNNs in natural language processing is the capacity of CNNs to capture local features such as n-grams, which are vital for tasks such as sentiment analysis. This is only one of the many benefits of utilising CNNs (NLP). In addition, CNNs are easily capable of being parallelized, which enables them to analyse enormous amounts of text input in a timely manner.

On the other hand, CNNs are subject to a number of limitations in terms of NLP. For instance, they are not as effective as recurrent neural networks (RNNs) at capturing long-term connections, which may be essential for tasks such as machine translation. This is because RNNs are designed to learn from previous data. RNNs are a subclass of neural networks that were first created by IBM. Moreover, convolutional neural networks (CNNs) may be sensitive to the choice of key parameters, such as the number and size of filters, and may need additional tuning in comparison to more straightforward modelling techniques.

In general, convolutional neural networks (CNNs) are a useful tool for natural language processing (NLP) applications, and they have achieved state-of-the-art performance on a variety of different benchmarks (shown in figure 2). Nevertheless, their utility is dependent on the characteristics of the data as well as the nature of the task at hand. In order to determine how well they perform in relation to other models, it is necessary to analyse them alongside other models.

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

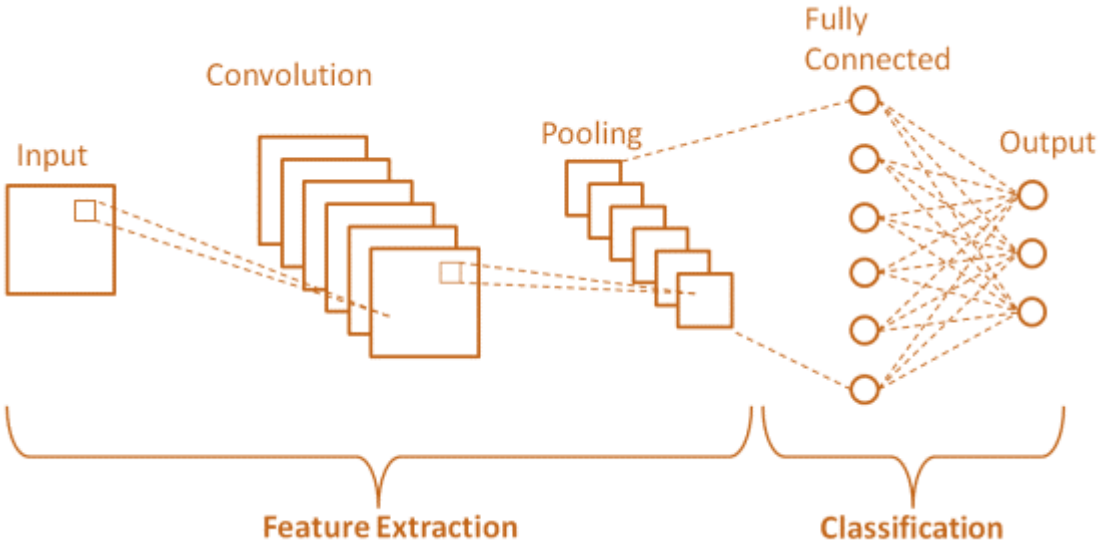

Figure 2: Basic architecture of CNN

RNN

In the area of natural language processing (NLP), recurrent neural networks (commonly abbreviated as RNNs) are a particular kind of neural network that is often employed in a variety of tasks. RNN is an abbreviation that stands for "recurrent neural network." RNNs, in contrast to feed forward neural networks, are able to process sequences of varying lengths, which distinguishes them from the latter. Feed forward neural networks only deal with inputs of a fixed length. As a consequence of this, the applications of language modelling, machine translation, and speech recognition are all ideally suited for use with these networks.

RNNs are constructed to imitate sequential dependencies by maintaining a hidden state vector that is updated at each time step. This allows the RNN to improve its performance over time. Because of this, the RNN is able to gain knowledge from its previous encounters. When the hidden state is first created, it is typically initialised to a vector of zeros. The hidden state is then updated by first applying a recurrent weight matrix to the input and the previous hidden state, and then passing the result through a non-linear activation function such as the hyperbolic tangent or the sigmoid function. When the hidden state is first created, it is typically initialised to a vector of zeros. When it is initially formed, the hidden state is often initialised with a vector of zeros as its value. Following that, the newly constructed hidden state vector is used in either the process of creating a prediction or the generation of the subsequent output in the sequence, depending on which of the two is being carried out.

One of the most important advantages of using these models is that RNNs are able to identify long-term correlations in sequences, which is one of the benefits of utilising these models. This is because the hidden state vector is able to keep information on the whole history of the sequence inside its memory. This has led to the current situation. This gives the network the capacity to generate predictions based on the context of events that occurred a very large number of time steps in the past. In addition, RNNs may be taught from scratch using a

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

technique called back propagation through time, which teaches them how to optimise the representation of their hidden states in line with the task at hand that is being carried out.

On the other hand, RNNs are subject to a variety of other issues. It may be challenging for them to acquire long-term dependencies due to the vanishing gradient issue, which occurs when the gradients that are used to update the weights become extremely small as they propagate back through time. This problem can make it challenging for them to acquire long-term dependencies. One of the challenges that you can have when attempting to teach them is dealing with this issue. There have been many other types of RNNs proposed as potential solutions to this issue. Such examples are long short-term memory (LSTM) networks and gated recurrent units (GRU) networks. In order to more successfully maintain gradient flow throughout longer periods, these networks make use of a variety of mechanisms.

There are a number of limitations associated with RNNs, one of which is that the process of training them may need a significant amount of computer power. This is true, in particular, for very extended sequences. This is because the hidden state vector has to be updated at each and every single time step, and the network needs to be unrolled throughout the whole sequence while it is being trained. Both of these steps are necessary when the network is being trained. When trained on restricted datasets, RNNs have a greater risk of falling victim to overfitting, which may lead to poor generalization performance.

In general, recurrent neural networks, also known as RNNs, are an effective tool for natural language processing (NLP) applications, and their performance on a variety of benchmarks has achieved the state of the art. Nevertheless, their utility is dependent on the characteristics of the data as well as the nature of the task at hand. In order to determine how well they perform in relation to other models, it is necessary to analyse them alongside other models. In addition, researchers are continually researching into new designs and training techniques for RNNs in an attempt to overcome the limitations of these models and raise their overall performance. This is being done in an effort to improve RNNs.

Cluster Modeling

Cluster Modeling, which is sometimes referred to as clustering, is a technique that is used in the area of machine learning to group data points that are similar to one another on the basis of the features that they have. Clustering is a method that is often used in the processes associated with unsupervised learning. While working on projects of this kind, the goal is to identify patterns or structures within the data without making use of labeled examples.

The K-means clustering method is a common approach that may be used for the process of grouping data; however, this is not the only method for clustering data; there are other alternative approaches. The purpose of the k-means clustering approach is to ultimately achieve the aim of iteratively segmenting the data into k clusters. It is possible to do this by finding a way to minimize the sum of squared distances between each data point and the centroid of the cluster to which it is geographically closest. The method begins by initially randomizing k centroids, and it then alternates between assigning each data point to the

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

centroid that is geographically closest to it and then updating the centroids based on the mean of the data points that are geographically closest to each cluster. Finally, the method concludes by assigning each data point to the centroid that is geographically farthest from it. The procedure is finished either when the cluster assignments are no longer being reshuffled or when a maximum number of iterations through the algorithm has been reached, depending on which of these two conditions is met first.

Clustering may also be done in a hierarchical fashion, which is a common approach. This approach depicts the degree of similarity between the data points by building a tree-like structure, which is referred to as a dendrogram. This structure is constructed using this method. While doing agglomerative hierarchical clustering, the algorithm first treats each data point as its own cluster before moving on to the next data point. After that, the method iteratively combines the two clusters that are geographically nearest to one another until all of the data points are a part of a single cluster. This process continues until all of the data points have been incorporated. It is common practice to employ a linkage criterion, such as full linkage, average linkage, or Ward's approach, when attempting to ascertain the distance that exists between clusters. Several distinct approaches are possible for achieving this goal.

Clustering has the potential to be used in a wide variety of contexts, such as the classification of customers, the detection of anomalies, and the segmentation of images. Clustering is a method that is used in natural language processing (NLP), and it can be used to group together phrases or documents that are similar based on their content. Clustering is a technique that can be used to group together phrases or documents that are similar. For instance, clustering may be used to gather news items that are linked to one another thematically, or it can be used to detect reviews or comments that are similar based on the tone that they communicate. Both of these applications are examples of how clustering can be used.

However, clustering does not come without its own set of challenges and limitations. When the data is high-dimensional or noisy, one of the issues is that it may be difficult to determine the appropriate number of clusters to use. This is one of the challenges that may be faced. This is especially important to keep in mind while dealing with complicated data. In addition to this, clustering may be sensitive to the choice of distance metric or linkage criteria, and it may need more tweaking than models with fewer moving components in order to provide accurate results. Due to the fact that clustering is a kind of unsupervised learning, it is likely that the generated clusters may be difficult to interpret without additional domain knowledge or further validation from the outside.

Clustering is a helpful way for revealing hidden patterns in data and has different applications in natural language processing as well as other domains. In general, clustering is a useful method for exposing hidden patterns in data. Yet, it is very necessary to pay careful thought to the specific problem at hand and the data that are available, as well as to evaluate the efficacy of clustering models by using the appropriate metrics and validation methods.

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

Topic Modelling

The discipline of topic modelling is a strategy that is used in the domains of natural language processing and machine learning. This approach is utilized as a way to locate topics or themes that are included within a massive corpus of text data. It comprises doing an examination of the co-occurrence patterns of terms within the text data and then clustering those words into groups based on the semantic similarity of the phrases in each category. This is followed by presenting the results of the analysis.

While the Latent Dirichlet Allocation approach is the one that is applied in topic modelling the majority of the time, there are a variety of alternative algorithms that may also be used in this process (LDA). The Latent Dirichlet Allocation (LDA) model is a generative probabilistic model that operates on the premise that each document in the corpus is a mixture of themes, and that each topic is a probability distribution over individual words. The Linguistic Data Analysis group is responsible for the creation of LDA.

The LDA algorithm kicks off its process by first randomly classifying the words in the corpus into the various themes. After that, it moves on to the next step in the process. After that, it performs a series of iterative adjustments to the topic assignments by first determining the probability that each word in each document is associated with the respective topic, and then modifying the topic assignments depending on the results of this calculation. This procedure will be repeated until the topic assignments are as exact as they can possibly be. Iterations will be performed by the algorithm repeatedly until the topic assignments arrive at a solution that is consistent.

When the LDA algorithm has established which topics are present in the corpus, it can be put to the duty of evaluating and summarizing the information that is found in the text data. This may be accomplished by using the LDA method. It is possible to use it, for instance, to find the most essential subjects or issues covered in a set of news items or to determine patterns in the data acquired from customer feedback. Another application for it is to decide which news items include the most important topics or problems.

The idea of topic modelling may also be used to the study of various kinds of data, including sensor data and photos, amongst others. For example, in the field of remote sensing, topic modelling can be used to recognise patterns in satellite data that are associated with land cover, land use, or environmental variables. This can be done, for instance, by analysing the data in order to determine whether or not the patterns are caused by human activity.

Topic modelling is a powerful approach that can be used for the study of large and intricate datasets, as well as the extraction of helpful insights from such datasets. In general, this may be accomplished by using topic modelling.

Evaluation Criteria for the Application of Artificial Intelligence to Natural Language Processing

It is possible to evaluate the effectiveness of machine learning models for natural language processing using a wide variety of various metrics (NLP). The F1 score, along with accuracy, precision, and recall, are some common examples of metrics that are used in evaluation. Because of their utility, text classification tasks such as sentiment analysis often make use of these measurements.

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

When assessing candidates for positions that include language translation, some potential metrics to employ include the BLEU score as well as the METEOR score. Although the METEOR score examines the degree of quality over the whole of the translation, the BLEU score determines how well the freshly generated translation matches a pre-existing corpus of reference translations.

The Challenges That Come With Using NLP Along With Machine Learning

When applying machine learning to natural language processing, one of the most significant challenges that immediately presents it is the issue of data bias. The results that are produced by machine learning models might be distorted due to the fact that the algorithms could learn from biased data and then continue to propagate those biases. A model of sentiment analysis that has been trained on biased data, for example, may give erroneous conclusions when applied to certain groups of persons. This may happen because specific groups of people are more likely to express certain opinions.

The need that these models must be explainable is another challenge that must be surmounted. Because of the intricacy of the models themselves, it may be difficult to understand the decision-making process that underpins NLP models. If the capacity to explain things is going to be preserved, it is very necessary to make certain that the model is able to provide judgments that are both accurate and impartial.

2. CONCLUSION

Jobs related to natural language processing (NLP) now need the use of machine learning approaches, and several algorithms have been developed to solve a wide range of issues. The task at hand and the desired result both have a role in determining the evaluation criteria that should be used to machine learning models that are employed in natural language processing (NLP). It is vital, however, to find solutions to difficulties like as data bias and a lack of explain ability in order to ensure that the results provided by these models are credible and impartial. Research and development efforts will undoubtedly result in activities involving natural language processing (NLP), in which machine learning will unquestionably play a role that is becoming an increasingly crucial component.

3. REFERENCES

1. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning (Vol. 1). MIT press.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
4. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics, 417-424.

Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN: 2799-1172
Vol: 03, No. 03, April-May 2023
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.33.1.9

5. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics, 311-318.
6. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 65-72.
7. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
9. Murty, R., & Challou, D. (2018). Practical machine learning for natural language processing: An introduction. Packt Publishing Ltd.
10. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
11. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson.
12. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12, 2493-2537.
13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
14. Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.
15. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).
16. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).
17. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media, Inc.
18. Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.
19. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.
20. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2383-2392).