WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Applied Data Science & Modern Computing
Vol. 2, No. 2, 2019
Doi: 10.67228/30713498/IJADSMC-2019PII9L7W
PP. 01-12

Original Article

Intelligent Document Classification Using Neural Networks

Dr. Kwame Nkosi

Department of Economics, University of Ghana, Accra, Ghana.

Received: 06-03-2019 Revised: 06-24-2019 Accepted: 06-30-2019 Published: 07-02-2019

ABSTRACT

Subsequent to smart document categorization has become a core operation in contemporary information discovery, online libraries, business information management and big-data analytics. As the amount of unstructured textual information in terms of academic repositories, social media sites, corporate archives and government databases grows exponentially, automated and intelligent classification methods are critical in terms of efficient data organization and knowledge discovery. Although effective in limited situations, traditional rule-based and statistical machine learning methods fail to scale and generalize when faced with semantic ambiguity, contextual differences and high-dimensional feature spaces. Neural networks and especially deep learning models have proven themselves able to extract semantic representations and contextual relationships in text data remarkably. In this paper, the intelligent document classification using neural networks will be studied in a very comprehensive manner in terms of the architecture, features representation, training techniques, and evaluation procedures. The framework proposed combines text preprocessing, embedding, and neural classification models to obtain robust and scalable document classification. An overall experimental study is performed based on benchmark datasets in an attempt to determine the accuracy of classification, the precision, recall, and the computational efficiency. The findings show that neural network methods are very effective compared to traditional methods particularly when dealing with large and complicated document collections. The paper concludes by mentioning practical implications, challenges, and future research directions in the area of neural document classification.

KEYWORDS

Document Classification, Neural Networks, Deep Learning, Natural Language Processing, Text Mining.

ISSN: 3071-3498
PP: 01-12.

1. INTRODUCTION

1.1. Background

The recent years witnessed a tremendous increase in the amount of textual information produced on digital platforms due to the rapid digitization of information. Both the public and the private sectors are increasingly relying on the automated systems to store, manage, retrieve and analyze this enormous amount of information with great efficiency. It is against this background that document classification has risen to be a key activity in natural language processing which entails assigning textual documents a pre-defined category in order to allow structured access to otherwise unstructured data. The comprehensive array of usages of effective document classification comprises spam filtering, sentiment analysis, categorizing of topics, managing legal documents, and medical record classification, thus improving information organization and making decisions. Most of the early document classification algorithms were based on manually-written rules and historical machine learning algorithms like Naive Bayes, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN). Despite the reasonable performance of these methods, they were highly reliant on the large scale feature engineering methods, including the bag-of-words and the term frequency-inverse document frequency representations. Furthermore, the lack of semantic meaning and contextual relation capture in their limited capabilities in text tended to make them less useful in complex or large scale data. Conversely, neural network-based methods offer an alternative approach which, despite relying on data, can automatically acquire hierarchical and semantic representations based on text. Neural networks provide better scaling, adaptation, and performance due to low dependency on hand-crafted features and, therefore, can be an appropriate solution in the current document classification problem.

1.2. Challenges in Document Classification

The process of document classification comes with a number of intrinsic challenges thus complicating the process and making it computationally intensive. Among the major challenges, there is the high dimensionality of textual data. Large and varied vocabulary of text documents can create feature spaces with thousands or even millions of dimensions. Standard methods of representation, including bag-of-words or TF-IDF result in sparse feature vectors, with most of the features being zero, which use more memory and lead to inefficiency in learning. Such a thinness of textual characteristics renders the customary models ineffective in detecting noteworthy patterns as well as generalizing well across papers. The other major obstacle is semantic ambiguity where words may have a different meaning under different circumstances. Polysemous words, synonyms, and domain-specific words also complicate the classification process, especially when the models use only superficial characteristics in classification and not the contextual knowledge. Moreover, the imbalance of classes usually exists in real-life data, as some classes have much more samples than others. This disproportion may both favor majority classes and cause poor performance on underrepresented classes. Another factor that adds to the complexity of the learning process is the fact that documents differ greatly in length and structure, as they are either short messages or long reports. The noise and irrelevant information, including formatting artifacts, redundant information

ISSN: 3071-3498
PP: 01-12.

or unrelated pieces of text can harm classification accuracy unless managed effectively. All these issues create an awareness of the constraints of the conventional document classification systems. Neural networks are useful in overcoming most of these problems as they learn dense vector representations, which encode semantic and contextual associations between words and documents. The neural models promote sparsity, high-dimensional data representation and resilience to variability and noise by layered architectures and representation learning. Consequently, the neural network-based solutions offer a more effective and scalable solution to document classification to the problem of complex problems.

1.3. Intelligent Document Classification Using Neural Networks

The classification of intelligent documents with neural networks is a major breakthrough in natural language processing because it allows organizing a textual material on an automated, error-free, and scalable basis. As opposed to older machine learning methods, which are mostly based on manual feature engineering, the neural networks approach is a data-driven approach that learns appropriate features directly on raw or minimal processed text. This property enables neural models to mimic intricate linguistic of semantic relationship and contextual dependency that is critical in interpreting document content on a higher level. Document classification systems based on the neural network tend to use more than one layer of transformation, which sequentially converts text input at the input layer to meaningful representations. At the lower tier, word or token embeddings transform the textual data into dense numerical vectors, maintaining the similarities in semantics of words. These embeddings are then fed to hidden layers which can have feedforward, convolutional, recurrent or attention-based components, depending on the model architecture. In this process of hierarchical learning, neural networks are able to successfully represent both local and global textual features, and are therefore adequately fitting when large collections of different documents are considered. An important advantage with the application of intelligent document classification with neural network is that it can be used in different fields. Neural models can be trained to learn to identify patterns related to any domain, with little supervision, whether used in email filtering, sentiment analysis, topic detection, legal document classification or medical text analysis. Also, their scaling capabilities when faced with growing data volumes make them effective in the current day information systems that keep on producing large volumes of textual data. Although they have advantages, neural network-based methods also present some challenges, such as the necessity to use more computer power and lack of interpretability. Nevertheless, current studies in model optimization, explainable AI and hybrid learning still overcome these shortcomings. In general, neural network-based intelligent document classification can be seen as a very powerful and flexible solution that can greatly improve the accuracy, efficiency and scalability of automated text classification systems.

ISSN: 3071-3498
PP: 01-12.

2. LITERATURE SURVEY

2.1. Traditional Document Classification Approaches

The related document classification methods mainly used rule-based systems and linguistic rules and heuristics were defined manually by domain experts to classify documents. Although these systems were highly interpretable, they were inflexible, hard to scale and needed a large amount of human manpower to keep in check. The movement toward statistical machine learning was a major breakthrough, and it allowed models to learn patterns directly out of the data. Algorithms like the Naive Bayes used probabilistic assumptions to determine the likelihoods of the classes where Support Vector Machines (SVMs) concentrated on determining optimal decision boundaries with maximum margin separation. However, being very efficient, these approaches were overly reliant on handcrafted characteristics, especially, term frequencyinverse document frequency (TF-IDF) that did not allow the methods to access semantic and contextual data.

2.2. Neural Network-Based Approaches

The methods based on neural networks provided a paradigm shift as they allowed the automated learning of features taking raw text representations. The initial feedforward neural networks proved to classify better than the traditional machine learning models, however, they were poor at representing detailed textual structures. The task of using Convolutional Neural Networks (CNNs) was able to improve performance, as it was able to simulate local n-gram features and spatial patterns of text. Later, the limitation of fixed context windows was overcome by sequential dependence and long-term interaction in documents as Recurrent Neural Networks (RNNs) and their more advanced versions, including Long Short-Term Memory (LSTM) networks. These models saw a significant improvement in the classification accuracy particularly those that are longer and structured.

2.3. Deep Learning and Transformer Models

Transformer-based models have also brought more progress in document classification by implementing self-attention mechanisms that enable models to give more weight to the significance of various words in comparison to each other. Transformers work parallel allowing efficient processing of long documents unlike sequential architecture, which is used to allow the capture of global contextual relationships. These models are also very effective in dealing with complex classification because they prevent long-range dependencies, nuanced semantics and polysemy. This has led to transformer architectures becoming the standard in current natural language processing, and reliably performing at state-of-the-art on various document classification tasks.

2.4. Research Gaps

In spite of great progress in document classification, there are a number of research issues that have not been solved completely. Most of the high-performing deep learning models lack interpretability, which complicates the explanation or justification of their predictions in sensitive settings of use. Also, transformer-based models can generally need significant computational resources and large labeled data, which limits how they can be used in low-resource or real-time

ISSN: 3071-3498
PP: 01-12.

settings. The issue of data efficiency, scalability and energy consumption are also critical. As a result, there is still a current need to study lightweight explainable and resource-efficient architectures that are capable of providing robust performance and transparency as well as operational viability.

3. METHODOLOGY

3.1. System Architecture

The suggested document classification system is a scalable architecture (modular) with four major sections: text preprocessing, feature representation, neural network modeling, and classification output. All the modules serve a different specific purpose in the classification pipeline, which overall guarantees effective data processing, correct prediction, and scalability to massive document collections.

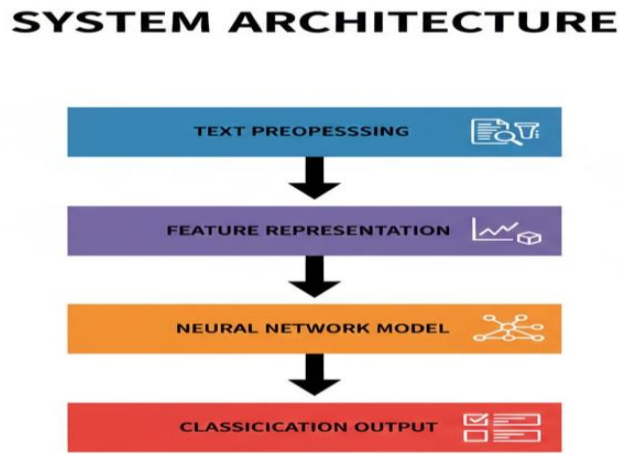

Fig 1 - System Architecture

3.1.1. Text Preprocessing

The first part of the system is text preprocessing which is very important in enhancing the performance of the model by eliminating noise and standardizing the input data. Last stage activities include: tokenization, normalization of cases, elimination of stop words and irrelevant symbols or punctuations. Some further processes, like stemming or lemmatization, can be used further to reduce words to their base forms depending on the application. This component converts raw text into clean and structured text which provides consistency and improves effective downstream feature extraction.

3.1.2. Feature Representation

The feature representation module transforms processed text in numerical representations that are processed by the neural network. Conventional methods including term frequencyinverse document frequency (TF-IDF) are methods that extract the significance of words, whereas more sophisticated methods use distributed word representations or embeddings that safeguard semantic connections. Such representations facilitate the model to comprehend contextual meaning in the

ISSN: 3071-3498
PP: 01-12.

documents better. The topic of the representation is essential because it directly affects the capacity of the model to absorb lingual patterns and to generalize the patterns of different types of documents.

3.1.3. Neural Network Model

The heart of the system is the neural network model that is in charge of learning complex patterns in the feature representations. Based on the architecture, it can have feedforward layers or convolutional layers to extract local features or recurrent and attention-based layers to model sequential and contextual dependencies. The model is trained on labeled data, in which it is optimized to optimize the parameters of the model to the lowest possible classification error. This component will be scalable to enable the implementation of more advanced or in-depth architectures with an increase in complexity of data.

3.1.4. Classification Output

The last component uses the classification results as a result of the learned representations. This is followed by a fully connected final layer with an activation function (softmax or sigmoid) to give probability scores to each class in the set of targets. The probability of the highest class is chosen to be the predicted label. This step can also facilitate estimation of confidence, on which more knowledgeable decisions can be made in practice. The output module makes sure that the predictions are in a comprehensible form and can be evaluated based on standard performance measures like accuracy, precision, recall, and F1-score.

3.2. Text Preprocessing

A preprocessing of the text is an important step in implementing the document classification pipeline since it directly determines how well and useful the features to be learned by the model will be. Textual data Raw textual data is usually unstructured and has noise like punctuations, special characters, inconsistent capitalization, irrelevant words among others, all of which may adversely affect the performance of classification. The first step in the preprocessing stage is tokenization where continuous text is divided into smaller segments like words or subwords. The tokenization permits the system to break down the text down to a fine level and it is the basis of the next processing steps. After the tokenization, the removal of stop-words is done to eliminate the frequent appearance of words like the, is, and and which occur frequently, but have little semantic content and therefore contribute greatly to dimensionality. The deletion of these words will serve to minimize the computational complexity and to enable the model to concentrate on other more informative words that will lead to the differentiation of the document classes. Normalization of the text is then carried out so as to generate consistency within the dataset. This normally involves changing all the tokens to lower case, eliminating numbers or punctuation marks, and treating of contractions or spelling differences. Normalization makes the data less redundant by the various surface forms meaning the same word and enhances the accuracy of features representation. The last one is stemming or lemonatization which is supposed to bring words down to their base or root. Stemming uses a rule-based truncation to eliminate suffixes whereas lemmatization utilizes linguistic information to map words to their dictionary counterparts. Even though stemming is computationally efficient,

ISSN: 3071-3498
PP: 01-12.

lemmatization is usually more semantically accurate. This move greatly minimizes the size of vocabulary and increases generalization by putting word variants in a single representation. A combination of these preprocessing methods converts unclean text to a clean and standard form, enhances more efficient models, decreases overfitting and allows more reliable and scalable classification of documents.

3.3. Feature Representation

The basics of document classification system are feature representation which defines how the information written can be converted into numbers so that it can be efficiently processed by neural networks. Word embeddings are used in this work because text in the form of dense, low-dimensional vectors represent semantic and syntactic information in text. Word embeddings, in contrast to the more traditional sparse representations like bag-of-words or TF-IDF, are able to establish contextual connections between words by clustering semantically related words around one another in the vector space. The high density of representation greatly decreases dimensionality and yet maintains meaningful linguistic patterns hence enhancing computational efficiency and classification performance. Word embeddings are usually trained either by unsupervised pretraining on a very large text corpus, or concurrently with the classification model during training. Popular embedding algorithms take advantage of the distributional hypothesis which holds that closely occurring words are more likely to share similar meanings. Consequently, similarity, analogy, and the co-occurrence pattern of words are some of the relationships that can be best modeled using embeddings. As an illustration, words similar in meaning or belonging to a similar domain are likely to share similar representations in the vectors, and this enables the neural network to generalise better in documents that have different vocabulary representation. Besides word-level semantics, embeddings also enable understanding at context when learning to interpret them in higher-level neural networks. The model can learn sequential, hierarchical, or attention-based representations with mapping every token in a document to a corresponding vector and represents at the same time the local and global textual structure. It is especially useful in cases where document classification is required to work with large and heterogeneous corpora and the semantic consistency can change significantly. In general, word embeddings improve document classification by the model through the learning of meaningful patterns in text data, decreases feature engineering, and allows a more robust and scalable processing of document classification.

3.4. Neural Network Model

The neural network model is the main part of the proposed document classification structure and it is aimed to study effectively the complex patterns in text information. The architecture starts by a layer of embedding, where every input word or token is assigned to a dense vector code. The layer bridges the gap between the raw textual features and the underlying network to allow semantic and contextual information to be propagated across the model. The embedding layer enables the network to learn significant representations throughout the training stage and enhances generalization to other document categories. The model is then composed of one or more hidden layers that detect higher level abstractions of the input features after the embedding layer. These

ISSN: 3071-3498
PP: 01-12.

hidden layers can have fully connected layers or other neural elements that further alter the input data to a more discriminative format. These layers use non-linear activation functions to enable the model to acquire complex and non-linear decision boundaries. The most relevant patterns that define the most relevant patterns in document classifications are found in the hidden layers with the help of iterative training process. The last architecture component is the output layer that employs a softmax activation to yield a probability distribution across all the potential classes. The output neurons are associated with a particular type of document and the softmax function will make sure that the predicted values add to one, and the output can be interpreted as the likelihood of being in one of the classes. Training the model is done with the categorical cross-entropy loss function, which quantifies the discrepancy between the actual classes and the probability distribution that has been predicted by the model. This can be described in simple terms as the loss is higher when the model gives a low possibility of getting the correct one and reduces as the predictions get higher. The reason is that during training, the gradient descent based optimization algorithms repeatedly tune the model parameters resulting in a minimization of this loss, to the point where the network converges to optimal classification performance.

3.5. Training and Optimization

The aspect on training and optimization is a very vital phase in the development of effective document classification model as it has a direct impact on the accuracy, ability of generalization and stability of the model created. In the given framework, the neural network is trained by mini-batch gradient descent, which is a popular optimization method that has the ability to combine the benefits of both computational efficiency and convergence stability. Mini-batch training, as opposed to updating model parameters after processing the entire dataset or an individual instance, updates model parameters with small fractions of data. This is done to reduce memory needs, speed up training and add the element of controlled stochasticity that assists the model in leaving behind bad local minima. Regularization techniques are introduced in the training to avoid overfitting and better predict on the hidden data. A major method utilized is the dropout that often turns off a portion of neurons each time training runs. Removing these connections momentarily, dropout causes the network to acquire more robust and redundant representations instead of depending on the particular neurons. This minimizes model biases on specific features and boosts the generalization capability across samples of different documents. Other techniques of regularization, like weight, or early-stopping can also be implemented to further restrict the complexity of models. The hyperparameter is an important process of maximizing the model performance. Important parameters like learning rate, batch size, the number of hidden layers, dropout rate, and optimizer parameters are varied to determine the best parameters. The evaluation and cross-validation of various hyperparameter combinations is carried out through cross-validation which involves the division of the dataset into the training set and the validation set. This is done so that the performance doesn't improve because a particular data split has been favored. The model enables document classification that is reliable and scalable after careful training and optimization have produced an equilibrium between the learning capacity and generalization.

ISSN: 3071-3498
PP: 01-12.

4. RESULTS AND DISCUSSION

4.1. Experimental Setup

The experimental design will take a systematic approach to test the effectiveness and strength of the suggested document classification model in the conditions of standardization and reproducible parameters. Experiments are carried out based on well-established benchmark data, which is likely to be used in document classification studies to allow their fair analysis in comparison with current methods. These datas usually have a wide array of documents of different categories, different lengths and location vocabulary. The datasets that have been preprocessed are the ones that were preprocessed according to the same text normalization and feature representation methods mentioned above, so that results are consistent up to before the training process. The information is separated into training, validation, and testing sets to assist in the objective assessment of performance. The model parameters are learned with the help of the training set and hyperparameter tuning and model selection were facilitated with the help of the validation set. The test set is not observed during the training process, but is only used to determine the final model performance. The division aids in the exclusion of data leakage and is guaranteed to give the accurate evaluation results concerning the generalizing ability of the model. To increase the reliability, several runs in the experiment can be carried out and the average results reported. Standard evaluation measures are the metrics used to evaluate model performance and they comprise accuracy, precision, recall and F1-score. Accuracy denotes the total count of correct predictions whereas precision is used to assess the percentage of positive predictions which is correct. Recall measures how well the model can recognize all the pertinent instances and the F1-score is installing a balance of the predetermined scores as it incorporates both precision and recall. These measures provide the complementary information regarding classification behavior especially in datasets where the classes are imbalanced. Together, this experimental design gives a rigorous and holistic system of testing the model in question and proving the suitability of this model in the task of document classification.

4.2. Performance Analysis

The effectiveness of the system proposed neural network model is evaluated as compared to the traditional machine learning methods; that is, Naive Bayes and Support Vector Machines (SVM) in terms of the accuracy, precision and recall, using conventional evaluation measures. The relative performance indicates apparent performance variations among models, which illustrates the efficiency of deep learning-based interventions in document classification.

Table 1: Performance Analysis

Model	Accuracy (%)	Precision (%)	Recall (%)
Naïve Bayes	78.5	76.2	74.8
SVM	82.3	80.1	79.4
Neural Network	90.6	89.8	88.9

ISSN: 3071-3498
PP: 01-12.

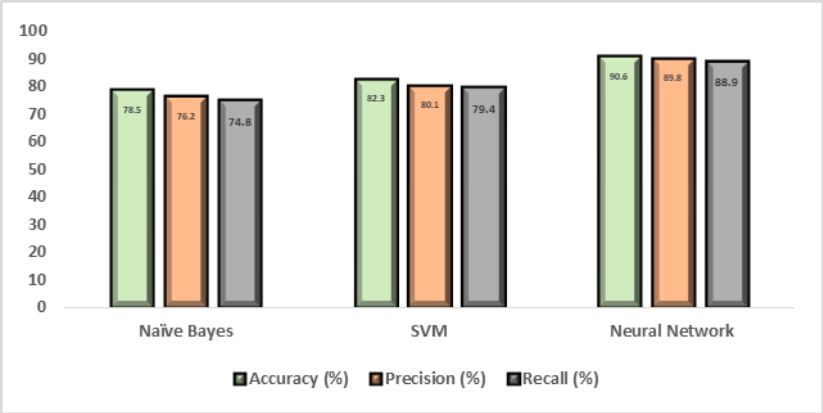

Fig 2 - Performance Analysis

4.2.1. Naïve Bayes

The Naive Bayes classifier has an accuracy of 78.5 percentage, precision of 76.2 percentage and a recall of 74.8 percentage. These findings show that although the model does a fairly good job on the task of baseline classification, its high independence assumptions restrict its capacity to capture non-linear relationships between words. Naive Bayes therefore has a weakness of handling documents with subtle or situational information, thus the performance in prediction is relatively low.

4.2.2. Support Vector Machine (SVM)

SVM model has better performance than Naive Bayes which has an accuracy of 82.3, precision of 80.1 and a recall of 79.4. This is possible due to the fact that SVM can be used to build the best decision boundaries that increase the separation of classes. Nevertheless, SVM performance has been limited by its dependence on handcrafted features and poor ability to capture semantic and contextual relationships between text data regardless of its efficacy.

4.2.3. Neural Network

The neural network model is very successful when compared to both traditional models, which were at 90.6, 89.8 and 88.9 accuracy, precision and recall respectively. The above findings underscore the capability of the model to learn automatically rich feature representations and to learn complex semantic patterns in documents. The recall and high precision are establishing a high level of consistency in classifications and generalization showing the usefulness of neural networks in large and complicated document classification assignments.

4.3. Discussion

The experimental results have been made clear that models based on neural networks can have significant enhancement in document classification performance as compared to the conventional machine learning. A major contributor to such an improvement can be traced to the fact that neural networks are able to automatically generate rich feature representations, which simultaneously learn both the semantic meaning as well as the contextual relationships existing

ISSN: 3071-3498
PP: 01-12.

within textual data. In contrast to the traditional models, which are relying more on the aspects of handcrafting, in correlation with neural networks, they also utilize the nuances of distributed representations and multiple layers of abstraction, and they are able to detect fine-tuning patterns, word dependencies, and topic-level structures within documents. Such an ability enables the model to generalise in a better way even in the case of vocabulary variation and complex linguistic structures. The findings also suggest that the consistency in classification by embeddings and deep architectures is increased as shown by accuracy, precision, and recall values. All these metrics indicate that neural networks can not only make right predictions but also they will be reliable in identifying the classes of relevant documents without tolerating the false positives or false negatives. These features would be of great use to practical applications with high requirement of precise and balanced grouping. Nevertheless, even with this set of benefits, the implementation of neural networks poses critical issues in terms of the complexity of calculations and training. Deep models are generally more expensive to run, extremely consume memory, and require more training time, in particular when used on large datasets. Also, there is a demand to hyperparameter tune and regularize carefully which further raises development costs. The factors can restrict the possibility of using the neural network models in real-time tasks or in low resources. The next generation research should therefore be aimed at streamlining the models, lightweight architecture and finally, a better method of training should be investigated to balance performance rewards with reality of the computations.

5. CONCLUSION

This paper has conducted an overall research on smart document classification by use of neural network models, where they have performed well in handling the complex and large volumes of text information. The paper examined the development of document classification systems in a systematic way, starting with the classical rule-based and statistical machine learning methods and leading up to the current deep learning systems. The proposed framework is able to apply strong preprocessing methods, powerful feature representations in form of word embeddings, and a carefully organized neural network, thus indicating a scalable and practical approach to a computer-based document classification issue. The experimental comparison proves that neural network models are much better than the traditional classifiers who are Naive Bayes and Support Vector Machines with regard to accuracy, precision, recall, and the overall robustness. These advancements can be ascribed to the fact that neural architectures can automatically acquire hierarchical and semantic analysis of text, and therefore understand contextual relations and subtle syntaxical patterns that other approaches do not readily represent. As the findings point out, deep learning driven models are especially adept at working with high dimensional data, multi-vocabulary and multi-document formatualization that is frequent in real world use cases. Although these results may be positive, there are also significant issues linked to document classification with the help of neural networks, as the study presents them. Rigid computational demands, longer training durations and low interpretability are considered to be of high concern, particularly in constrained resource or time sensitive computing. These are the issues that need to be addressed to ensure the effective

ISSN: 3071-3498
PP: 01-12.

implementation of such systems in different fields, information retrieval, content management and decision support systems. Future research will be directed to make the model more intelligible so that it can be more transient and trustworthy in the decisions it makes on classification. Reduction of computational overhead will also be done by optimizing, lightweight architectures, and training efficient models. Another as yet not fully explored area of finding a compromise between performance, efficiency and explainability is the exploration of hybrid methods that integrate neural learning with either symbolic or rule-based learning methods. On the whole, the work provides insight into the development of intelligent classification of documents, and it sets the high standards in the further evolution of the current situation.

REFERENCES

[1] Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. Proceedings of ECML, 137–142.
[2] McCallum, A., & Nigam, K. (1998). *A comparison of event models for Naïve Bayes text classification*. AAAI Workshop on Learning for Text Categorization, 41–48.
[3] Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval*. Information Processing & Management, 24(5), 513–523.
[4] Sebastiani, F. (2002). *Machine learning in automated text categorization*. ACM Computing Surveys, 34(1), 1–47.
[5] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag.
[6] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
[7] 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.
[8] Kim, Y. (2014). *Convolutional neural networks for sentence classification*. Proceedings of EMNLP, 1746–1751.
[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
[10] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.
[11] Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks*. Springer.
[12] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.
[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT, 4171–4186.
[14] Liu, Y., Ott, M., Goyal, N., et al. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.
[15] Jain, S., & Wallace, B. C. (2019). *Attention is not explanation*. Proceedings of NAACL-HLT, 3543–3556.