a

International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Long Document Classification using Hierarchical Attention Networks

Ms. Ayesha Mariyam, SK. Althaf Hussain Basha, S.Viswanadha Raju

Submitted: 23/10/2022 **Revised:** 24/12/2022 **Accepted**: 21/01/2023

Abstract: Online comments and reviews are the primal exponents of the current era, which needs thorough conceptual discriminations and opinion analysis. Full-text analysis using document classification methods are traditional approach. Deep learning methods are employed on long legal documents for text classification. The word level classification semantically signifies the sense of classification. Statistical methods exist like TF, TF/PDF, and TF/IDF. Document classification is sentence-level classification using sentence vectors mentioned in hitherto successful research consensus. The challenge lies in long legal documents classifying based on the sense in sentences. The article represents challenges in the existing propositions and ideas for implementing long document classification using attention learning. A CNN-based attention learning model is described for classification on BBC Web News datasets. The results are appraised using performance evaluation metrics and RoC graph and have accomplished estimated accuracy of 96%.

Keywords*: long document classification, word2vec, sentence2vec, attention model*

1. Introduction

As the computing society is experiencing rapid developments in various methods and methodologies, traditional cultures are extinct, no matter their excellent applications [1]. In such prevailing conditions, promoting the traditional culture lure much demand, and the literature connotes the excellence of culture just being inherited. More resolutions, deliberations, and extensions are evolved for the traditional cultures with time. Text classification has been essential, particularly in disorganized kinds of literature. Traditionally they are done by hand [2], where manual extraction could not meet people's needs when the growth of text is at an exponential pace, and deriving valuable information accurately with no time demands the utilization of computers, as to work large quantities [3].

Automatically writing articles, précis, comprehensions, and summaries is proposed by Dr. Luhn in the 1950s. Statistically counting the frequency counts of terms, words, and sentences with spatial distribution is a typical traditional method. Bibliographic indexing and library searching to include in statistics, although developing into statistical text classification technologies. Setting the rules appropriately, applying the huge texts for classification, and summarizing the text characteristics are the foundations of text classification.

Text Analytics

A great surge is experienced in the world of unstructured and text data. The ubiquity and continuous conceptual growth in text and data contribute to proliferation. The rapid growth of text data is experienced in social media. Valuable information and nuggets are buried in unstructured text data. They can be used and exploited for decision-making supporting various enterprise activities if identified and properly extracted.

Text analysis attempts to understand the meaning of the written word. This isn't easy because it depends on the communicational situation of humans. Therefore, the development of social media surrounds a level of communication inability.

Document Classification

Classification of documents into predefined categories or groups is the essence of document classification. Document classification is automated to manage a large number of text documents. Documents are classified by labeling methods and unsupervised learning methods when the category is not readily available for the classification task [5][6][7].

Business activities of an enterprise evolve with possible categories which support supervised learning. Feature selection, feature extraction, vector forms of document representation, and feature analysis with the application of machine learning algorithms [7][8][9].

Document classification is classically categorized into two phases: deriving feature selection metrics and predicting the document label. While deriving feature selection, several metrics like information gain, correlation coefficient and other statistical document metrics are used. The features are given as input to the different classification algorithms like support vector machines and Naïve Bayes classification methods to predict the document label [5]. Documents may also be converted to features by feature selection methods and later fed into the text categorization algorithms.

Classification in Long Documents

Any particular label cannot classify documents with lengthy concepts and content that are. Still, they can be limited to adhering to the set of labels that belong to a context. Context-based classification can categorize long documents, which are lengthy unstructured documents [7][8][9][10]. An agent is trained to identify the groups of words which are discriminative that support a particular category of context in the document. Attention-based learning [4] is applied for categorizing the long document with the help of an agent, where groups of words that belong to a context are identified, and the document is discriminated as a category.

2. Related Work

Learning in Classification

Classification is a predictive technique where data from unknown sources is categorized into different classes. A label defines each class, and the classifier is trained and tested on the datasets for the best fit. A learning mechanism is implemented to train and test the classifier on the datasets. Classification is a supervised learning technique that discriminates data into categories, and new observations of classification experiments can draw the basis of training data. Observations from the classification experiment are learned to evolve new classes or groups [5].

Classification using the supervised learning technique is applied with a single classifier or multiple classifiers. A single classifier mechanism derives a binary classification method, categorizing datasets into a class or no class. A multiple-classifier mechanism derives multi-class or multi-label classification where many categories of datasets are derived. A classifier can define one or more labels, whereas a multi-label classification can undergo the process with a classifier that determines the groups of data into multiple labels.

Deep Learning for Classification

Bankers' credit approval, medical image analysis, and diagnosis deriving competitive intelligence and target marketing are the applications of salience for classification using machine learning. It is called lazy learning when classification is performed based on the data sets that are stored for training and waiting to prove the classification methodology until the test data is ready. Data sets are prepared for training the classifier before the data sets arrive, and a classification model is ready before the classification process is undergone, called eager learning [5][6].

Further, the classification model is built based on the type of input datasets and objectives, viz., predictive classification modeling, binary classification, multi-class classification, multi-label classification, and imbalanced classification [10][11][12][13].

Natural Language Processing and elementary statistics play an important parallel role in document classification. Conventionally, one or more predefined categories are assigned to the sequence of texts, and a feature extraction follows. Classification of texts is undertaken after the feature extraction. More popular TF-IDF methods are used to build the text vector in a document, which acts as an input feature to the subsequent classifier in the classification process. Deep Learning networks are proven to perform such predictive tasks more efficiently. Feature extraction and classification for document classification tasks include input vector extraction and passing them into the convolutional layers of deep learning. The anatomical components of a document are letters, words and sentences, and then paragraphs. Each word as a rudimentary element is projected into a continuous vector space sufficiently multiplied with a weighted matrix. The weighted matrix evolves to generate a sequence of dense, real-valued vectors. Further, the sequence is fed into the deep neural network, which processes in multiple layers, almost resulting in positive prediction probability. The chain of activities in the deep neural network is refined and tuned to maximize the classification algorithm's accuracy and robustness using a training set [14][15].

Deep learning has recently ushered into the process flows based on the level of words, where there were many shortcomings in previous applications of deep neural networks [16][17]. Each word is a token considered separately, and parameters are estimated, even though many words share the same root, as either prefix or suffix.

Deep learning architectures for text analysis work on different levels of words and sentences. Word-level approaches have many major hindrances, as each word is considered as a separate token and estimates by the number of parameters such as suffixes and prefixes [17][18][19]. Word-level classification always has a shortcoming with the words having similar roots. This problem is solved by using an exclusive language-dependent mechanism of segmentation of words into various components such as root, prefix, and suffix, which will be ideal for language-dependent document classification. Word-level text analysis tasks cannot handle the problem of out-of-vocabulary situations if the word is not available in the training corpus, which will be hazardous that it will be mapped to any unknown or wrong word. Typos are not handled in the word-level text analysis methods. A deep learning model is limited to a training corpus that cannot handle combinatorial problems arising from the non-availability of words and non-descriptive typos from informal documents on social network sites. Therefore, applying a trained model for

the new domain is difficult, delivering huge mismatches between the training corpus and the target domain.

Attention Learning

Application of Natural Language Processing and Convolutional Neural Networks in document classification has much interest in Research in recent years, where CNNs are proven to be applied directly to the words embedded in the texts [21]. There are word-level and character–level models, where character-level models achieve competitive results. CNN cannot work out long sequential information in text classification, which impacts the performance, especially amongst the sentences representing semantic transitions and negations [22]. Recurrent Neural Networks emerged with Natural Language processing which uses recurrent neural networks as gated networks in document models and sentimental analysis. A hierarchical model of recurrent neural networks and long-short term memory is evolved in [21] for text classification. Contextual information in the documents can be extracted by RNN, especially complying with the semantics of the long texts. RNNs cannot perceive certain important contexts in the long text or determine the deep semantics and sense in the long texts [23].

"Attention" is the recent model akin to human attention. The complete mechanism of human attention is attempted to simulate through a long-short term memory and recurrent neural networks, which forms a hierarchical learning network and defines a new Hierarchical Attention Model of document classification.

Sequence-to-sequence Model:

A seq-2-seq model is a deep learning model that is used in the translation of sequences during classification. An input is a series of words; an output is a series of translated words in a neural machine translation.

Fig. 1. A seq2seq model

This is a model composed of encoders and decoders. The context of the input series of words is captured and into an input sequence, which is represented as a sequence vector and sent to the decoder to produce the output sequence. At the same time, using the hidden states for translating the input vector to the output vector, a recurrent neural network consisting of the two hidden layers is employed.

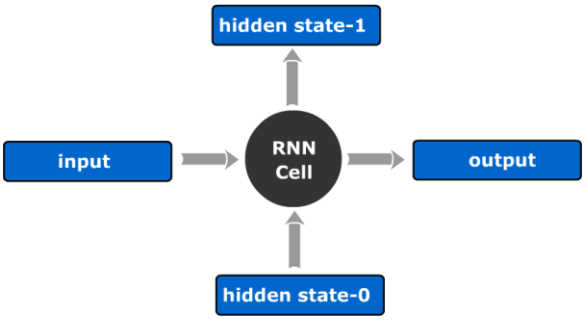


Fig. 2. A RNN with hidden layers

Natural Language Processing has achieved the possible results in recent years in text summarization, caption identification and generation, machine translation of texts, and conversational models, satisfactorily accomplished with the sequence-to-sequence (seq2seq) model. But seq2seq models do not fit all skewed long text documents, where the entropy of the documents is very high. The recurrent neural network model is introduced as a hidden layer perceptron, as an 'attention' during the inputs to mitigate the issue. The results are improved from the consensus of the Research works related to convolutional neural networks in classification compared to the said attention model of the recurrent neural network. The only reason is in the seq2seq model, and each word becomes a member of the bag having a label. Multi-language tasks can be handled with a seq2seq model with relevant corpus and dictionary using 'machine translation.' Such frameworks are also suitable for parts-of-speech tagging tasks. As for the fact, in text classification, there will not be any predefined labels for the words for the training set documents, although indicating the relevance of the category and challenging the efficiency of text classification with perceptron in one hidden layer.

CNN-based Attention Model

Attention could be vested only in the long documents' small amounts of information related to a particular context. From the consensus of cognitive and neuroscience Research, a typically human phenomenon of attention can be vested as 'task-oriented-templates.' The attention mechanism thoroughly studied and drawn from neuroscience is proven adaptable to computational linguistic text analysis. As convolution is the template-matching process, the convolutional neural network is the natural model that stimulates human attention.

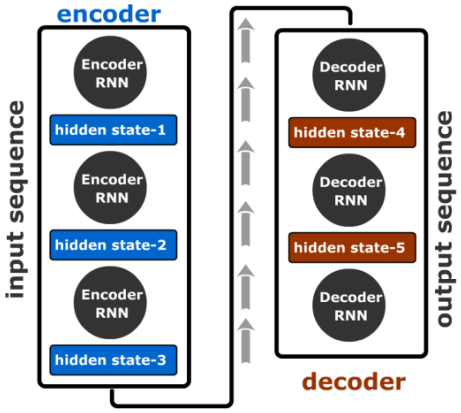


Fig. 3: An Encoder and Decoder model of seq2seq.

One-dimensional convolution is applied in text analysis. The concatenated representations of vectors of words, which are the snippets of texts and 'attention search templates,' are the convolution input to discover the similarities. A text sequence of length T is also represented as $s_{0:T-1}$, which means $s_0 \oplus s_1 \oplus s_2 \oplus \ldots \oplus s_{T-1}$, where s_t is a d-dimensional vector R^d and representation of the T-th word in the text sequence and $\oplus$ is the vector concatenation operator. A window of l words is applied with the convolution operator, filter w. where $w = w_0 \oplus w_1 \oplus \ldots \oplus w_{l-1}$ of the original text sequence to obtain convolution similarity c_t. A transformation called c_t is applied to the text sequence $[s_0, s_1 \ldots s_{T-1}]$. The output of each convolution is a signal of attention to the original text. Filters may be applied to the noise signals to separate the text sequences. For the convolutions' m', the filter is denoted as $[w_1, w_2 \ldots w_m]$ with supported attention similarity of $[c_1, c_2 \ldots c_m]$. Smooth attention may be arrived at after averaging the attention similarities concerning the filters $c \in R^T$, representing the importance of each word in the sequence as $c = \sum_{i=1}^{m} c^i$.

Convolutional-Recurrent Attention Network Model

The convolutional model and the recurrent neural network are combined to form a convolutional-recurrent attention network model. The traditional CNN uses a pooling layer on the whole input of long sentences, resulting in a sequence represented as a vector. For a CNN, capturing long-distance dependencies in sequences containing transition and negation is very difficult/time-consuming.

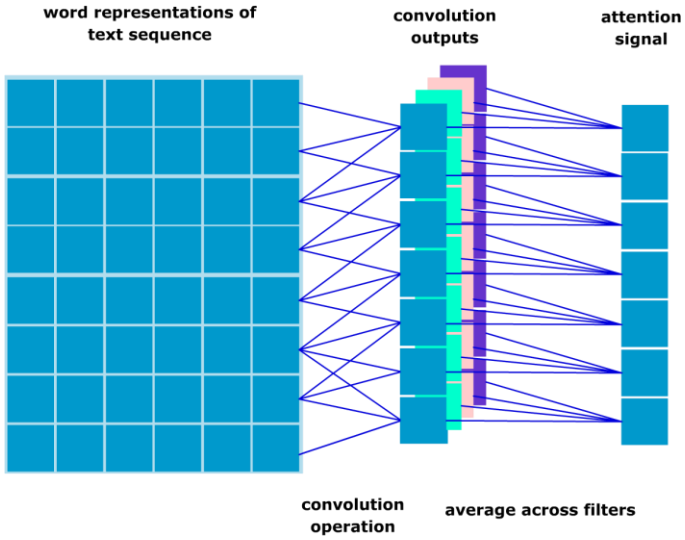


Fig. 4: Attention in CNN Model

As the RNN is associated with LSTM, which is exclusively designed for handling long-distance dependency, combining CNN with the RNN can deliver good results in text analysis.

As the lengths of the sequences from the sources are arbitrary, RNN processes by recursively applying the function onto its hidden state vector for each element in the input sequence, where the hidden state vector at a time-step is h_t in the d-dimensional space. The hidden state vector of the input sequence at a given time, t, depends on the input symbol in the sequence x_t. Thus hidden state vector of the last time-step is h_{t-1}.

$$h_t = \{\ 0\ g(g_{t-1}, x_t) \quad t = 0\ otherwise$$

The limitation with the recurrent networks is that gradients processed and propagated over many steps tend to get exhausted. In contrast, the RNN exhibits difficulty learning in long-dependency correlations in the sequence. To overcome this problem in RNN, the LSTM is associated with the architecture. The LSTM works with three gates along with a memory cell. At any given

time-step ***t*** in the d-dimensional space, the vectors are the *input gate, ft is the forgotten gate, ot is the output gate, and ct is* the memory cell.

Attention Model

In learning networks, cognitive attention at any stage of learning is called attention. An interface between encoder and decoder modules where the decoder is provided information from hidden states of every encoder to devise a model that focuses selectively on valuable parts of the input sequences and establish an association between the word embeddings and the sentences of the documents.

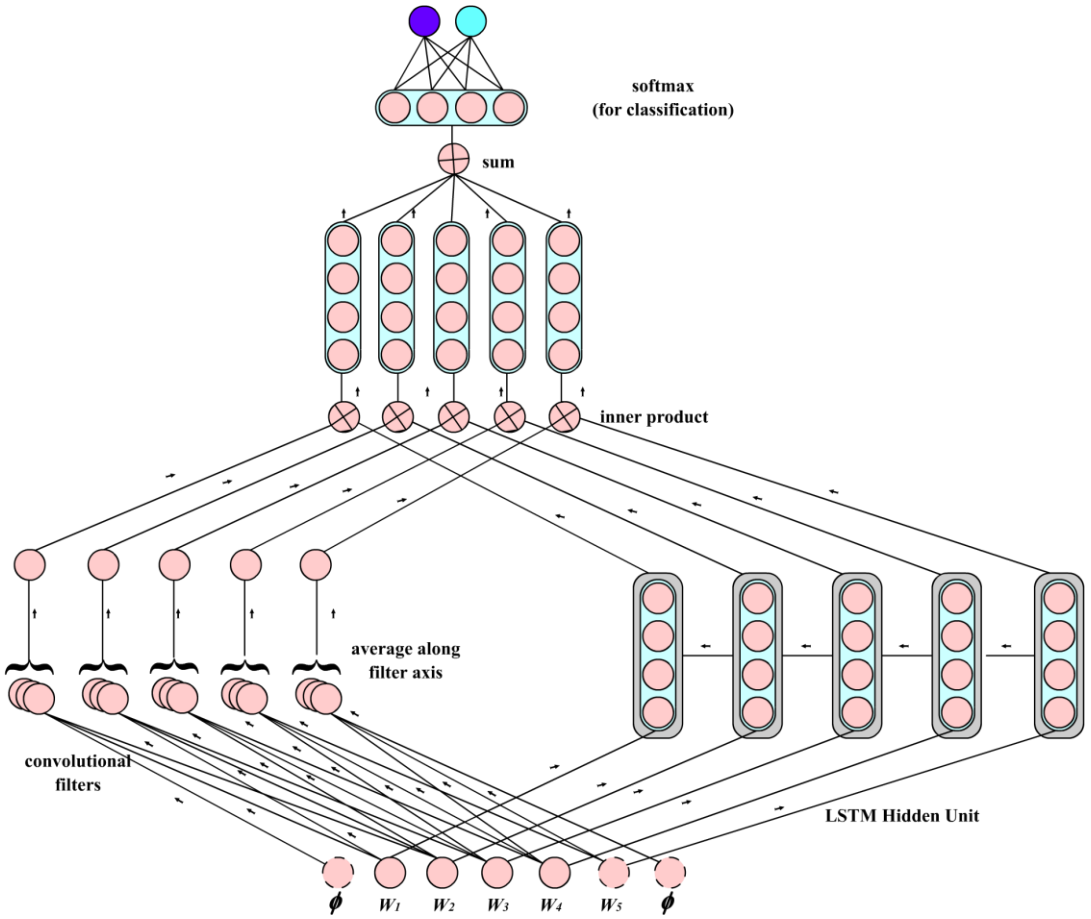


Fig. 5: Convolutional Attention Neural Network

This model competes with the long document classification models on long input sequences compared to traditional classification models of data engineering. The effect of attention enhances the quality of learning but has limitations in other aspects, like quick learning. Small and detailed parts of data are focused on to mark attention flags during the learning process. Often, the importance of details becomes questioned during learning. In the context of long document classification using attention, the attention is a trained vector developed from the input sequences.

Types of Attention Learning

A state definition is required in the input sequence to derive attention. The type of attention is determined based on the nature of the source states in the input of sequences. Global attention is derived from being attention established on many source states in the input of sequences. For attention established on more than one selected state of a particular context in the input of sequences, local attention is derived. Attention is placed on a single source with a specific version of context. Hard attention is derived. For applications where selected states are focused on deriving new contexts depending on the contexts, especially for building a classification model, soft attention is derived.

Anatomy of Attention Learning

The important anatomical components of Attention Learning include encoder-decoder modules, LSTM, and RNN. Based on the input processing capability, an LSTM or RNN is strongly selected for the attention learning model's built-up. Attention is defined on the source state, and the encoder processes the input sequence. If the encoder mistakes the translation of the input sequence, the decoder will follow the same mistaken translation of the input sequences. The phenomenon of translating a particular state of the input data while accomplishing learning for classification or any predictive applications is 'attention modeling.' Typically in an attention learning model, there are encoder-decoder modules, learning modules with translation layers in the neural networks comprising as main components in the anatomy of attention learning.

Even complex developments of the attention-learning model include transformers as constituent elements. Memory-augmented and neural graph networks are also auxiliary components in the attention-learning model based on specific applications containing input sequence data such as images, video, sound, etc.

Applications of Attention Learning

The characteristic of cognitive attention in attention learning has become an important asset for most learning networks. Adding an attention model adds up a notch of quality for all the applications in artificial intelligence. The demand for attention mechanisms amassed for their versatility in computer vision and natural language processing applications. The attention models have attracted researchers in computer vision and natural language processing. In medical and healthcare systems, a tool such as attention has a strong need that supports improvements in routine clinicians' job of conversations with patients, where a summary of medical and health status can be prepared before the treatment.

As the limitation in the CNN and RNN, which they exhaust at a point where learning reaches the gradient of genericity, the attention model plays a very important role where the source states are recorded to limit the definition of knowledge learned from the documents. The attention models are ideal for bearing with the long dependency of sequence text data. Working on the legal documents and summarizing the legal decision or judgment will be an ideal application for the attention models. Word2Vec model is employed in defining the attention model for legal documents where a glossary of terms is built as a meta-data or vector, and further concept embedding shall take place to match the legal resolutions from the documents based on the keywords. Strongly motivated by the behavior of observation and identification of the characteristics of the states of an input sequence, the attention mechanism is introduced to computer vision with the sheer aim of imitating a human visual perception system. The attention model is deployed on the real-time unclassified data sets. It can be called 'an attention model with dynamic inputs where dynamic weights are adjusted from the features derived in the images.'

Attention learning networks also are popular but have yet to invent more prospective frameworks and methods in criminal documents. The attention mechanism that works on categorical meta-data, which is renowned as a categorical attention learning mechanism, evolved into categorical dependencies and fully connected networks.

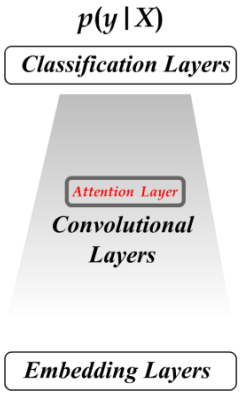
3. Proposed Work

Classification using Attention Learning

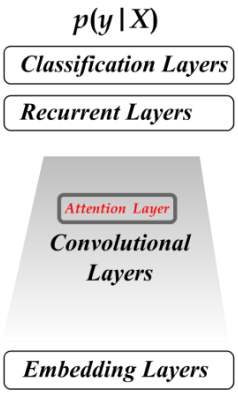
Out of classical classification problems, document classification is a popular and interesting area of Research in natural language processing. The main purpose of this novel perspective methodology is to mark the documents category. A wide range of applications supports document classification, such as defining a bunch of topic tags, sentimental classification, polarity classification of opinions, etc. We the rapid developments in convolutional neural networks, long-short-term memory machines, and recurrent neural networks – several methods are evolved for document classification.

The challenge of document classification is primitively categorizing based on vocabulary using word embedding methods. Still, new challenges are faced with the classification of documents based on the sentences emanating from the sense of sentences. Word2vec and seq2seq methodologies are rudimentary for sentence-based modeling and sentence-based classification of documents. The advanced version of transforms and BERT can only do the concepts of sentence modeling and sentence-based classification of documents. Nevertheless, methods are advanced into document classification with a sense of sentences. An attention network can do some scratch work on the documents before the documents are input into the corpus of bots to engage in the answering or informal sessions. Practically hierarchical attention networks have achieved 75.8% of accuracy so far on Yahoos' Answer data sets and 71% accuracy on the Yelp'15 datasets.

Further many machines for bots are developed using Convolutional Neural Networks to gain an accuracy of 81.5% on the movie review datasets and 85% accuracy on customer review data sets which support businesses in competitive intelligence. The benchmark datasets from UCA, Kaggle, etc. Support the average number of words that any method achieves 83% accuracy in categorizing documents based on word-topic modeling.



Convolution Neural Network with Custom Attention Layer



Recurrent Neural Network with Custom Attention Layer

Fig. 6: Framework for Document Classification

The state-of-art document classification methods are attention networks, whereas the existing approaches are based on neural networks and statistical methods. The advent of CNNs, LSTM, RNN, and HAN ushered in the concept of document classification and has made it easy. A bidirectional RNN with recurrent gated units (GRUs) combined with the attention mechanism, developing the weighted sum of the RNN outputs at each time step, reiterating the process through the attention, and narrowing the input using CNN is a typical formation of HAN. During data preparation for the long document classification, the long documents are broken into mini documents and into sentence components. The processing is initiated sentence-wise individually before processing the entire document. Breaking a long document into chunks of sentences makes the document manageable and improves the ease of understanding the context of the sentence, which is useful for classification. This preprocessing method surpasses all the trials of document classification processes, where the HAN stands conceptually strong with the blend of CNN and RNN.

Datasets

Datasets used in the context of long document classification usually belong to news channels, sports, technology innovations, marketing, and sentimental / opinion studies. To interpret the proposed framework's performance characteristics, specific data sets that project the propensity of processes are selected.

Source	Documents	Classified	Language
BBC	2225	Yes - 5	English
WOS	46985	Yes - 134	English

Documents from BBC News Websites from Kaggle are available in five categories: technology, business, sports, politics, and entertainment. WOS-46985 is the Web of Science Documents collection, which contains 46985 documents with 134 categories and seven parent categories. Compared to other data sets limited to information retrieval and word-level classification, these two datasets support document categorization and text mining [24].

All the documents in the datasets are converted into vector files. Considering the BBC News Website documents from Kaggle, the documents from the 'technology' category are selected for application into the experiment.

The basic preprocessing tasks include removing invalid data and converting all documents' paragraphs into tuples of the input key vector. Lemmatization and elimination of stop words are done on the datasets, creating a new vector. The key vector and new vector are the columns in the candidate data set.

Long Documents Classification

Long document classification is different from text classification or text categorization problems. A long document consists of lengthy paragraphs. Paragraphs are lengthy due to complex sentences forming various kinds of expression, unlike a subject+verb+object or subject+predicate. The contexts in sentences are linked to each other, successively narrating a legal case or a medical history. Therefore all the sentences and more words are conjugated to give the impression of the context. Because the documents are larger, it becomes difficult to maintain the consistency of the contexts due to the large length. Signals are chosen for defining the group of words composed in sentences to simplify the large documents to sustain the quality of the context. In the consensus of Research, many classification algorithms are available that will group the words of a context. In studies related to deep learning, long documents are classified based on the text in the word or the sentence derived into a vector form called

embeddings. These embeddings are vectors of fixed size. Information relevant to the context and the semantics of representation of the context is well defined by these fixed-size vectors. Further, a model is proposed using fixed-size vectors to represent the model and interpret the context of the texts. The best possible category building is done for the contextual groups of word embedding using efficient ground-level classification algorithms [25].

The proposed method maximizes the likelihood of a group of sentences and sentences represented as groups of words. The logarithmic relation is defined between words and the group as a sentence. Further into groups of sentences is evaluated as fact during preprocessing and grouping long document datasets to connect the contextual dependencies. The sentence and their respective word groups are defined as multi-dimensional vectors with logarithmic dependencies. The document vector for each document is represented as sentence vectors.

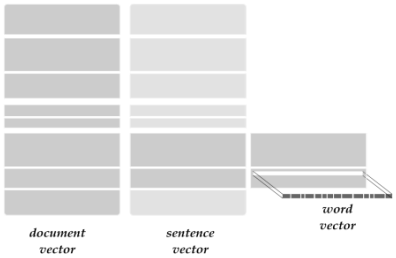


Fig. 7. Vectorization of a long document

From a d-dimensional document space, the document is a vector of sentences, a sentence is a vector of words, and a word vector defines the category of words. Word vectors define the class of sentences; sentence vectors define the class of documents.

4. Experimentation – Framework

The maximum number of features for the experiment on documents data sets from BBC New Website is 6000, and word embedding categories are 128 from NLTK corpus. Stop words and Lemmatization are applied from WordNet API to support preprocessing in the long document data sets. The APIs for Keras, Tokenizer, and Sequence padding for preprocessing, LSTM, and Convlution1D layers are drawn from TensorFlow. Keras has an ideal set of initializers, regularizers, constraints, optimizers, and the ability to define custom layers. To be uniform in the datasets, all the documents are converted to lowercase after filtering stop words and punctuations, and then tokenization is done. This process aims to generate smoothly non-ambiguous encodings among the words to avoid misclassification problems. Sentences are certainly of variable length consisting of words connected in a non-uniform sequence, which is stated as logarithmic dependencies among the sentences and words. However, when generating candidate input sequences for the network, the uniform length shall be maintained, where padding is used for such vectors with limited length. The max length of the sentence size is predefined in the framework. The mean of the lengths for the word vectors is considered and rounded to the nearest integer to determine the max length of the vectors.

Attention Model

In an input sequence of text vectors containing long document dependencies, a set of tokens is labeled sequentially using an index value i. With the help of a neural network, a soft weight is computed called w_i, for each token i in the sequential set, the weighted average. Given a sequence of tokens labeled w_i. Word embedding searches for an equivalent word weight to build the vector with weights; a query-key mechanism is applied to find the appropriate weights. Equivalent word embedding is done on each w_i of the input sequence based on the meaning of words, and a vector is generated. A value vector v_i, is generated of the i^{th} token. Therefore a weighted average of the input sequence with attention is $\sum_i \quad w_i v_i = 1$. An encoder and decoder are applied to configure an attention model. The encoder module transitions the input data, and the decoder module performs another transition on the data from the encoder module. At each iteration, the output m_i is defined as $m_i = F(C_i, m_1, m_2, \ldots m_{i-1})$, where C_i is the semantic encoding information corresponding to the data in the input sequence.

$$C_i = \sum_{j=1}^{T} \quad a_{ij} S(n_j) \ \ldots (1)$$

Considering $S(n_j)$ as the hidden status of the data, at the hidden status after the process by the encoder-module, K quantifies the input data, and the a_{ij} denotes the probability of attention in the input sequence; $a_{ij} = \frac{e^{r_{ij}}}{\sum_{p=1}^{K} \quad e^{r_{ip}}}$ And r_{ij} is the influence evaluation score, h_j is the hidden state of the input j in the encoder module, and s_{i-1} is the output of the decoder module in the previous time step.

The class Attention is built for determining the score, and the attention weights in the networks are computed from softmax in the dense layer. The context vector is built based on the attention weights and features. The features are derived from the LSTM. The hidden states are from each bi-directional RNN to compute the attention weights. A simple concatenation of two RNNs, one processing the sequences left-to-right and another right-

to-left, to get more reliable encodings for the words and build a consistent word vector. The computed context vector is assumed to be a sentence vector for sentence-level modeling. The new context vector is derived in iterations on each document.

The model is trained and tested with all performance evaluation metrics, and a ROC (Repeater Operating Characteristic) curve is drawn, demonstrating the framework's efficiency. The complete network is trained with attention for five epochs with small batches of inputs sized 100 documents of samples. Just five epochs have been derived with less loss and high accuracy from the 20% of validation sets. From an instance of the experiment, 1128 documents are collected as a large sample, and the performance evaluation parameters are derived as follows:

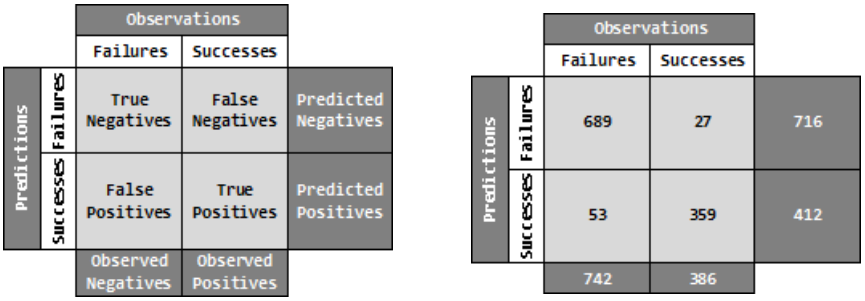


Fig. 8. : Performance Evaluation Parameters of the Framework on a sample size of 1128 documents

True Positive Rate (Sensitivity)	0.930052
True Negative Rate (Specificity)	0.928571
Accuracy	0.969078
False Positive Rate	0.071429
Positive Predictive Value	0.871359
Negative Predictive Value	0.962291

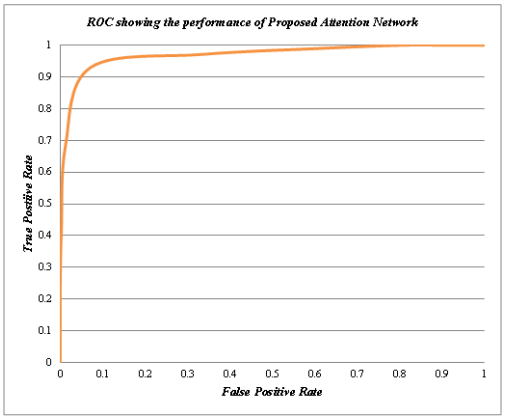


Fig. 9. : Repeater Operating Characteristic curve of the Proposed Attention Network on a sample size of 1128 documents

5. Conclusion

The limitations of CNN-RNN are that they exhaust at a point of genericity due to misclassification of text documents or the documents are generic, which causes imbalance. The concept of sentence modeling resolves using the corpus of sentence sense vectors and classifies according to sense in sentences. The attention mechanism identifies and observes the characteristics of states of input sequences and designates the appropriate classification. The sentences of long documents are vectorized and generalized semantically considering the syntactic constraints. One thousand one hundred twenty-eight documents from the BBC Web News datasets were sampled into 100 sizes, experimented with the proposed model, and arrived at an accuracy of 96.91% approximately.

References

[1] Arar, Moab, Ariel Shamir, and Amit H. Bermano. "Learned queries for efficient local attention." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10841-10852. 2022.

[2] Coquenet, Denis, Clément Chatelain, and Thierry Paquet. "End-to-end handwritten paragraph text recognition using a vertical attention network." IEEE Transactions on Pattern Analysis and Machine Intelligence 45, no. 1 (2022): 508-524.

[3] Dogra, Varun, Sahil Verma, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. "A complete process of text classification system using state-of-the-art NLP models." Computational Intelligence and Neuroscience 2022 (2022).

[4] de Santana Correia, Alana, and Esther Luna Colombini. "Attention, please! A survey of neural attention models in deep learning." Artificial Intelligence Review 55, no. 8 (2022): 6037-6124.

[5] Gnanavel, S., Vinodhini Mani, M. Sreekrishna, R. S. Amshavalli, Yomiyu Reta Gashu, N. Duraimurugan, and Namburi Srinivasa Rao. "Rapid Text Retrieval and Analysis Supporting Latent Dirichlet Allocation Based on Probabilistic Models." Mobile Information Systems 2022 (2022).

[6] Lan, Fei. "Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method." Advances in Multimedia Vol. 2022 (2022).

[7] Liu, Minqian, Lizhao Liu, Junyi Cao, and Qing Du. "Co-attention network with label embedding for text classification." Neurocomputing 471 (2022): 61-69.

[8] Madyatmadja, Evaristus Didik, Bernardo Nugroho Yahya, and Cristofer Wijaya. "Contextual Text Analytics Framework for Citizen Report Classification: A Case Study Using the Indonesian Language." IEEE Access 10 (2022): 31432-31444.

[9] Muaad, Abdullah Y., Hanumanthappa Jayappa Davanagere, D. S. Guru, JV Bibal Benifa, Channabasava Chola, Hussain AlSalman, Abdu H. Gumaei, and Mugahed A. Al-antari. "Arabic Document Classification: Performance Investigation of Preprocessing and Representation Techniques." Mathematical Problems in Engineering 2022 (2022): 1-16.

[10] Bhavani, A., and B. Santhosh Kumar. "A review of state art of text classification algorithms." In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1484-1490. IEEE, 2021.

[11] Dong, Hang, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. "Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialization." Journal of biomedical informatics 116 (2021): 103728.

[12] Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep learning--based text classification: A Comprehensive Review." ACM computing surveys (CSUR) 54, No. 3 (2021): 1-40.

[13] Naseem, Usman, Matloob Khushi, Shah Khalid Khan, Kamran Shaukat, and Mohammad Ali Moni. "A comparative analysis of active learning for biomedical text mining." Applied System Innovation 4, no. 1 (2021): 23.

[14] Pintas, Julliano Trindade, Leandro AF Fernandes, and Ana Cristina Bicharra Garcia. "Feature Selection Methods for Text Classification: A Systematic Literature Review." Artificial Intelligence Review 54, No. 8 (2021): 6149-6200.

[15] Poulos, Jason, and Rafael Valle. "Character-based handwritten text transcription with attention networks." Neural Computing and Applications 33, No. 16 (2021): 10563-10573.

[16] Sun, Qian, Aili Shen, Hiyori Yoshikawa, Chunpeng Ma, Daniel Beck, Tomoya Iwakura, and Timothy Baldwin. "Evaluating Hierarchical Document Categorisation." In Proceedings of The 19th Annual Workshop of the Australasian Language Technology Association, pp. 179-184. 2021.

[17] Wagh, Vedangi, Snehal Khandve, Isha Joshi, Apurva Wani, Geetanjali Kale, and Raviraj Joshi. "Comparative study of long document classification." In TENCON 2021-2021 IEEE Region 10 Conference (TENCON), pp. 732-737. IEEE, 2021.

[18] Bansal, Neha, Arun Sharma, and R. K. Singh. "An Evolving Hybrid Deep Learning Framework for Legal Document Classification." Ingénierie des Systèmes d'Information 24, No. 4 (2019).

[19] Linmei, Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. "Heterogeneous graph attention networks for semi-supervised short text classification." In Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4821-4830. 2019.

[20] Kakade, Arpit, Kunal Dhumal, Sachin Das, Shikhar Jain, and N. M. Ranjan. "A neural network approach for text document classification and semantic text analytics." Journal of Data Mining and Management 2, No. 2 (2017): 1-5.

[21] Semberecki, Piotr, and Henryk Maciejewski. "Deep Learning methods for Subject Text Classification of Articles." In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 357-360. IEEE, 2017.

[22] Dharmadhikari, Shweta C., Maya Ingle, and Parag Kulkarni. "Empirical studies on machine learning based text classification algorithms." Advanced Computing 2, No. 6 (2011): 161.

[23] Baker, L. Douglas, and Andrew Kachites McCallum. "Distributional clustering of words for text classification." In Proceedings of the 21st annual

international ACM SIGIR conference on Research and development in information retrieval, pp. 96-103. 1998.
[24] Kumar, M. R., Chakravarthy, V. D., Ranganatham, T. N., & Ramana, K. (2021). WITHDRAWN: Personal finance transaction index scoring using machine learning model.
[25] Kumar, M. R., & Gunjan, V. K. (2020). Review of machine learning models for credit scoring analysis. *Ingeniería Solidaria*, *16*(1).