International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET2411143

Natural Language Processing and Deep Learning Approaches for Multiclass Document Classifier

Shruti A. Gadewar[1], Prof. P. H. Pawar[2]

[1]ME Aspirant, Department of Computer Science and Engineering, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

[2]Associate Professor, Department of Computer Science and Engineering, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

ARTICLEINFO

Article History :

Accepted: 08 Feb 2024
Published: 22 Feb 2024

Publication Issue :
Volume 11, Issue 1
January-February-2024
Page Number :
278-283

ABSTRACT

With the recent growth of the internet, the volume of data has also increased. A large section of the internet is full of documents, which may contain data, big data, formatted and unformatted data, structured data, and unstructured data. The increase in the amount of this unstructured data results in making it difficult to manage data. As it is difficult to classify the increasing volume of data for various purposes manually, automated classification is required. This paper overviews different approaches to Natural Language Processing and Deep Learning for content-based classification.

Keywords :- Classification, Natural Language Processing, Deep Learning.

I. INTRODUCTION

Data is widely spread over the internet as we are generating so much data every day. Google Maps, uploading documents in various formats from various websites. Content in the form of documents, posts, blogs on various issues, videos all are posted on the internet. The data on which this paper is focusing is document. The very first step of deriving information is to have related data classified and clustered at one place.

Classification is a data mining technique which is a process of classifying or categorizing data on the basis of similar features or attributes. It is a supervised learning technique that uses labelled data to build a model that can predict the class of new, unseen data. Document classification is structurally different from sentence classification. Documents consist of multiple sentences. Sentences may have ambiguous and complex semantic relationships, which makes it difficult to classify documents. In addition, as the number of document categories increases, their management becomes more difficult. Automatic document classification is a supervised machine learning technique that involves determining whether a particular document belongs to a specific category by analyzing the words or terms used in the document and comparing them to those associated with the category [2]. Moreover, as the number of classes

increases, the number of decision boundaries will also increase. Therefore, it will be difficult for the algorithm to solve the problem. This is particularly difficult if the data is imbalanced[1].

A. Methodologies

In a study conducted with the aim of increasing classification success rate, it was emphasized that reducing the size of the feature vector is important for increasing the success of the model. In the study, size reduction algorithms were divided into feature selection and feature extraction algorithms [3]. Feature extraction algorithms can be divided into two types: linear and nonlinear algorithms [4]. These algorithms perform data transformation using optimization techniques. The most important method is Principal Component Analysis (PCA) [5], which produces new features.

Developments in deep learning methods have resulted in developments in the field of text classification. Experimental studies have been carried out with the aim of increasing the success of classification models, especially with the emergence of the BERT model. In one study, the authors reported that methods using attention mechanisms such as BERT have the ability to capture contextual information present in the document. The use of WordNet in future studies was also suggested by the authors, as the vocabulary graph can provide useful global information for BERT [6]. Therefore, it can be concluded that the use of WordNet could be beneficial. In some recent studies, the WordNet lexical ontology and BERT language model were used together to perform document classification, where the role of WordNet was as a source of semantic knowledge, such as with respect to word embeddings, e.g., path2vec and wnet2vec, while that of BERT was to extract the local feature information of the documents and to classify them [6,7].

This paper is discussing different techniques and methodologies that are used to classify the documents and label all on the basis of content in the documents, using Natural Language Processing and Deep Learning techniques which is subset of Artificial Intelligence.

II. MULTI-CLASS VS MULTI-LABEL TEXT CLASSIFICATION

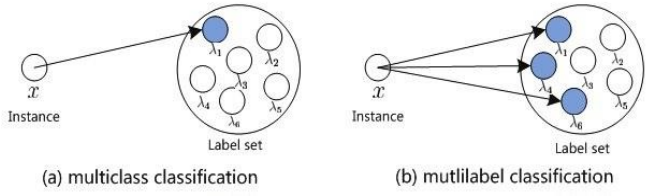

Fig. 1 : Supervised Learning

Multi-class text classification is the process of classifying text into more than two classes or categories. Instances of texts can only be categorized under one label at a time. In other words, the data instances are mutually exclusive. For example, a classification of new articles can be given. The news can be classified into various categories such as sports, politics, weather, entertainment, business, etc.

On the other hand, multi-label categorization assigns a set of target labels to every sample and the data instances are not mutually exclusive. As an example, we can take a model predicting movie genres. One movie can fall under more than one category out of a pool of labels such as romantic, comedy, horror, thriller, and drama.

III. MULTICLASS DOCUMENT CLASSIFICATION TECHNOLOGIES

A. WordNet

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser.

WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

A document can be considered as an ordered conglomeration of words. We start with n documents and with two existing groups of words to be used for classification. Call them w0G = {w1G,...,wnG} and w0B = {w1B,...,wnB}. We can then proceed for classification. Pick the document di = {wi1, wi2,...,win}.

For each word in the document we have to calculate the distances from the words of w0G and w0B. We consider the proportion of classification of words to each group. We prefix {ϵ1, ϵ2} in such a way that pA > ϵ1 we classify to group 1; pB > ϵ2 we classify to group 2, anything in between we fail to classify.

Each word from the WordNet are taken and its distance from categories A is compared with the distance from category B. The distance between two words is considered in terms of number of nodes (intermediate words) between them.

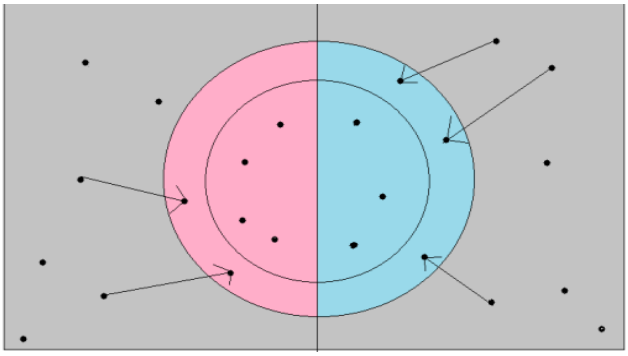

Figure 1: Inserting words into the two sets- good and bad from the WordNet.

If the distance from the category A is greater than the distance from category B then the word which has been taken from the WordNet will be a similar sounding word of category B and thus it will get appended to category B otherwise it gets appended to category A[8].

The pink colour represents good set, the blue colour represents bad set and the grey represents the words of the WordNet which are not present in either of the two sets[8].

B. Text- GCN:-

TextGCN is a GCN-based text classification model that uses a large text graph based on the whole corpus. To understand the concept properly, we first explore the GCN process.[9]

TextGCN constructs a large corpus-level graph but with textual information, documents and words as nodes so it can model the global word-document co-occurrence. The constructed graph includes documents and words nodes from training sets and test sets. TextGCN aims to model the global word-document occurrence with two major edges: 1) word-word edge: calculated by co-occurrence information point-wise mutual information(PMI) , 2) document-word edge: TF-IDF. One-hot vectors are fed into a two-layer GCN model to jointly learn the embedding for the documents and words. The representations on the document nodes in the training set train the classification model while those in the test set are used for prediction.[9]

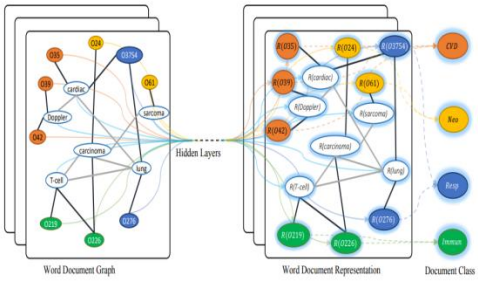

Figure 3: Schematic of Text GCN. Example taken from Ohsumed corpus

C. Bi-LSTM:-

A bidirectional LSTM, often known as a Bi-LSTM, is a sequence processing model that consists of two LSTMs, the first model takes the input as it is, and the second model takes a backward direction copy of the sequence.

This special architecture of Bi-LSTM effectively increases the quantity of data available to the network, giving the algorithm better context.

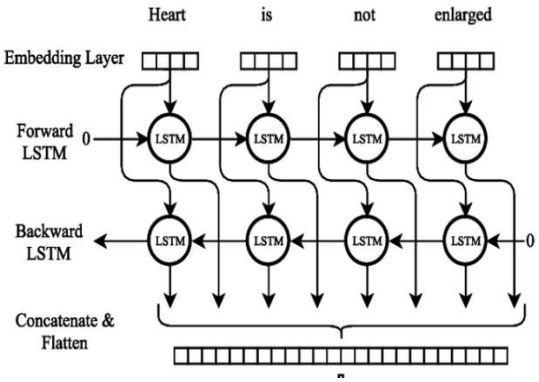

Figure 4:- Architecture of Bi-LSTM

As you can see from the above architecture, there are two LSTM models— one working in a forward direction, and the other in the reverse direction. Because of this, biLSTMs enable additional training by passing the text sequence twice. Therefore, the biLSTM model completes additional training on a given dataset than LSTM, which helps to offer better predictions.

D. Word2Vec:-

Word2Vec is a popular algorithm used for natural language processing and text classification. It is a neural network-based approach that learns distributed representations (also called embedding's) of words from a large corpus of text. These embedding's capture the semantic and syntactic relationships between terms, which can be used to improve text classification accuracy.

Word2Vec can be a powerful tool for text classification, primarily when combined with other machine learning techniques. However, it is vital to remember that there may be better approaches than this, and different algorithms like BERT or transformers may outperform them in some cases. The Word2vec method has two different learning models. These are CBOW (continuous bag-of-words model) and skip-gram (continuous skip-gram model) [10]. The architectures of these models are illustrated in Figure 5

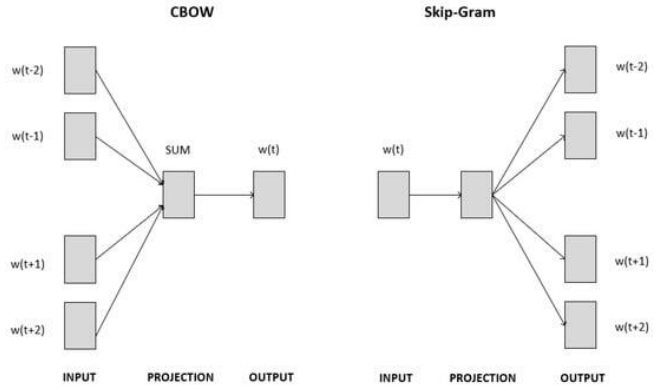

Figure 5:- Architecture of Word2Vec models: CBOW and skip-gram.

E. BERT:-

Transformers were first introduced by Google in 2017. At the time of their introduction, language models primarily used recurrent neural networks (RNN) and convolutional neural networks (CNN) to handle NLP tasks.

Although these models are competent, the Transformer is considered a significant improvement because it doesn't require sequences of data to be processed in any fixed order, whereas RNNs and CNNs do. Because Transformers can process data in any order, they enable training on larger amounts of data than ever was possible before their existence. This, in turn, facilitated the creation of pre-trained models like BERT, which was trained on massive amounts of language data prior to its release. In 2018, Google introduced and open-sourced BERT[12].

The goal of any given NLP technique is to understand human language as it is spoken naturally. In BERT's case, this typically means predicting a word in a blank. To do this, models typically need to train using a large repository of specialized, labeled training data. This necessitates laborious manual data labeling by teams of linguists.

The transformer is the part of the model that gives BERT its increased capacity for understanding context and ambiguity in language.

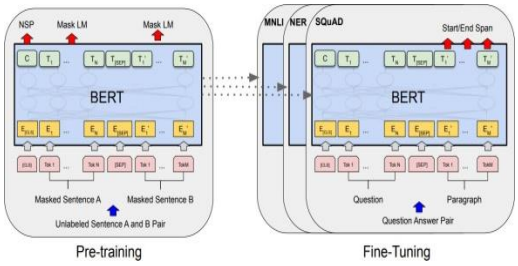

Figure 6:- BERT architecture of pretraining and fine-tuning tasks

The transformer does this by processing any given word in relation to all other words in a sentence, rather than processing them one at a time.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have studied some of the natural language processing and deep learning approaches for multiclass document classification.

From the study, we can conclude that the using WordNet in feature extraction reduces the dimensionality problem by avoiding the repetition of word as well as the occurance of words with same meaning. Whereas Text-GCN is a Graphical Convolutional Network - based text classification model that uses a large text graph based on the whole corpus With the development of GNN, some graph based classification models are gradually emerging and achieved state-of-the-art results on several mainstream datasets. However, Text-GCN has the disadvantages of high memory consumption and lack of support online training. While Bi-LSTM effectively increases the quantity of data available to the network, giving the algorithm better context. However, BERT was more successful than the other classical methods, because it provides deeper learning.

In future works, we aim to extend these studies using other BERT models such as ALBERT, RoBERTa, XLNet, etc., with WordNet and various multi-class imbalanced datasets. Furthermore, since the Graph Convolutional Network (GCN) has recently achieved successful results in text classification, it will be used together with and independently of BERT in order to measure its success on multi-class imbalanced datasets.

REFERENCES

[1]. Ilkay Yelmen, Ali Gunes, and Metin Zontul on "Multi-Class Document Classification Using Lexical Ontology-Based Deep Learning" Appl. Sci. 2023, 13(10), 6139; https://doi.org/10.3390/app13106139

[2]. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. Artif. Intell. Rev. 2019, 52, 273–292. [Google Scholar] [CrossRef]

[3]. Kumbhar, P.; Mali, M.A. Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification. Int. J. Sci. Res. 2016, 5, 1267–1275. [Google Scholar]

[4]. Mwadulo, M.W. A Review on Feature Selection Methods for Classification Tasks. Int. J. Comput. Appl. Technol. Res. 2016, 5, 395–402. [Google Scholar]

[5]. Zhang, T.; Yang, B. Big data dimension reduction using PCA. In Proceedings of the 2016 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, 18–20 November 2016; pp. 152–157. [Google Scholar] [CrossRef]

[6]. Lu, Z.; Du, P.; Nie, J.Y. VGCN-BERT: Augmenting BERT with graph embedding for text classification. In Advances in Information Retrieval, Proceedings of the 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 14–17 April 2020; Springer:

Berlin/Heidelberg, Germany, 2020; pp. 369–382. [Google Scholar] [CrossRef]

[7]. Barbouch, M.; Verberne, S.; Verhoef, T. WN-BERT: Integrating WordNet and BERT for Lexical Semantics in Natural Language Understanding. Comput. Linguist. Neth. J. 2021, 11, 105–124. [Google Scholar]

[8]. Koushiki Sarkar and Ritwika Law on "A Novel Approach to Document Classification using WordNet" arXiv:1510.02755 [cs.IR] or arXiv:1510.02755v2 [cs.IR] for this version) https://doi.org/10.48550/arXiv.1510.02755

[9]. Kunze Wang, Soyeon Caren Han, Josiah Poon on "InducT-GCN: Inductive Graph Convolutional Networks for Text Classification" arXiv:2206.00265 [cs.CL] (or arXiv:2206.00265v1 [cs.CL] for this version) https://doi.org/10.48550/arXiv.2206.00265

[10]. Ren, Y.; Wang, R.; Ji, D. A topic-enhanced word embedding for twitter sentiment classification. Inf. Sci. 2016, 369, 188–198. [Google Scholar] [CrossRef]

[11]. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1188–1196. [Google Scholar]

[12]. Nozza, D.; Bianchi, F.; Hovy, D. What the [mask]? making sense of language-specific BERT models. arXiv 2020, arXiv:2003.02912. [Google Scholar]

Cite this article as :

Shruti A. Gadewar, Prof. P. H. Pawar, " Natural Language Processing and Deep Learning Approaches for Multiclass Document Classifier, International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 11, Issue 1, pp.278-283, January-February-2024. Available at doi : https://doi.org/10.32628/IJSRSET2411143