JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Study on Text Mining Using Natural Language Processing and Artificial Intelligence Algorithms

[1]Sowmya N, [2]Veena V, [3]Laxmidevi H M, [4]Manjunath R
[1,2,3]Assistant Professor, Department of CSE, R R Institute of Technology, Bengaluru, Karnataka,
[4]Professor, Department of CSE, R R Institute of Technology, Bengaluru, Karnataka,

Abstract—Natural Language Processing (NLP) is a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks. Classification of text and recognition is considered as a fundamental research area in the field of natural language processing, which is a discipline that merges computer, mathematics, and linguistic experience. In the context of the big data era, how to effectively classify text information in the face of a sea of text-based data is the focus of current research. This paper describes the theoretical knowledge of text classification concepts, text representation methods and text classifiers. Firstly, the basic concepts of text classification and the classification process are introduced. Then the model structures of convolutional and recurrent neural networks and their variants are introduced, followed by the structure and implementation principles of two classical word embedding models, Word2vec and BERT**.**

Index Terms: **Natural Language Processing, Text Classification, Models, Word2vec, BERT**

I. Introduction

The field of natural language processing (NLP) combines linguistics, computer science, and machine learning. NLP is all about teaching computers to comprehend and produce human language. This area concentrates on communication between computers and humans in natural language. NLP techniques are used for text-filtering and machine translation as well as voice assistants like Apple's Siri and Amazon's Alexa. The latest advances in machine learning, especially deep learning methods, have significantly benefited natural language processing. There are three sections to the field:

- Speech recognition is the process of turning spoken words into written ones.
- Natural language understanding—a computer’s ability to understand language.
- Natural language generation is the process through which a computer produces natural language.

Due to its intricacy, understanding human language is seen as a challenging undertaking. There are countless potential ways to organize words in a phrase, for example. Furthermore, since words can have several meanings, context is important for the accurate interpretation of phrases. Each language has its own characteristics and ambiguities just take a look at the following newspaper headline “The Pope’s baby steps on gays.” This sentence clearly has two very different interpretations, which is a pretty good example of the challenges in natural language processing. In recent years, domestic research on text classification has developed more rapidly, and the direction of research has shifted more towards combining with deep learning. liu et al. [4] proposed three LSTM-based multi-task learning architectures in 2016, which can explore the information sharing mechanism between different tasks in a text sequence model, and the method performed very well in several experiments.
The techniques Natural Language Processing (NLP) uses to extract data from text are:

- Sentiment Analysis.
- Text Classification.
- Keyword Extraction

II. Type Style and Fonts

A. Text classification

A technique based on machine learning called text classification provides a list of predetermined categories to open-ended text. Text classifiers may be used to organize, arrange, and categorize almost all kinds of text, including files from the web, medical research, and publications. In other words, given a pre-labeled text dataset D = {(D1, y1), (D2, y2), ⋯ , (Dm, ym)}, where Dm represents the mth text in the text dataset, ym ∈ {0,1} k represents the text category label, and k represents the number of categories, text classification takes part of the data in the text dataset D as the training set, learns the potential relationship between text and categories through an algorithmic model[2] and establishes a mapping function f to realize the mapping from text to categories f(Dm) → ym. Text classification and recognition generally depend on theoretical understanding of machine learning, using a dataset with a predefined category as the training set, and mining the relationship[5] between the dataset's features and its corresponding category labels to build a model for classification. Figure 1 shows the primary procedures and steps of automated text classification and recognition.

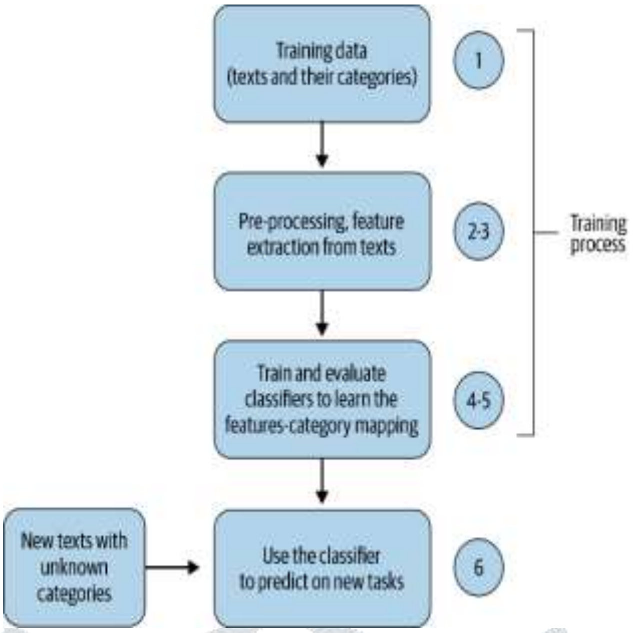
Training data
(texts and their categories)
1
Pre-processing, feature extraction from texts
2-3
Training process
Train and evaluate classifiers to learn the features-category mapping
4-5
New texts with unknown categories
Use the classifier to predict on new tasks
6

Figure 1: Flow chart of text classification

One typically follows these steps when building a text classification system:

1. Collect or create a labelled dataset suitable for the task.
2. Split the dataset into two (training and test) or three parts: training, validation (i.e., development), and test sets, the2n decide on evaluation metric(s).
3. Transform raw text into feature vectors.
4. Train a classifier using the feature vectors and the corresponding labels from the training set.
5. Using the evaluation metric(s) from Step 2, benchmark the model performance on the test set.
6. Deploy the model to serve the real-world use case and monitor its performance.

B. Sentiment Analysis.

Sentiment analysis requires analyzing digital text to identify if the message's emotional tone is good, negative, or neutral. Companies are collecting a lot of text data, such as emails, chat transcripts from customer service, comments on social media, and reviews. A classification challenge in the field of natural language processing is sentiment analysis. Sentiment analysis models turn the views included in spoken or written language data into useful insights, a process also referred to as "opinion mining." It is one of the first problems that many developers who are new to machine learning attempt to address in the field of NLP. This is true because both conventional and deep learning solutions currently exist, and the notion is straightforward and practical.

Sentiment analysis involves determining whether the author or speaker’s feelings are positive, neutral, or negative about a given topic. For instance, you would like to gain a deeper insight into customer sentiment, so you begin looking at customer feedback under purchased products or comments under your company’s post on any social media platform. You would like to know if the customer is pleased with your services, neutral, or if he/she has any complaints, meaning whether the customer has a neutral, positive or negative sentiment regarding your products, services or actions. Figuring this out is called sentiment analysis.

C. Key Extraction

The method of extracting key information from a series of paragraphs or texts is known as keyword extraction. Automated text input may be processed to extract the most essential words and phrases using keyword extraction. It is a text analysis technique that automatically extracts the most crucial words and sentences from a page. It helps in summarizing a work's substance and identifying the main issues raised. In order for machines to understand and determine human language, machine learning artificial intelligence (AI) and natural language processing (NLP)[5] are utilized in keyword extraction. It is used to extract keywords from a variety of sources, such as traditional papers and business reports, comments posted on social media, internet forums and reviews, news articles, and more.

One may easily find the most crucial terms and phrases in huge datasets by using keyword extraction. Additionally, these words and phrases might help you understand the topics that customers are talking about. Businesses need automated keyword extraction to help them process and analyze customer data more effectively since more than 80% of the data people produce every day is unstructured, meaning it is not organized in a certain way and is thus extremely challenging to review and process. Let's say someone want to browse through thousands of internet product reviews. The ability to swiftly filter through a lot of data and extract the terms that best define each review is known as keyword extraction. As a consequence, company can quickly and simply determine what topics are being discussed by customers the most, saving employees many hours of manual processing.

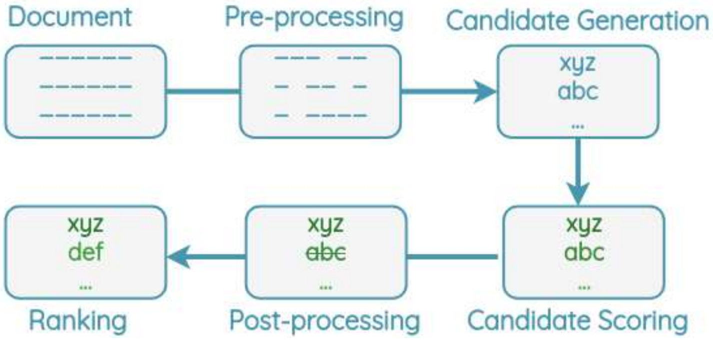

Figure 2: Key Extraction Model

III. PROPOSED SYSTEM

According to the enormous expansion of data brought on by the popularity of the Internet, machine learning models must use a lot of resources, frequently produce unsatisfactory results, and are eventually removed. Since Hinton [10] and others proposed the deep neural network training method in 2006, deep learning technology has advanced quickly, producing ground-breaking results in the fields of image processing and natural language processing. Text classification methods based on deep learning have also begun to draw the attention of academic researchers.
The two traditional deep learning models, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), are frequently employed as classifiers for text categorization tasks.

A. CNN-based text classification models

Convolutional neural networks are efficient at capturing specific features, were originally employed extensively in image processing, and are also fiercely competitive in the field of natural language processing. The output of the upper layer is frequently connected to all the inputs of the lower layer in fully connected neural networks, eventually forming a dense interaction structure that is very likely to cause parameter explosion and cause the model to converge at an extremely slow rate. In contrast, in a convolutional neural network, each output node in the convolutional layer is only partially connected to the previous layer, forming a local self-organized map. CNN model is trained with 2 layers one is convolution and other is softmax layer. These two layers are on top of word embedding layer. With little tuning of parameters we achieve significant change in overall accuracy of model. In contrast, in a convolutional neural network, each output node in the convolutional layer is only connected to some of the nodes in the previous layer, forming a local perceptual field, reducing the number of model parameters and capturing local features; secondly, the parameters are shared, meaning that different layers of the same model share the weight parameters and no longer need to update the weights for each location, greatly accelerating the model operation. Convolutional neural networks can automatically extract and merge N-gram level features from text to get multi-level local semantic information in the field of text categorization due to the sparse interaction property. In order to do text categorization, CNNs frequently have the following four layers.

(1) Input layer: N × K word vector matrix, where N is the total number of words and K is the word vector dimension.

(2) Convolutional layer: The convolutional layer produces feature information with varying granularities by convolutioning the input matrix with a number of convolutional kernels of various sizes.

(3) Pooling layer: By deleting unnecessary features and adding crucial information, the output of the convolutional layer is pooled to create a fixed-length representation of the text, therefore simplifying the model and accelerating convergence. [9]; typical pooling procedures include of Max Pooling, Average Pooling, Minimum Pooling, etc.

(4) Output layer: Combine with fully connected layer and use Softmax function to complete the classification. Similar to the traditional CNN structure, the Text CNN also consists of four basic layers, namely the input layer, the convolutional layer, the pooling layer and the output layer, and its structure is shown in Figure 3. The input layer consists of two main channels, both using Word2vec pre-trained word vectors as the word embedding layer, but the training method is different. One of the channels directly initialises the non-occurring words at random, softmax layer are combined to perform classification. As a pioneer, Text CNN uses the powerful local feature extraction ability of CNN to obtain the relationship of adjacent words in text, and further enriches the semantic features of text by means of multiple convolutional kernels, triggering a boom in the application of CNN to text classification tasks.

(5) Fully connected neural network: A fully connected neural network is made up of a number of layers, each of which connects every neuron in one layer to every other layer. The main benefit of completely linked networks is that they are "structure agnostic," meaning that no particular assumptions about the input are required.

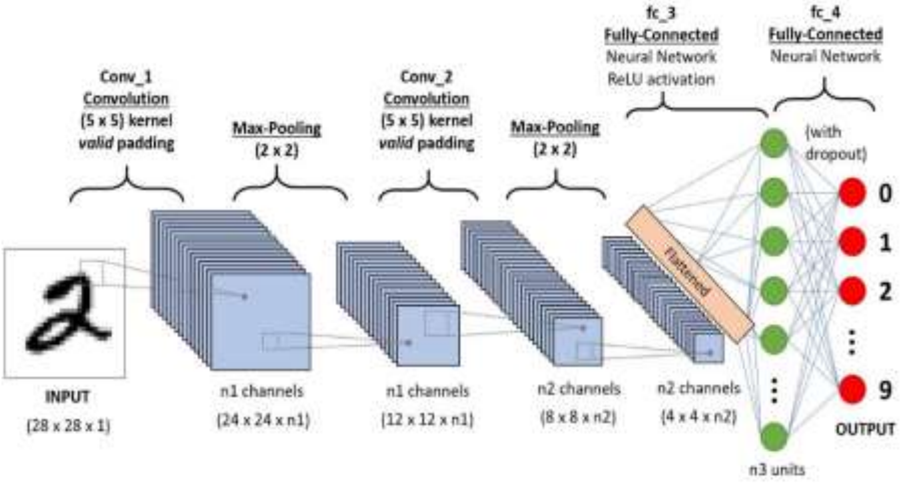

Figure 3: Text CNN model structure

B. **RNN-based text classification models**

As conventional feedforward neural networks cannot successfully exploit historical data, recurrent neural networks (RNN) were created to address these issues. Recurrent neural networks can analyse lengthy sequences because of their special chain structure, which allows them to compute while taking into consideration both the current input and the preceding hidden layer's output. Figure 3 depicts its typical structure.

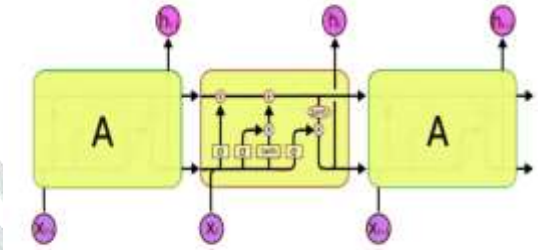

Figure 4: RNN model structure

IV. METHODOLOGY

Textual representation is a technique for representing textual material digitally. Deep learning-based word embedding models have supplanted traditional text representation techniques, which suffer from excessive computational effort, a lack of semantic information, and substantial resource waste. The pre-training model BERT and the widely used Word2vec word embedding model are the main topics of this research.

A. Word2vec

The natural language processing, or NLP, tool Word2vec was introduced in 2013. With the help of a huge text corpus, the word2vec technique employs a neural network model to learn word connections. Once trained, a model like this may identify terms that are similar or propose new words to complete a phrase. One of the most popular word embedding models, Word2vec was developed by Google and comprises of the two primary components, CBOW and skip-gram, which are seen in Figure 4 below. Figure 5 demonstrates that both CBOW and skip-gram are actually constructed as shallow neural networks, each of which has an input layer, a projection layer (implicit layer), and an output layer. CBOW predicts the current word in context, while skip-gram predicts the context in terms of the current word. The sliding window size in the graphic is 2, and the current word is labelled as "wt." Other words in the figure include "wt-1," “wt-2” etc.

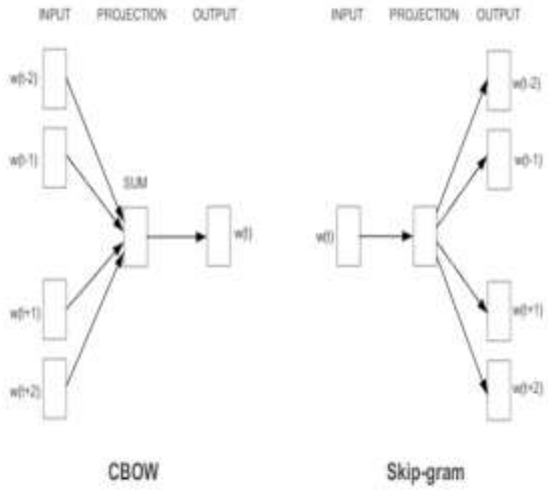

Figure 5 : Word2vec model architecture

B. BERT

BERT uses the surrounding text to provide context in order to assist computers understand the meaning of ambiguous words in text. With the use of question and answer datasets, the BERT framework may be adjusted after being pre-trained on text from Wikipedia. In an unsupervised way of language modeling, BERT leverages the idea of pre-training the model on a bigger dataset. The context of the input sentence may be understood by a pre-trained model using a bigger dataset. The model may fine-tune the task-specific supervised dataset after pre-training to get good results.

In this step, we can use any of two strategies: feature-based or fine-tuned. Elmo utilizes the feature-based model since it is there that the model architecture will be task-specific. Different models and trained models for language representations will be used for each challenge. BERT makes advantage of the idea of fine-tuning. Its utilization of bidirectional layers of transformer encoders for language comprehension gave rise to the moniker BERT. We must be aware that BERT is capable of comprehending a word in its entirety. The phrase that comes before and after the word will be analyzed by BERT to determine their connection.

BERT works with the help of the below steps:

Step 1: The BERT was developed unambiguously for use on increased word counts and has a large quantity of training data. BERT has improved its fluency in several other languages, including English, thanks to the huge informational databases. When utilizing a larger dataset, BERT training takes longer. Training BERT is possible because to the transformer architecture, and the training process may be accelerated with Tensor Processing Units.

Step 2:The second step is the Masked Language Model (MLM), which enables bidirectional text learning. We can do this by concealing a word in a phrase and making BERT use the term in both directions. In order to forecast the concealed word, we might attempt to comprehend the words that come before and after it as shown in the below figure.

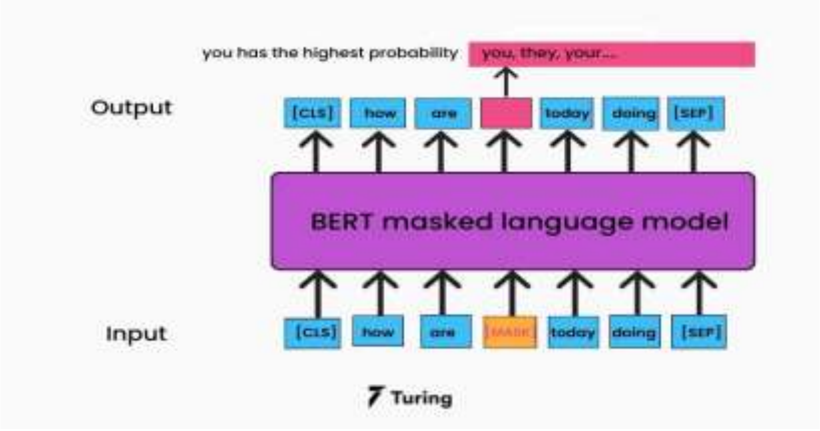

Figure 6: BERT model framework

V. CONCLUSION

With the Continual development of internet technology, the twenty-first century has evolved into an era of big data, with the internet serving as the primary source. Text classification methods are beginning to be the subject of research and debate among academics as hot challenge that has to be solved is how to swiftly acquire the necessary data from the hundreds of millions of enormous data. Text categorization techniques have been regularly introduced and enhanced as a core job of natural language processing. This article discusses the theories and particular techniques linked to text classification. Text classification has advanced significantly in recent years, and new concepts including word embedding models, pre-training models, and attention processes have been put

REFERENCES

[1] Zou Y, Zhao T D, Qian W B. An improved model for spam user identification[P]. DEStech Transactions on Computer Science and Engineering, 2018.

[2] Nawangsari R P, Kusumaningrum R, Wibowo A. Word2Vec for indonesian sentiment analysis towards hotel reviews: an evaluation study [J]. Procedia Computer Science, 2019, 157: 360-366.

[3] Feng G., Zhang X., Liu S. Research on Chinese text classification based on CapsNet [J]. Data Analysis and Knowledge Discovery, 2019, 2(12).

[4] Zhao Q., Du Y. H., Lu T. L., et al. A text similarity analysis algorithm based on capsule-BiGRU [J]. Computer Engineering and Applications, 2020, 11(27):1-9.

[5] Lei K., Fu Q., Yang M., et al. Tag Recommendation by Text Classification with Attention-Based Capsule Network [J]. Neurocomputing, 2020.

[6] Wang Shuang. Research on automatic text classification based on machine learning [D]. University of Electronic Science and Technology, 2020.

[7] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [J]. ar Xiv preprint ar Xiv:1810.04805, 2018.

[8] Liu Wanjun, Liang Xuejian, Qu Haicheng. Study on the learning performance of convolutional neural networks with different pooling models [J]. Chinese Journal of Graphical Graphics, 2016, 21(9):1178-1190.

[9] Zhang Q, Gao T Z, Liu X Y, et al. Public environment emotion prediction model using LSTM network[J]. Sustainability, 2020, 12(4):1-16.

[10] Chung J, Gulcehre C, Cho K H , et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [J]. Eprint Arxiv, 2014.

[11] Meenakshi, Neha Goutam, Raghavendra S., Manjunath R., Santosh Kumar J , "Development Of Framework To Recognize Akhara-Muni Character Using Ann", Turkish Online Journal of Qualitative Inquiry, ISSN: 1309-6591 , Vol. 12 No. 9, pp. 1528-1533, 2021.