CE807 – Assignment 1 - Interim Practical Text Analytics and Report

1901423

Introduction

Text classification is an important aspect of language processing and has led to the rise of the many interventions. A lot of studies have explored the neural networks, but only limited number has been able to provide the most flexible neural networks. There are various techniques of classifying text in various data, documents, and research articles such as Deep learning, neutral networks and Recurrent Neural Network.

There is a need to classify documents in relevant forms according to their text data and should be interested in many practical reasons. We have the Domain-specific SKOS vocabularies, advanced methods and the traditional methods. A deeper analysis helps to determine the best method of text classification.

The advanced methods of text classification are the automated method of text classification. The advance method are considered as the most vital and used in managing and processing huge amounts of data in digital forms [5]. The basic text classification methods are the non-digital methods that are archaic paper sources like the magazines, books among others. Finding the suitable method of classifying text is a big challenge due to technological changes and increased complexity of data [5]. This hence calls for the adoption of suitable advance d method as it is able to classify the complex and large amounts of documents electronically [16].

The Linked Open Data Cloud is a system consisting of the metadata concerning research articles and documents. The Simple Knowledge Organization (SKOS) in this system are used to describe the semantics of the documents [10]. The Metadata of the document such as the title are used to compare the semantic annotations with the metadata of the document obtained from a deeper analysis of the full texts [2]. This technique is good quality because it is well maintained and crafted by the expert and made easily available for users [1]. The challenge with this classification method is that the successful use of the SKOS and the annotations to provide clear information and enhance the text classification.

Graph Neural Network

The Graph Neural Network is the method that has recently received attention globally as it is used in a regular grid structure. This system is used to encode the syntactic structure of a sentence [5].The GCN achieved the best classification results on various benchmark graph datasets. This system also explores the various roles in the relation classification and the semantic role labeling [9].

This method majors in use of the tittle of the text. It is proven to reach this system regards the texts and word as nodes (heterogeneously embedded) and hence does not need inter-document relations [2]. Evidence shows that the improvement of the GCN over the state-of-the-art methods became prominent as it lowers the percentage of training data. The challenges with this method are the need for more knowledge and skills to perform it.

Graph Convolutional Network

The Graph Convolutional Network is a method that entirely operates on the graph and introduces vector of nodes which are embedded based on the neighborhood traits. It is a multilayer neural network as it incorporates the neighbor information by piling the multiple GCN layers [1]. This method is useful as it has reduced errors and has more accuracy as compared to other methods. It is also very easy to use and understand. The challenge with this system is the need for proper training knowledge and skills. It is also a bit expensive method. This method is complex hence understanding it is a bit hard task.

Text Graph Convolutional Network

The Text Graph Convolutional network is building the heterogeneous text graph which comprises of the word nodes. This allows the global word concurrence to be modeled as well as graph convolution can be adopted. This method measures the word associations to calculate the weight of the two nodes [11].

It creates a heterogeneous graph for the whole corpus and turns all the documents to the node classification question. It is a generative method which extends to auto encoder of literature [8]. This method is useful as it can capture the whole word concurrence information and uses limited labeled articles well. This method is flexible making classification achievable.

The prototypical network is proposed for the shot classifications whereby a classifier must be used to accept the new classes skipped or not viewed in the real training. Approach those are basic such as retraining is used in this aspect [11]. It is proven that humans can perform this type of classification with only a single example offered with proper accuracy [3].

This approach uses sampling patches called episodes and mimicry in the training to help in creating a good test environment and gives room for generalization [6]. This approach is a simple and efficient method that enhances text classification. The challenge with this method is that it is prone to errors hence requires a lot of care [19].

Recurrent Neural Networks

Recurrent Neural Network is an artificial neural network and it involves connection of nodes from a given directed graph along with a given pattern allowing for the exhibition of dynamic temporal behavior for a given sequence [9]. RNN comprise of a sequence of neural network blocks linked like a chain and each passes a message to the successor [7]. This method can be enhanced by use of external embedded knowledge as it integrates the new information referred to as the lexical and semantic elements [18].

The disadvantage of this method is that it is complex to understand however it is beautifully designed and overcomes the traditional neural networks dealing with text sequences [19]. It has high accuracy and consistency [2]. The challenge with this system is the complexity and a need for training which is costly.

Hierarchical Attention Network (HAN)

This method is used to provide guide for document classification, the preprocessing is done by magnificent embedded sequence. This method is accurate but does not provides consistency. It is more flexible and uses limited data giving room for generalization [4]. The challenge with this system is the inconsistency of the data.

Deep Learning

Deep learning is used extensively in the natural language learning as it is suitable for learning the complex structure of semantic proxy and sentences. Currently deep learning is used to capture the hard model linguistic concepts like the negations and mixed sentiments[15].The deep learning has various advantages over the other logarithms [13].Deep learning is more flexible than other methods [16].This method allows for building of models with flexible outputs.

This helps in allowing for the development of models that help in understanding the complex linguistic structures useful in developing NLP applications like the translations, chat box and text to seek applications [20].This method requires less domain knowledge hence advantageous as it uses the natural language [14]. Additionally, this method is easy to train on a case where new data comes in, updating, this makes this process a preferred simple classification method [3].

A simple, non-parametric and scalable approach of short text classification is highly proposed. This approach leverages the well-studied and scalable retrieval framework and mimics humans in the labeling process of a short text [12]. This approach identifies the topical indicative word from a given text as the question and searches the matching query words [17].

The short text classification search vote and the query search offers accuracy in comparable data and clarity [19]. This classification method is flexible and has a great potential of achieving better results. It enhances the short text enrichment and relevant ranking techniques specific for short texts [20].The disadvantage with this system is the lack of accuracy; a lot care must be observed to minimize errors.

Traditional Learning Technique (Machine Learning)

The traditional learning methods include the logistic regression, decision tree Scale Invariant Feature Transform and Speeded up Robust Features which connects all the neurons of a sentence to show the connections and how they are supposed to model [21]. The traditional methods are the basic of text classification that reveals why the results are [16]. The disadvantage of these methods is that they are inaccurate and are not suitable for large data [8].

Deep learning is the most preferred choice used as the alternative classification because it accommodates both the traditional and modern methods to achieve the best text classification. I learned that the best text classification method should be simple, accurate, flexible and should give room for generalization. Deep learning executes features by itself. In this method, an algorithm scans the data to identify features which relates and combines them to enhance faster learning. Increased training is a way to overcome the challenges in deep learning and make it the best alternative.

REFERENCES

[1] Abadi, M, Agarwal, A, .B. Paul, Brevdo,E, Chen, Z. Citro, C, Corrado, G S,D. Andy, D.Jeffrey, D, M, et al. (2016). *Tensorflow:Large-scale machine learning on heterogeneous distributed systems.* arXiv. arXiv:1603.04467.

[2] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Verydeep convolutional networks for text classification. In ECACL, Vol. 1. 1107–1116.EMNLP. 1746–1751.

[3] Mai, F. Galke. L. & Ansgar, S. (2018). Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text. Joint Conference on Digital Libraries. (JCDL); FortWorth, TX, USA,pp169-178 doi.10.1145/3197026.3197039.

[4] Hu, X Sun,N Zhang, Cand Chua. T,S(2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In CIKM, pages 919–928,

[5] Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: *Literature review and current trends*. webology, 12(2).

[6] Koch, G. (2015). Siamese neural networks for one-shot image recognition. *ICML Deep Learning Workshop.*

[7] Liang Y. Chengsheng M. & Yuan L. (2019). *Graph Convolutional Networks for Text Classification.* AAAI7370-7377.

[8] Li F, Wang C, Liu X, et al (2018) A Composite Model of Wound Segmentation Based on Traditional Methods and Deep Neural Networks. *Comput Intell Neurosci* 2018:1–12. https://doi.org/10.1155/2018/4149103

[9] Lukas G, Florian M, Alan S, Dennis B, and Ansgar S. (2017). *Using Titles vs. Full-text as Source for Automated Semantic Document Annotation.* In K-CAP.9.

[10] Parisotto, E, Ba, J. Lei, and Salakhutdinov, R. (2016) Actor-mimic: Deep multitask and transfer reinforcement learning. *International Conference on Learning Representations* (ICLR).

[11] Rei, M. (2015) *Online representation learning in recurrent neural language models*. arXiv preprint arXiv:1508.03854.

[12] Salimans, T. and Kingma, D P.(2016) Weight normalization:A simple reparameterization to accelerate training of deep neural networks. *In Neural Information Processing Systems* (NIPS).

[13] Saxe, A, McClelland, J, and Ganguli, S. (2014) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations (ICLR*),

[14] Snell, J. Swersky, K. & Zemei, R.S(2017). *Prototypical Networks for Few-shot Learning*. NIPS. 4077-4087.

[15] Veillard, A., Morére, O., Grout, M., & Gruffeille, J. (2018, June). Fast 3D seismic interpretation with unsupervised deep learning: Application to a potash network in the North Sea. In 80th EAGE Conference and Exhibition 2018 (Vol. 2018, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.

[16] Vinyals, O. (2016). Blundell, Charles, Lillicrap, Tim, Wierstra, Daan, et al. *Matching networks for one shot learning. In Neural Information Processing Systems* (NIPS),

[17] Wenjie Z, Liwei W, Junchi Y, Xiangfeng W, and Hongyuan Z (2017) *Deep Extreme Multi-label Learning*. arXiv:1704.03718

[18] Wenpeng Y, Katharina K, Mo Yand Hinrich S. (2017*). Comparative. Study of CNN and RNN for Natural Language Processing.* arXiv:1702.01923.

[19] Xiang Z, Junbo Z, and Yann L. (2015). *Character-level convolutional networks for text classification. I*n NIPS. 649–657.

[20] Yarin G and Zoubin G. (2016). A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In NIPS. 1019–1027.

[21] Yoon K. (2014). Convolutional Neural Networks for Sentence Classification.

Appendix

Specific Research papers used for research (Task1 and 2).

1: Text Classification Basics: 4,5,8,9

2: Graph Neural Network: 2,9,5

3: Graph Convolutional Network: 1

4: Text Graph Convolutional Network: 3,6,8,11,19

5: Recurrent Neural Networks: 2,7,9,18,19

6: Hierarchical Attention Network (HAN): 4,10

7: Deep Learning: 3,4,12,13,15,16,17,19,20

8: Traditional Learning Technique (Machine Learning): 8,16,21