Comparative Study between basic methods and advanced methods for Text Classification

School of Computer Science
and Electronic Engineering
University of Essex
CE807-7-SU : Text Analytics
registration number:2110366

ABSTRACT
Text classification is a very important solution. Which is indispensable in natural language processing, we compared all 6 models including Random Forest Classifier, Support Vector Machine, Logistic Regression, Convolutional Neural Networks, Recurrent Neural Networks, and long short term memory to compare the pros and cons. , including the system model which each model There are different advantages and disadvantages. but if measured with accuracy The most efficient models are CNN and LSTM.

Author Keywords
Text classification; machine learning; deep learning

CCS Concepts
•**Text classification** → *Basic approach; advanced approach; advantage; disadvantage;* •**Basic approach** → *Random Forest Classifier; Support Vector Machine; Logistic Regression;* •**advanced approach** → *Convolutional Neural Networks; Recurrent Neural Networks; Long short-term memory;*

1 INTRODUCTION

Text Classification is the classification of documents to be in the given category. One document may be in more than one category. by converting the text data into numeric format Then do it to predict various models. Most of the problems in the world today of natural language processing are Information Retrieval, Information Filtering, Sentiment Analysis, Recommender Systems, Knowledge Management, and Document Summarization can be solved with text classification. So learning text classification is very important to do natural language processing. There are many studies describing the process and how text classification works in basic and advanced methods of every model But there are very few studies comparing the advantages and limitations of all models. So we will compare all 6models. It is divided into basic approaches for text classification 3 models: Random Forest Classifier, Support Vector Machine, Logistic Regression, and advanced methods approaches for text classification 3 models: Convolutional Neural Networks, Recurrent Neural Networks, long short term memory. to understand the text classification, this research is the most important reason we research the procedure and how it works, including comparing the advantages and limitations of all models related to text classification. This research wants to compare and explain the working principles and advantages and limitations of every model. To decrease the complexity of the text, then we make all texts lowercase with an adaptive capitalization method to lowercase. However, doing so is often problematic in some words.

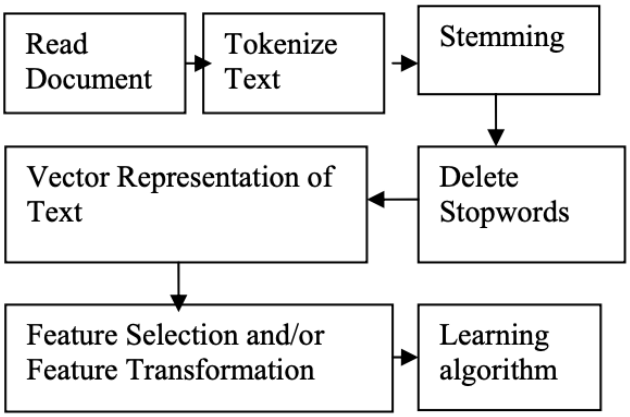

Figure 1. step of cleansing text classification[2]

2 METHODOLOGY

how to make text classification Can start from Tokenization is the taking of words form sentences or texts separate into phrases, words, and symbols that have meaning. after that, we will Parsing which is then the process of specifying the structure of the text. by analyzing the constituent words with the grammar of the language and will eliminate words don't use by Stop Words, which helps to remove words that we find frequently in sentences or documents, but rarely helps to convey meaning. So we have to use the Slang and Abbreviation procedure to solve the problem of slang. It is a word used in informal conversation, and Abbreviation is a shortening of words. This made it very difficult to translate languages. The solution is to convert the unofficial language to the official language. Sometimes we will encounter punctuations or special characters, which can make the model less correct and accurate, so we have to use the Noise Removal method to delete punctuations or special characters. Then we have to do the Spelling Correction because humans don't always write correctly. spell check and correcting spelling mistakes is important. By looking at the context of the word and looking for misspelled words. Then we will use Stemming. it will do a

process of cutting the end of the word coarsely with Heuristic, which works quite well. For most, but not all, English words, stemming reduces form. leaving only the front part of the same words in the same group of words and Lemmatization is the process of converting a word with a list of words in a dictionary, proper grammatical analysis of the language. to eliminate the inflection of words, then we use Word embedding to convert words to numbers. where the result will come out in the form of Vector and do Dimensionality Reduction in order to make model Various models that we will be comparing are as follows.

3 BASIC APPROACHES FOR TEXT CLASSIFICATION

3.1 SUPPORT VECTOR MACHINE

Support Vector Machine is one of the supervised learning algorithms, we will use the number of keywords as the number of dimensions in the vector. while vector is the transformed text of the sentence. It uses the principle of finding the coefficients of the equations to create a line that separates the data group fed into the SVM model. It will try to find the best dividing line. SVM is easy to use for text classification because it can automatically find a good parameter without having to adjust the parameter. Advantages of Support Vector Machine are Support Vector Machine Can be used with non-linear decision boundaries and it is extremely efficient on memory and high dimensional space. In addition, there is a margin for clearly separating classes. However, it is not suitable for large data, works inefficiently when there is noise in the data, and Support Vector Machine doesn't work well with complex classes[1].

3.2 Logistic Regression

Logistic Regression is to do Classification, but the output is actually a Regression, then mapping it to the Sigmoid function will get the data as a class. which can only be 2 classes. The output of Logistic Regression is actually a number, but when used with Sigmoid, the Y-axis is the probability of that class being in the range 0–1. The X-axis is the predicted regression value. If the probability is in the range of 0.0–0.49 then it is not that class such as yes or 1. If the probability is in the range of 0.49–1.00, it is that class such as no or 0. Logistic Regression is Easy to use, easy to interpret train data, and Very accurate in small volumes of train data. Additionally, Classes can be extended with multinomial regression. However, it Can't deal with many features or categorical variables and it Can't solve non-linear problems, and the Possibility of overfitting in case of observations less than the number of features[9].

3.3 Random Forest Classifier

Random Forest generates a model from multiple subsets of Decision Trees, with each model receiving a different data set, which is a subset of the entire data set. When making a prediction, let each Decision Tree make a prediction and compute the prediction result with a vote output that was selected by the Decision Tree the most of the mean from the output of each Decision Tree. Each Decision Tree model in Random Forest is considered a weak model but when we combine each decision tree to make a prediction together, we get a combined model with proficiency and more accurate than a Decision Tree that makes an alone prediction, which Random Forest is very well

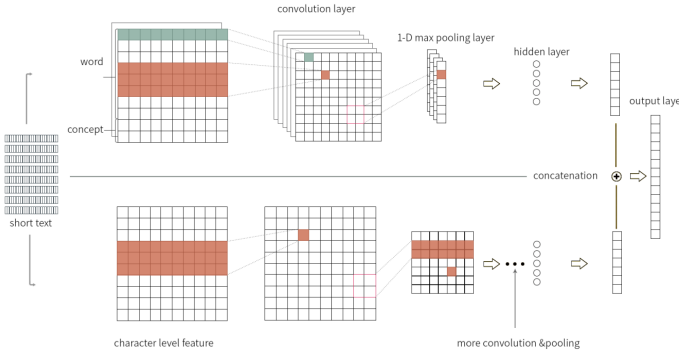

Figure 1: Architecture of the Overall Model

Figure 2. Architecture of Deep Learning Model[10]

suited to text classification with high dimensional noisy in data. The Advantage of Random Forest Classifiers is High prediction accuracy and Able to handle large volumes of data efficiently and the Reduced variance compared to the decision tree. The Disadvantage of the Random Forest Classifier is Train a model takes more time than most any other models in machine learning and Possibility of overfitting easily and It is very difficult to interpret inside the model[7].

4 advanced methods and approaches for text classification

4.1 Convolutional Neural Networks

Convolutional Neural Networks are a popular model of text classification that can be classified by hierarchical document classification. The documents are in matrix format, where the input layer is a vector of word embeddings, while the convolution layer is a vector of word embedding. it is a multi-layer filter. it makes text classification very effective Advantages of Convolutional Neural Networks are Very accurate in short messages, they Can be used with emoji messages, and they can improve model text classification by adding a layer After doing the pooling layer step. The Drawback is Long messages are much more difficult than short messages, they need a lot of vocabulary to train, and big data size. It takes a lot of power to process[3].

4.2 Recurrent Neural Networks

A recurrent neural network is a neural network that takes the output from the previous state as the input. It works similarly to a loop operation, which makes it look like a simple neural network with multiple outputs and connects the outputs to a new neural network. Because RNNs use data from previous networks, it works well with Time Series data. Time Series includes both text data and audio data. Can memorize the results of the text and use them to calculate in the future continuously. Because of the process of sequential information, the more information. if it has many data before it has The more accurate It helps to predict the next word or sentence more accurately. This Advantage, it can process the input as needed and Although the input size is larger, the model size remains the same. Nevertheless, Calculations are very slow due to repeated calculations from previous data. There is a

chance of gradient vanishing. Processing will be very long and very difficult to understand if using relu or tanh[6].

4.3 Long short-term memory

In order to solve the problem of RNN dealing with long sequences of data, the use of Long Short-Term Memory was proposed. it looks like RNN but it is more detailed. RNN is viewed as a Neural Network with simple memory beside it. In order to save the previous hidden state, long short-term memory also has internal memory. But what's better than RNN is that memory can also have a descriptor. When should write, forget, or read is allowed? Working principle of long short term memory First, cell state holds the state of the memory cell in long short term memory. Next, Gate controls the flow of data. This is an analog value that controls when to read, write, or forget. The good point is it can solve vanishing problems The gradient is almost completely gone,-Long short-term memory does not need to adjust parameters because there are many parameters Therefore, it is not complicated to adjust the weight. Nonetheless, It takes a lot of memory to train. It is easy overfit. Takes a very long time to train. Long short-term memory is very sensitive to different random weight initialization, so start with small weight initialization[8].

5 CONCLUSION

In comparing the different models, it is clear that the advanced methods approach for text classification took longer to train the model than the basic approaches for text classification. However, advanced methods approach for text classification are more accurate than basic approaches for text classification and feature extraction can be performed along with classification. while basic approaches for text classification. The feature selection must be done before training. by the best accuracy models are CNN and LSTM, it can be seen that in translating the text from English to French the performance of CNN is as good as RNN[4],[5].

REFERENCES

[1] Atreya Basu, Carolyn Watters, and Michael Author. 2003. Support Vector Machines for Text Categorization. 103. DOI: http://dx.doi.org/10.1109/HICSS.2003.1174243

[2] M Ikonomakis, Sotiris Kotsiantis, and V Tampakas. 2005. Text classification using machine learning techniques. *WSEAS transactions on computers* 4, 8 (2005), 966–974.

[3] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018a. Understanding Convolutional Neural Networks for Text Classification. 56–65. DOI: http://dx.doi.org/10.18653/v1/W18-5408

[4] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018b. Understanding convolutional neural networks for text classification. *arXiv preprint arXiv:1809.08037* (2018).

[5] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

[6] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* (2016).

[7] Tomas Pranckevicius and Virginijus Marcinkevičius. 2017. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing* 5 (01 2017). DOI: http://dx.doi.org/10.22364/bjmc.2017.5.2.05

[8] Winda Kurnia Sari, Dian Palupi Rini, and Reza Firsandaya Malik. 2019. Text Classification Using Long Short-Term Memory with GloVe. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)* 5, 2 (2019), 85–100.

[9] Kanish Shah, Devanshi Sanghvi, and Manan Shah. 2020. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research* 5 (12 2020). DOI:http://dx.doi.org/10.1007/s41133-020-00032-0

[10] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. 2017. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification.. In *IJCAI*, Vol. 350.