Understanding and Responding Text Data with Categories Using Bi-Directional LSTM and NLP

Akshaya K[1] Rakshitha T[2] Swathi J[3] Julia Faith S[4]
[1,2,3]UG Scholar [4]Assistant Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]S. A. Engineering College, Chennai, India

Abstract— The old text classification methods are based on machine learning. It requires a large amount of artificially labelled training data as well as human participation. However, it is common that ignoring the contextual information and the word order information in such a way, and often happen some problems such as data sparseness and latitudinal explosion. With the development of deep learning, many researchers have also been using deep learning in text classification. In our project we investigates the application issue of NLP in text classification by using the Bi-Directional LSTM method. The impact of Bi-Directional LSTM with native LSTM model will also be experimented in our project.

Keywords: Natural Language Processing (NLP), Long Short Term Memory (LSTM), Data Sparseness, Bi-directional LSTM, Latitudinal Explosion

I. INTRODUCTION

Text classification, also known as text categorization, is a classical trouble in natural language processing (NLP), which aims to allot labels or tags to textual units such as sentences, queries, paragraphs, and documents. It has a wide range of applications including question answering, spam detection, sentiment analysis, news group, user intent classification, content moderation, and so on. Text data can come from different sources, including web data, emails, chats, social media, tickets, assurance claims, user reviews, and questions and answers from consumer services, to name a few. Text is an extremely loaded source of information. But extracting insights from text can be demanding and time-consuming, due to its shapeless nature. Text classification can be performed either through manual explanation or by automatic labeling. With the growing scale of text data in commerce applications, automatic text classification is becoming gradually more important. Approaches to regular text classification can be grouped into two categories:

- Rule-based methods
- Machine learning (data-driven) based methods

II. PROPOSED WORK

The proposed model is based on recurrent models to deal with text as chronological information. Long Short-Term Memory (LSTM) models use the information from the preceding status of neurons. More related information can be extracted using a bidirectional LSTM (Bi-LSTM). The data processing flows forward and backward at same time and the output of each LSTM are merged using their sum. Bi-LSTM models have shown to be effective in problems of speech recognition, being a state-of-the-art model to classify sequential data into multiple classes. Documents are sequences of words and our difficulty is also multi-class, therefore Bi-LSTM is a appropriate model. Our architecture it is collected of three main layers with 1000 tokens of input. We followed and used word embedding as an input layer of the Bi-LSTM. This layer transforms each token in a multiply array of 100 dimensions. The recurrent layer has two unknown LSTM, a forward and backward layer model, each with 200 memory blocks and one-cell. The output of this layer uses ReLu activation. The two unknown LSTMs are combined by adding their outputs. The final layer is dense, with 6 output neurons and a Softmax activation function.

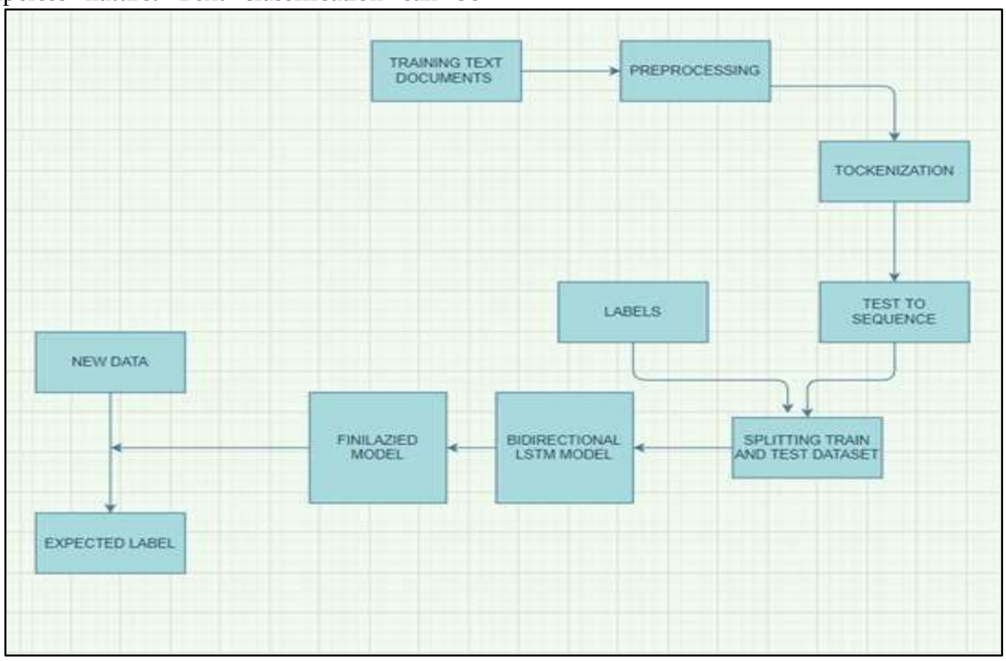

Fig. 1.1: System Architecture

III. RELATED WORK

High accuracy of text classification can be achieved through instantaneous learning of multiple information, such as sequence information and word importance. In this article, a kind of flat neural network called the broad learning system (BLS) is functioning to take two novel learning methods for text classification, including recurrent BLS (R-BLS) and long short-term memory (LSTM)-like architecture: gated BLS (G-BLS). The proposed two methods possess three advantages:

1) Higher precision due to the simultaneous wisdom of multiple information, even compared to deep LSTM that extracts deeper but single information
2) Extensively faster training time due to the noniterative learning in BLS, compared to LSTM and
3) Simple combination with other discriminant information for further improvement. The proposed methods have been evaluated over 13 real-world datasets from a variety of types of text categorization. From the experimental results, the future methods achieve higher accuracies than LSTM while taking extensively less training time on most evaluated datasets, especially when the LSTM is in deep architecture. Compared to R-BLS, G-BLS has an extra overlook gate to control the flow of information (similar to LSTM) to further improve the precision on text classification so that G-BLS is more valuable while R-BLS is more efficient.

IV. MODULES DESCRIPTION

A. Module 1: Dataset Collection

For text classification we have used a large collection of data set which is more than 1,00,000. In this project, the bag-of-words is used which is a representation of text that describes the happening of words within a document.

B. Module 2: Pre-processing

All the null values and duplicate records in the dataset are handled. Categorical data is converted in to numerical form with the help of label encoder. Reprocessed data is separated as training and testing. In this model, we are handling a noise removal and processing data retrieval, text tokenization and stop words.

C. Module 3: BI-LSTM Implementation

Bi LSTM to extract the contextual information from the features outputted by the convolutional layer. In our project, Bi LSTM networks are trained using backpropagation through time. The basic idea is to split a sequence into several length-equal blocks, and apply an intra block self-attention networks (SAN) to each block independently. The outputs for all the block are then processed by an inter-block SAN

D. Module 4: Evaluation of Creating Models

Evaluation metrics used to estimate the generalization accuracy of a model on the future data. Confusion matrix is a matrix representation of the prediction results of any binary testing that is often used to describe the performance of the classification model. After the matric representation we split the data set into two sets which is testing and training set. By using bi directional LSTM we got 98 % of accuracy.

E. Module 5: Implementation of Model in django Framework

Django framework helps us to show our project in virtual environment. In our project Django provides a dynamic CRUD (create, read, update and delete) interface, which means the interface facilitates viewing, changing, and searching for information. The main reason for using django framework is we can do more than one iteration at a time without starting the whole schedule from scratch.

V. CONCLUSION

This paper principally introduces the existing model for text classification task in deep learning and NLP. NLP is a field full of opportunity and It has a tremendous consequence on how to analyse text and speeches. NLP is doing better and better every day. Knowledge mining from the large data set was impossible five years ago. The augment of the NLP technique made it possible and easy. There are still many opportunities to discover in NLP. In addition, the deep learning model enhances performance by improving the presentation learning method, model structure, additional data and knowledge. Then we introduced dataset with summary table and evaluation of creating modules for label task. Finally, we summarize the possible future research challenges of text classification.

REFERENCE

[1] Novel Efficient RNN and LSTM-Like Architectures: Recurrent and Gated Broad Learning Systems and Their Applications for Text Classification
[2] Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis
[3] A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis
[4] Learning Generic Context Embedding with Bidirectional LSTM-
[5] Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach
[6] Text Classification Using Machine Learning Techniques