News Classification

Vu Dang Nghia
Advanced Program in Computer Science
Falcuty of Ìnformation Technology
VNU-HCM University of Science HCM City, Vietnam
vdnghia@apcs.vn

Le Quoc Anh Duy
Advanced Program in Computer Science
Falcuty of Ìnformation Technology
VNU-HCM University of Science HCM City, Vietnam
lqaduy@apcs.vn

Tran Huu Thien Luong
Advanced Program in Computer Science
Falcuty of Ìnformation Technology
VNU-HCM University of Science HCM City, Vietnam
thtluong@apcs.vn

***Abstract*—News articles, especially online, are increasing rapidly. Getting them automatically classified can help save time for both the readers and the publishers. Many machine techniques were developed over the year to do this classifying task for human but not effective enough. We propose a new method using machine learning technique, which is building a neural network. The result doesn't disappoint us. This method can get the accuracy over 95 percents, which is quite surprising since we use very little resource.**

***Index Terms*—plate localization, plate detection,character segmentation, license plate detection**

I. Introduction

News information was not widely available until the year of 2000s. But now, accessing the news is very simple via newspaper or online. A huge amount of information exists in form of text in various diverse areas whose analysis can be beneficial in several areas. Classifying them is quite a challenging field in text mining. It requires steps to convert unstructured data to structured information. With the increase in the number of news, it is difficult for user to access the news of his interest. It is necessary to classify news into appropriate categories so that it can easily be accessed. Categorization refers to grouping allow easier navigation among articles. Especially, the e-news needs to be divided into different categories. This will prevent user from wasting time accessing the news that he wants to read. Talking about the news, it is such a difficult job to classify since they keep appearing and need to be processed, even some of them have never been seen before and could fall in a new category. In this paper, a review of different news classification methods based on its contents is presented.

II. Method

A. System 1

B. Data processing

C. Main algorithm

We use a 2-hidden-layer neural network, which is shallow but simple and effective. The dataset is the well-known BBC news, which include 1500 documents in total. All documents are labeled into five categories: business, politics, entertainment, sport and technology. First thing to do in Data Preprocessing is to clean the documents. We convert all the letters to lowercase, remove the punctuation and stopword, then load the dataset into a Dataframe. Next, we use a Tokenizer object from the Keras framework to create a bag-of-word dictionary. Then, we vectorize the dataset using our dictionary and split it into the train/dev/test sets. Now we are ready to build the model. Since the dataset is small and contains only text, a shallow neural network with two hidden layers is enough. It takes little time to train and iterate. We use GridSearchCV to find the best values for the model's hyperparameter. The optimizer is Adam. To reduce overfitting, we add a dropout regularization to each hidden layer. Since we're finding the best model, multiple evaluation metrics can cause confusing. Therefore to evaluate, we use the accuracy as the only tool to judge the performance of the model.

D. algorithm 2

E. Process the output

Add equation (rise the number of text and show more knowledge about the method).

III. Result

IV. Conclusion

Acknowledgment

References