FRI Natural language processing course 2021
Project report

University *of Ljubljana*
Faculty *of Computer and*
Information Science

IMapBook Collaborative Discussions Classification

Jan Ivanovič, and Grega Dvoršak

Abstract

Keywords
Natural language processing, Text classification, IMapBook

Advisors: Slavko Žitnik

Introduction

This report presents our practical work for the Natural Language Processing course at the Faculty of Computer and Information Science at University of Ljubljana. Our task was to classify chat messages from the IMapBook platform into several predefined categories such as content discussion, greeting, general comment etc. The IMapBook platform's target users are elementary school pupils, who use the platform to discuss and answer questions related to a book they were required to read. The platform can also be used to study the pupils' learning process and other aspects of their conduct while answering and discussing said questions.

To solve the text classification task, our goal was to first use more traditional approaches such as [1], [2] and [3], and then compare the results to some methods closer to state-of-the-art such as [4], [5] and [6]. The dataset contains the discussion messages which are to be classified, along with plenty additional information, such as message senders, groups and timestamps, grades for the submitted answers and more, so an important task was also to decide which of these pieces of information would be useful for classification and to use them efficiently. We report on the methods used, the results and our findings in the following sections.

Related Work

To get a better understanding of text classification and its approaches a number of other articles with similar goals were reviewed. We reviewed articles where the authors used traditional classification techniques and articles about using neural networks for text classification.

In the first article [1], the work aims to show how text from news articles can be used to predict intraday price movements of financial assets using support vector machines. They observe that while the direction of returns is not predictable using either text or returns, their size is, with text features producing significantly better performance than historical returns alone.

The authors of the second article [2] utilized named entities as features for classifying news articles into a pre-constructed hierarchy about international relations. The feature selection was implemented based on named entities associated with local categories. The documents were represented by the selected features using two types of models, Boolean model and Vector model. The experimental results show that the use of named entities improves the performance of hierarchical text classification for news articles.

The third article [3] illustrates the text classification process using machine learning techniques. It covers different topics and steps which we need to follow when doing text classification, with the goal to present different techniques and their results.

Focusing more on neural network based approaches, the next article [4] presents a deep learning architecture suitable for text classification where extreme multi-class and multi-label problems are considered. The authors use a hierarchical label set and define a methodology called Hierarchical Label Set Expansion (HLSE) to regularize the data labels and then analyze the impact of different word embedding models. The results are said to prove the usefulness of the HLSE methodology and provide insight to some combinations of word embedding models.

The next paper focuses on text classification for medical data using a neural network approach [5]. It is a special case of text classification, as it uses specific records and literature, thus causing a high dimensionality and data sparsity problem. To solve these problems, the authors propose a unified neural network method which extracts features from sentences via a convolutional layer with a bidirectional gated recurrent unit (BIGRU) and an attention mechanism to generate the sentence and document representations. The final step is a classifier of medical text categories. The method is said to be effective.

The final paper [6] uses a graph convolutional neural network for the text classification task. The authors build a single text graph for a corpus based on word co-occurrence and document-word relations which is used for learning. The network uses one-hot representations for words and documents and then jointly learns the embeddings for both. Results show, that the main advantage of this approach when compared to other state-of-the-art methods is that it performs better, when the amount of training data is reduced, suggesting robustness.

Methods

Results

Discussion

Acknowledgments

Here you can thank other persons (advisors, colleagues ...) that contributed to the successful completion of your project.

References

[1] Ronny Luss and Alexandre D'Aspremont. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012, 2015.

[2] Yaocheng Gui, Zhiqiang Gao, Renyong Li, and Xin Yang. Hierarchical text classification for news articles based-on named entities. *Advanced Data Mining and Applications*, pages 318–329, 2012.

[3] Emmanouil Ikonomakis, Sotiris Kotsiantis, and V. Tampakas. Text classification using machine learning techniques. *WSEAS transactions on computers*, 4:966–974, 08 2005.

[4] Francesco Gargiulo, Stefano Silvestri, Mario Ciampi, and Giuseppe De Pietro. Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, 79:125–138, 2019.

[5] Li Qing, Weng Linhong, and Ding Xuehai. A novel neural network-based method for medical text classification. *Future Internet*, 11(12), 2019.

[6] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7370–7377, Jul. 2019.