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

IMapBook collaborative discussions classification

Januš Likozar, Janja Koželj

Abstract

Keywords
text classification, collaborative discussion

Advisors: Slavko Žitnik

Introduction

In this assignment we try to classify messages from IMapBook collaborative discussions. Users of IMapBook were divided into groups. Each group had to read a book and had to answer given questions. Before writing final answers, each group communicated through chat messages. We try to use text processing tools to classify each message into one of the predetermined groups.

For the text classification task, we will use some baseline methods as well as some more advanced methods and compare their performance. Most of the approaches are divided into three parts: feature extraction, dimensionality reduction and classification. For feature extraction, two baseline methods would be TF-IDF and Bag of Words. These two methods do not account for word order, so there exists an improved method called continoius bag of words. To get features that better represent word meanings and so that we can use these features in machine learning algorithms we use word embeddings, like word2vec [1] or GloVe [2]. To reduce the dimensionality of feature space we can use well known algorithms like principal component analysis (PCA) or linear discriminant analysis (LDA).

For the most important step of classification we can use traditional methods like Naive Bayes Classifier (NBC), which can be extended for unbalanced classes, but has several limitations. More complex classifier methods are support vector machines (SVM) [3] and conditional random field (CRF). Using these two methods in text classification means that we will have high dimension feature space, which can lead to additional challenges. A better simple model that is fastText [4]. Currently the methods that give best results are deep learning approaches, where we use neural networks. The most widely used neural network architectures used are transformers based, mostly different improved versions of Bidirectional Encoder Representations from Transformers (BERT)[5, 6].

Methods

- SVM + bow/tfidf
- fasttext + bag of n-grams
- BERT + word embeddings

Results

Discussion

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[3] Unang Rio et al. Text message categorization of collaborative learning skills in online discussion using support vector machine. In *2013 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, pages 295–300. IEEE, 2013.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.