University of Ljubljana
Faculty of Computer and
Information Science

Text classifications using IMapBook dataset

Tara Patricija Bosil, Kristijan Šuler and Miha Arh

Abstract

TODO

Advisors: Slavko Žitnik

Introduction

In the era of digital documents, automatic text classification has been an important topic for researchers and also its use in production applications. The IMapBook concept is a web-based application, designed to improve reading comprehension among elementary and middle school students and adults. It has an integrated e-reader and games which can be played when reader completed reading certain stage. This improves reader comprehension [1]. In our case readers first read a book and are later formed into groups where they talk to each other based on specific topic which is usually question and at the end they need to provide a final answer from that topic. They talk using chat inside IMapBook application, we want to classify these messages into appropriate classes, this is later used for grading criteria. In section `Related work` we will look over different researches and methods used in NLP and data prepossessing. Then in section `Data` we examine IMapBook data gathered from students chat and what are the classes that we will be used for classification. In section `Methods` we will describe what methods will be used to classify messages and how they work. Later in section `Results` we will present results with following `Discussion` and `Conclusion`

Related work

Text classification is a very popular and widespread field, so a lot of different researches have already been done. It is mainly focus on text pre-processing, feature extraction and using different machine learning and deep learning methods. In this section we present some related works and used methods.

One of the most popular machine learning methods for text classification is Naive Bayes and support vector machines (SVM). Yu et. al. [2] compare the performance of both algorithms on two literary text classification tasks, the classification of Dickinson's poems and the sentimentalism classification. For the pre-processing they use namely stemming, stopword removal and statistical feature selection. Both algorithms achieved high accuracy for sentimental task, but for poem classification Naive Bayes performs better than SVM.

Recently, deep learning methods have become increasingly popular and they have achieved state-of-the-art results across many domains, including natural language processing. Conneau et. al. [3] propose new deep learning architecture, called very deep convolutional neural network VD-CNN. This architecture uses many layers with small convolutions and 3 pooling operations. They shown that with increasing the model depth up to 29 convolutional layers the performance drastically improves.

Zhou et. al. [4] implement a model called C-LSTM, which consist of two main components, convolutional neural network (CNN) and long short-term memory recurrent neural network (LSTM). The model uses CNN to extract higher-level sequences of word features and then this is fed to LSTM model to learn long-term dependencies. With this new model they have achieved very promising results.

Data

Data used in this report was gathered from students chatting about a specific book within IMapBook system. Before these conversations took place students had to read one of the three books given (The Lady or the Tiger, Design for the Future When the Future is Bleak or Just have less). After that small book clubs were formed where each group had to discuss about the given book and at the end provide a collaborative answer to the given question.

After the project was finished they ended up with approximately 800 chat messages and final responses. Data was then annotated and formatted into three categories. Crew data includes the whole chat history of all users where each message was annotated by the type. This type is represented in column "CodePreliminary" (we will infer to this as Term) which is exactly what our algorithm had to predict. There are also other attributes like what book did a specific group discuss about, user which posted a message, timestamp of that message and an id of the final response. Discussion only data is similar to crew data tab, but here users did not provide a final answer. The last tab consists of the final responses of each book club with some grades.

Figure 1 represents distribution of terms in raw data. We can see that the majority of messages were annotated as content discussion which means that students were mostly discussing the assigned book. There are 28 different terms that annotate all the messages, but some of this include typos so we merged some of those together to get the distribution in Figure 2.

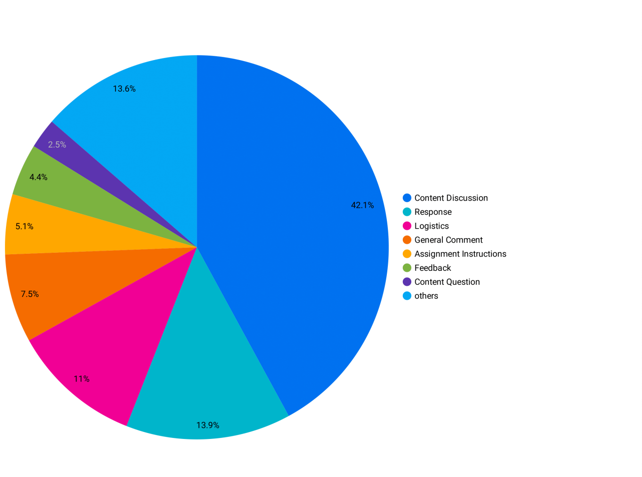

Figure 1. CodePreliminary distribution in raw data. Graph showing CodePreliminary distribution in raw data.

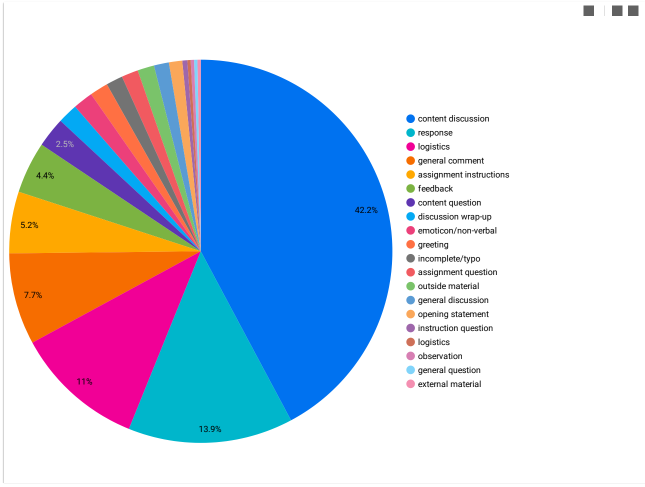

Figure 2. CodePreliminary distribution after merging. Graph showing CodePreliminary distribution after merging together attributes that include typos.

Methods

Text classification is mainly focus on three parts, text preprocessing, feature extraction and using different machine learning and deep learning methods.

Text Processing

The first and very important part of text classification is preprocessing. Text may contain many unnecessary letters and errors, so we will need to correct this misspelling. After that we will use tokenization and we will remove the stopwords. At the end we will use stemming and lemmatization.

Feature Extraction

- bag of words (BOW)
- term frequency inverse document frequency (TF-IDF)

Dimensionality Reduction

After pre-processing it is very recommended using dimensionality reduction. In that way we remove the irrelevant features and we reduce the size of the feature space. The most commonly used methods for dimensionality reduction are principal component analysis (PCA) and linear discriminant analysis (LDA).

Algorithms

- Naive Bayes
- Support vector machine
- Deep learning (deep neural networks, convolutional neural networks)

Results

Discussion

Conclusion

References

[1] Imapbook. https://www.imapbook.com/blog/. (Accessed on 03/26/2021).

[2] Bei Yu. An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327–343, 2008.

[3] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*, 2016.

[4] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.