FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and
Information Science

IMapBook Collaborative Discussions Classification

Luka Bračun, Klavdija Veselko, Demian Bucik

Abstract

In this paper we explore natural language processing approaches that aim to classify replies from discussions into predefined categories. Classes range from content discussion, greeting, logistics to feedback, response and others.

Keywords

...

Advisors: Slavko Žitnik

Introduction

Natural language processing (NLP) is a field of research where artificial intelligence, computer science and linguistics meet. Text classification is one of the NLP applications. It is defined as the process of categorizing free text according to its content. In this project we will address the text classification of collaborative discussions in online chat. For testing data we will use conversations from IMapBook [1], a web-based technology that allows reading material to be intermingled with interactive discussion. IMapBook users have access to a chat and text box where they collaborate to formulate an answer to a given question. Each message in the testing data was manually annotated with some classes, based on the information in the message. The goal of our project is to build a classifier, that would annotate messages with these classes. Such classifier could then be implemented into this platform, and could help with keeping pupils focused on discussion about the book.

Related work

Short texts compared to documents have less contextual informations, meaning they are more ambiguous, which poses a great challenge for short text classification. Examples of short texts are tweets, chat messages, reviews, search queries,...

The most used vector representations of words that capture well the semantic information are Word2Vec [2] and GloVe [3]. Both models learn vector of words from their co-occurrence information (how often they appear together in large text corpora) and are pre-trained. Word2Vec uses neural networks and it does not have explicit global information embedded in it by default. GloVe does not use neural networks and has a global co-occurrence matrix with an estimate of the probability that a certain word will appear together with other words.

There are also contextual embeddings, such as ELMo and BERT. They assign each word a representation based on its context, thereby capturing uses of words across varied contexts and encoding knowledge that transfers across languages [4]. We can also mention other methods like Bag of Words (BOW, neural embedding), WordNet (graph embedding), TF-IDF.

Data

As we can see from figure 1 the distribution of classes is skewed, which might present additional challenges during model fitting. This can be especially problematic when the distribution in the training set differs from the distribution seen in the real world. Luckily for us, training and testing sets will be constructed from the same data and will have similar distributions.

By far the most common class is content discussion, followed by response and logistics. The bottom half of the classes are practically non-existent.

Initial ideas

First we will implement a few simple approaches, like majority voting based on words or softmax regression. Then we will train and fine tune a few models that are considered state the art, and compare the results against the simple baseline models.

Some of the best performing models known today use some form of contextual or non contextual word embeddings

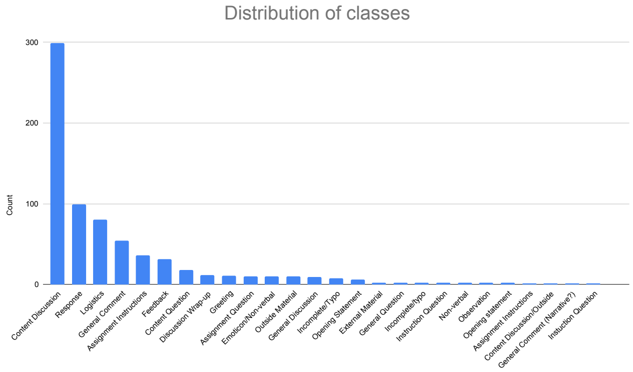

Figure 1. Distribution of classes in the provided dataset of collaborative discussions.

followed by fully connected layers for classification.

Since comments and replies are short pieces of text, we will explore the possibility of using a few of their predecessor and successors in order to improve performance. It would also be interesting to see how we can use incorporate relevant book content into our models, probably to improve embeddings, feature extraction or some other comment preprocessing in early stages of the pipeline.

We will have to pick the right metrics to evaluate model's performance. If the data distribution is skewed, we can obtain great results in metrics like accuracy by just predicting the most common class.

References

[1] Grandon Gill and Glenn Gordon Smith. Imapbook: Engaging young readers with games. *Journal of Information Technology Education: Discussion Cases*, 2(1), 2013.

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

[3] 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.

[4] Qi Liu, Matt J Kusner, and Phil Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.