

IMapbook: Automating Analysis of Group Discussions

Patrik Kojanec and Marko Rus

1 Introduction

Nature Language Processing has seen a huge rise in popularity in recent years. It is now broadly studied topic with many successful applications. In this project we touch subfield Text Classification and apply its methods to the data from IMapbook (Gill and Smith, 2013), a web-based technology that allows reading material to be intermingled with interactive games and discussions. Some portion of discussions from this platform were manually annotated, each reply was given more categories based on the information in the reply. Our goal is to take this data and try to build a classifier which would predict these categories. Such classifier could then be used to automate analysis of discussions at this platform, recommend the time for the teacher's intervention and more.

1.1 Related work

The domain of our problem is short-text classification. Unlike the common text classification problems, the short-text classification comes with a variety of restrictions that might most of the common techniques, e.g. bag of words representations, not so effective.

1.2 Proposed Work

Our goal is to construct and implement a model for predicting relevancy of replies and their category, based on the dataset from IMapbook (see Section 2). Due to relatively small size of the dataset, models based on neural networks can not be trained, and pre-trained models are not an option because of the vivid language used in conversations. Throughout the work we will only use original (Slovene) replies and ignore English translations, because they were not automatically generated.

Keeping above in mind our work can be divided into three main categories:

1. *Data Preprocessing*: Due to many errors in the language, intensive preprocessing is required. Many words are from the Slovene dictionary, and first thing to do would be to classify these words, for instance into mistakes, slang and gibberish. For other words regular text normalization will be applied.
2. *Feature Extraction*: Many features will be hand-crafted, based on the presence of the certain symbols, lengths of words, number of correctly spelled words, number of messages by that person up to observed reply, and so on.
3. *Model Fitting & Evaluation*: We will try more models, such as Naive Bayes (Lewis, 1998), Logistic Regression (Kleinbaum et al., 2002), Support-Vector Machines (Joachims, 2002), Random Forest (Liaw et al., 2002), and evaluate them by accuracy and F measure.

2 Dataset

The dataset is provided by IMapBook and includes the discussions between students and teachers on the topics of the book they are reading. The dataset includes approximately 3500 Slovene messages, from 9 different schools and on 7 different books, which were also translated to English. Students in each school were divided in "book clubs", where the conversations occurred.

The data was manually annotated, with three main tags:

- *Book Relevance*: Whether the content of the message is relevant to the topic of the book discussion.
- *Type*: Whether the message is a question (Q), answer (A), a mix of the two (AQ, QA) or a statement (S).

- *Category*: Whether the message is a simple chat message (C), related to the book discussion (D), moderating the discussion (M), wondering about users' identities (I), referring to a task, switching it or referring to a particular position in the application (S), or other cases (O).

The *Category* category can be further on split in sub-categories; *chats* may be in the form of greetings (G), related to the book (B), they could be encouraging (E), talk about feelings (F), contain cursing (C) or others (O), *Discussion* messages could be questions (Q), answers (A), answers to users, still related to the discussion topic (AA) or encouraging the discussion (E); *identity* messages can be answers(A), questions (Q) or their combination (QA).

The dataset is suitable for both binary and multi-class classification, whether the target variable is the relevance or the category of the message respectively.

References

- Grandon Gill and Glenn Gordon Smith. 2013. Imapbook: Engaging young readers with games. *Journal of Information Technology Education: Discussion Cases*, 2(1).
- Thorsten Joachims. 2002. *Learning to classify text using support vector machines*, volume 668. Springer Science & Business Media.
- David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.
- David D Lewis. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer.
- Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.