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, which is closely related to social media. Unlike the common text classification problems, where the documents are usually long and written in formal language, it deals with texts of few sentences, written in informal language. The amount of context information carried in the texts is usually very low, thus classification and information retrieval become challenging tasks to perform efficiently (Song et al., 2014). Furthermore, the low co-occurrence of words induced by the shortness of the texts often results problematic for machine learning algorithms, which rely on word frequency.

With the arise of social media this branch of text classification became a well researched problem, and people tried different approaches to overcome its constraints. In a survey in 2014, (Song et al., 2014) pointed out the main methods of short text classification, which mostly relate on semantic analysis, since it pays more attention to the concept, inner structure semantic level, and the correlation of texts to obtain the logic structure, which is more expressive and objective. Currently, the most widely used vector representations of words (or embeddings), that proved to capture well the semantic information are GloVE (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013).

Although standard machine learning approaches often resulted problematic with short text, Sriram et al. (Sriram et al., 2010) showed that their model with hand-crafted features, related to user's tweets[1], efficiently filtered irrelevant tweets from the users, thus suggesting that by adding extra sources of context information increases the performance. Similarly, this concept was also recently shown by Yang et al. (Yang et al., 2018). Furthermore, they have also shown that Support Vector Machines performed almost equally well in classification when using word embeddings or TF-IDF, but they were outperformed by deep neural networks.

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

[1] Short text message on the Twitter platform (www.twitter.com).

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 multiclass 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.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Jeffrey Pennington, Richard Socher, and Christoper Manning. 2014. Glove: Global vectors for word representation. volume 14, pages 1532–1543.

Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. 2014. Short text classification: A survey. *Journal of Multimedia*, 9.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. *Proceedings of the*

33rd international ACM SIGIR conference on Research\r\nand development in information retrieval, pages 841–842.

Xiao Yang, Craig Macdonald, and Iadh Ounis. 2018. Using word embeddings in Twitter election classification. *Information Retrieval Journal*, 21(2-3):183–207.