Helix (2021) 11 (2): 33-39

https://doi.org/10.29042/2021-11-2-33-39

Group Discussion Analysis and Digression Intervention

[*1]Sahiti Cheguru, [2]Y Vijayalata

[1, 2] Dept of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, Telanagana, India

[*1]*sahitichegurú2000@gmail.com*

[2]*vijaya@ieee.org*

Received: 8th March 2021, Accepted: 9th April 2021, Published: 30th April 2021

Abstract

It is in common knowledge that reading is one of the richest sources of knowledge in this world. Reading empowers you with the light that leads you through the dark. Therefore, we attempt to promote this valuable skill with this study. In this paper, a platform is developed that facilitates the exchange of thoughts and information among students. We have leveraged NLP to develop this application and categorize texts into various categories. Further, various textclassification methods are introducedto derive meaningful insights from written communication among students regarding books. We go on to apply the information drawn from text classification to a technology that engages readers through interactive games and discussions, IMapbook. The conversational text acquired through these discussions is further classifiedinto various categories based on the context. Here, we aim to build aclassifier that can predict these categories. Our study shows that thefine-tuned BERT, outperforms all the other methods used in this research.

Key Words: *NLP, BERT, Text classification, IMapbook*

Introduction

Nature Language Processing is a burgeoning field that has seen plenty of research breakthroughs recently. 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[1], 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. 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 Furthermore, the low co-occurrence of words induced by the shortness of the texts of ten results problematic for machine learning algorithms,which rely on word frequency.

With the rise of social media this branch of text classification became a well-researched problem, and people tried different approaches to overcome its constraints. Currently, the most widely used vector representations of words (or embeddings), that proved to capture well the semantic information are GloVE [2] and Word2Vec [3].Although standard machine learning approaches often resulted problematic with short text [4], showed that their model with hand-crafted features, related to user's tweets, efficiently filtered irrelevant tweets from the users, thussuggesting that by adding extra sources of context information increases the performance. Similarly,

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

this concept was also recently shown by Yang et al. [5]. 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.

Dataset: The dataset is provided by IMapBook andincludes the discussions between students andteachers on the topics of the book they are reading. The dataset includes approximately 3500Slovene 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) or a statement(S). In original data mixture of these classes also appear (QA and AQ), but because of their low frequency(together they appear only three times inentire dataset), we changed QA occurrences toQ and AQ to A.
- *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), refer- ring to a task, switching it or referringto a particular position in the application (S),or other cases (O).

The *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.

Research Methodology

In this segment, we explore the techniques for three different message classification tasks:

1. Book relevance classification(binary)
2. Type of message classification(3-class)
3. Broad category classification(6-class)

Input data to all classifiers are exchanged messages. To provide information about whether users are discussing about relevant topic, each message also has information about the question provided to users before the discussion.

Baseline

As a baseline model we decided to use Majority Classifier. In each task it classifies every instanceas the most representative class in training set.

Hand-Crafted Feature Models

The first group of models that we present is basedon a hand-crafted feature set. These features werethen used as an input to different classification algorithms.

Features Extraction: The aim of the features was to simply and intuitively capture the relevance to the question, while filtering gibberish and inappropriate messages. Thus, the following set of features wasde- signed:

- Tokens in a message.
- Mistakes in a message; this was computed by matching words with the words in a lexicon [6]
- Maximal length of the token in the message.
- Characters in a message.
- Question marks in a message.
- Exclamation points in a message.
- Commas in a message.
- Periods in a message.
- Capital letters in a message.

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

- Capital letters within the interior of the words in a message.
- Peculiar characters in a message.
- Numbers with in the interior of the words ina message.
- *Levenshtein distance:* Number of all pairs of words from the question and the message, whose Levenshtein distance is less than half the length of the longest of the two words.
- Interrogative words in a message.
- "kdo" in a message.

In the case of *Levenshtein distance* feature, the messages were initially tokenized and stop-words [7] were removed, while for other cases regular expressions were used to extract the features.

All features were designed while looking at thedata, having some sense in how the feature could increase the classification success. For instance, many messages had "kdo" word in it, asking for identity of somebody. Those messages have the same class. But nevertheless, we observed only small portion of the data, so that chosen features would not be over fitted.

Classification Algorithms: We decided to feed the features to four different classification algorithms to see how they perform. We chose a Naive Bayesian (NB), random forest (RF), support vector machine (SVM) and a logistic regression (LR) classifiers. We used the implementations from scikit-learn library [8].

When selecting the parameters we observedtrain and test accuracy and paid close attention to detecting over fitting. For NB we left the de- fault parameters. For the SVM we used the RBF kernel and set the parameter *gamma* to "auto" and*C* to 5, while for the LR we decided to use "lbfgs" optimizer with maximum 1000 iterations. In the case of LR the input data was standardized to ensure equal class importance. For the RF we set the number of estimators to 150, while *min_samples_leaf* to 3 and *min_samples_split* to10. This way we managed to reduce the overfitting to the training data. We kept the same parameters for all the tasks.

ELMo Embeddings

We handcrafted features by looking at the messages and observed what could potentiallydiscriminate different types of messages. For the next experiment we wanted to know, how good features can we extract automatically, so that such human interaction and understanding of messages wouldn't be necessary.

ELMo [9] is model for creating contextual embeddings. We have chosen it as it can also be used to embed entire message. Firstly, we put discussion topic into it, and then message, so thatmessage's embedding also contains information about the relevance to the topic.

We have used pre-trained ELMo model for Slovene language [10]. For classification we tried all models discussed and also KNN [11] with cosine distance, as it is natural distance to use in ELMo embeddings. Random Forest classifier ended up having the highest performance.

Fixing Typos in Messages: Messages in the input data contain a lot of words, that have typos in them and are not part of the Slovene lexicon [6]. Also, a lot of mistakes come from users deliberately leaving out carrot (e.g.'s' instead of 'š'). That is why we decided to writean algorithm for correcting typos that are away from the correct word for at most Levenshtein distance of two. We also calculated probabilities of the words and removed words with probabilityless than10^{-8}.

BERT Fine-Tuning

Another approach we propose is using a pre- trained BERT[12],by fine-tuning it for our classification tasks. We avoided customizing BERT models, because they require notorious amount of data which was not available. We trained our BERT model for Slovenian, Croatian and English languages for sequence classificationfor three epochs on training data that consisted of80% of our dataset, while the remaining 20% was left for testing. Out of these 80%, 15% were used for validation. We trained one model for each task, for both Slovenian and English- translated messages.

Results and Discussion

Baseline Models

Scores for computationally less expensivemodels (Majority Classifier and different models with handcrafted

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

features) are shown in Figure 1.We notice that all feature-based classifiers outperform the Majority Classifier. Furthermore, as expecting, the classification accuracy dropswithin creasing number of target classes. The best performing classification algorithm on this dataset is Random Forest, which outperformed the others in both "Category Broad" and "Type" classification tasks. Its performance on the "BookRelevance " task was also on average higher than the rest. However SVM and LR obtained comparable results.
Initially, RF yielded very high performance on the training set, reaching 95% accuracy. However, the performance on the test set was lower, showing signs of over fitting. Thus, with a more careful selection of the parameters, we dropped the training accuracy for about 10% and reached the current test performance.

Features Importance

RF is often used as a feature's selection tool, a sit ranks the importance of the features. The featuresoccupying the upper section of the tree are majorly decisive in predicting the output. Theinputs taken for this purpose can be used analyze the most important features or gauge theirrelative importance. Figure 2 spotlights the vitality of every feature in the decision process ofthe Random Forest model.

We evaluated the models using F1 evaluation metric. At multi-class problems we used weighting over different classes to compute it. Because of the complexity of our models, we opted for two different evaluation techniques: on models that are not so computationally expensive to fit, 5-fold cross validation was applied, where our performance estimator was the average resultof the five test sets. This technique also points out the variance of our estimator, hence quantifying to some extend the uncertainty of theperformances of our models. The second evaluation is a simple hold-out evaluation, where we split train and test sets at 80%, thus losing information about the variability of the performance of our predictor.
Lev. distance between answer and question, general length of message, and number of mistakes are shown as important features.

As we notice, each classification task focuses on different features. However there are some common ones that are discriminatory for all three tasks, i.e. last five in the plot. As expected, *Lev. Distance* works particularly well on the "Book relevance" problem since it performs a naïve kind of matching of the text messages with the questions. However, it results also as the most discriminatory feature for "Type" classification and third for "Category Broad" classification.

It is not surprising that some features are particularly relevant to some classification tasks, since they were designed for that purpose. It is also known that good features increase performance. Here we showed that some features are particularly suitable for some specific tasks, while others behave well over different classification problems. One future improvement that could be done is trying to define some other features that would boost the performance, removing the irrelevant ones.

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

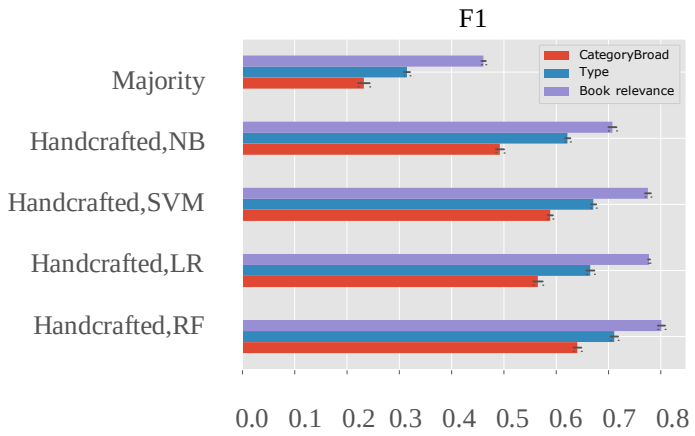

Fig 1: F1 scores of baseline models on three different classification tasks

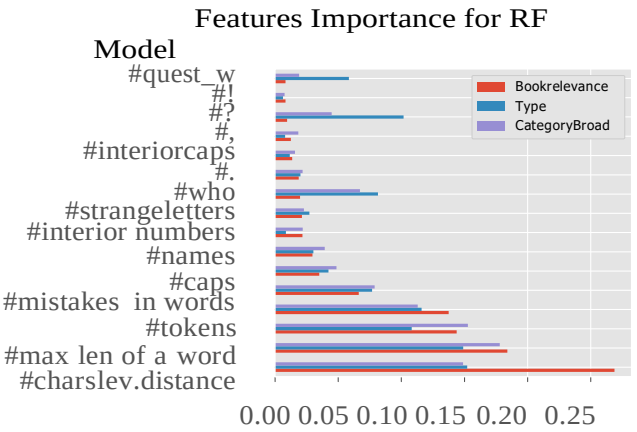

Fig 2: Features Importance

Deep Models

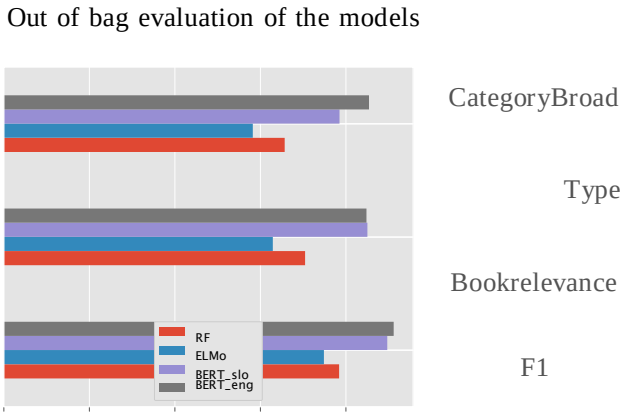

Fig 3: Hold-out performance evaluation

Comparing performances on the test set of BERT, Hand- crafted Features Model and ELMo model. F1 scores from hold-out evaluation are shown in the figure above and in this table below.

	Relevance	Type	Category
Handcrafted	0.78	0.70	0.66
ELMo	0.75	0.63	0.58
BERT	0.90	**0.85**	0.78
BERT (Eng)	**0.91**	0.85	**0.85**

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Here BERT (Eng) is BERT trained on English translations. Note that these translations were made by human and wouldn't be present in unseen data.

ELMo

We can see that ELMo has worse performance than baseline model with handcrafted features. But it is important to note here, that handcrafted features may be overfitted to the given data. If model was applied to discussions from older children, same features may perform worse. In the other hand, ELMo features are generated automatically and may generalize better.

BERT

When analyzing performance of the BERTmodels, we can clearly see15-20% increase in performance compared to model with hand crafted features. BERT model that uses English translations is even more successful, especially inthe classification of the category, where we can observe nearly 29% increase in performance. This clearly demonstrates dominance of BERT models.

We would like to mention, that here we did notmeasure uncertainty of the scores. But scores are still comparable, as we evaluated models on the same test set.

Analyzing Predictions

A lot of messages are asking for identity of somebody, and such messages were mostlysuccessfully classified by all models. Lots of messages contain a lot of gibberish and are as such distinguishable from other messages. Harder to predict are messages that are short and contain only few words. Models performed worse also on messages with a lot of unidentified mistakes in words.

Conclusion

Reading opens the gates to knowledge and wisdom like nothing else in this world. This paper progresses with this idea while drawing the benefits of Natural Language Processing. We experiment through text classification methods tounderstand the degree of accuracy to which they can automatically assign relevant categories to pieces of text. As a baseline model, we decidedto use Majority Classifier. We chose to feed features to Naïve Bayesian (NB), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) classifiers for this purpose. Further, we worked with Elmo Embeddings where Random Forest delivered the highest returns. We also fine-tuned the end-to- end BERT neural network, yielding a significant increase in performance. Future work for this research involves the recognition of messages that are direct replies to a particular message.This would improve classification with additional context that will make the categorization more meaningful.

References

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

[2] Jeffrey Pennington, Richard Socher, and Chris toper Manning. 2014. Glove: Globalvectors for word representation. volume 14, pages 1532–1543.

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

[4] 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 and\d\evelopment in information retrieval*, pages841–842.

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

[5] Xiao Yang, Craig Macdonald, and IadhOunis. 2018. Using word embeddings in Twitter elec-tion classification. *Information Retrieval Jour- nal*, 21(2-3):183–207.

[6] Kaja Dobrovoljc, Simon Krek, Peter Holozan, TomažErjavec, Miro Romih, Špela Arhar Holdt ,Jaka Cˇibej ,Luka Krsnik ,and Marko Robnik-Šikonja. 2019. Morphological lexicon sloleks 2.0. Slovenian language resource repository CLARIN.SI.

[7] JožeBucˇar.2017.Automaticallysentimentan- notated Slovenian news corpus Auto Senti News1.0. Slovenian language resource repository CLARIN.SI.

[8] F. Pedregosa, G. aroquaux, A. Gramfort, V. Michel, B.Thirion, O. Grisel, M.Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van- derplas,A.Passos,D.Cournapeau,M.Brucher,

M. Perrot, and E. Duchesnay. 2011. Scikit- learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[9] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contex- tualized word representations. *arXiv preprint arXiv:1802.05365*.

[10] MatejUlcˇar.2019.ELMoembeddingsmodelsforseven languages. Slovenian language resource repository CLARIN.SI.

[11] Keinosuke Fukunaga and Patrenahalli M. Narendra. 1975. A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7):750–753.

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

© 2021 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495