DETECTING THE CATEGORY, THE FRAMING, AND THE PERSUASION TECHNIQUES IN ONLINE NEWS IN A MULTI-LINGUAL SETUP

Rosina Baumann
University of Tübingen
rosina.baumann@student.uni-tuebingen.de
Sabrina Deisenberger
University of Tübingen
sabrina.deisenhofer@student.uni-tuebingen.de

Abstract

Fake news and false arguments are everywhere. In twitter posts or in news articles people use persuasion techniques to influence the opinion of other people. Our task is to automatically detect these non-neutral Texts (first subtask), evaluating the frame of the argumentation on article level (second subtask) and finding out which techniques were used on sentence level (third subtask). The data consists of online news and twitter posts of English, French, German, Italian, Polish and Russian. The system should generalize to sources of other languages, that it has never encountered before, as well.

1 Former Literature

In the earlier approaches we read about the authors used a BERT model and added a linear layer that they fine-tuned to the task (Martino et al., 2019). The authors had two classifiers: One was a sentence-level classification that classified if the sentence contains at least one persuasion technique and one a fragment-level classification which classified which persuasion technique was used and the corresponding place in the sentence. They evaluated different models that were different in how seperate they evaluated these subtasks.

2 Baselines

2.1 Subtask 1

The baseline classifies if a article is “satire“, “opinion“ or “reporting" by useing a SVM with a linear kernel. They also used n-grams on character level.

2.2 Subtask 2

The second subtask is a multilabel classification of 23 framings, like for example "Economic", "Crime and Punishment", "Political", or "Quality of Life". These were also classified by useing a SVM with a linear kernel. They used n-grams on word level.

2.3 Subtask 3

The third subtask is to classify the persuasion techniques like for example "NameCallin", "Simplification", "Strawman Argument" or "Guilt by Association" and mark the concerned tokens in the text. They also used a n-gram on word level.

The F1 scores for the baselines are between with 0.3 and 0.4 depending on task and language at a medium level.

3 State of our Work

We did first look at the annotation guidelines to better understand the data. Then we tried out different simple machine learning algorithms on the subtask 1 data that were similar to the given baseline. For example we tried Random Forests and different easy ensamble methods. But these methods overfitted completly and were worse than the given baseline. A simple decision tree turned out to be better than the baseline for the English but not for the Italian developmental set. The problem with these simple methods are that they overfit to one language. So if they are better in one language of the given baseline, than they are worse on other languages. This is very reasonable, because the structure of the given languages is quite different. We also tried different ranges for the n-grams, but apparently the baseline already uses a good range, which was also applicable on the other methods we tried.
Currently we are working on a method that also incorporates a small BERT model. We first tried to stay in the sci-kit learn framework and just use BERT as a preprocessing before feeding the linear models. But maybe it is better to use Pytorch directly.

4 Our Ideas

The data is unbalanced so it would be important to compare the accuracy of the different classes in subtask 1 for example with a ROC or AUC curve. It seems to be the case that also the developmental set is unbalanced, beecause if one removes the argument for unbalanced in the training the F1 accuracy increases. So for the system we will have in the end it will be important to double-check, if we make useful predictions and not only to have a high F1 score.
Our idea to not only to classify as good as possible but also to look as a second step at the features of the text that leaded to this classification. For the subtask 3 it would be very interesting to see if the system attends to reasonable features in the text. The idea is to visualize this in a attantion heatmap. For example if the following sentence is classified as "Guilt by Association (Reductio ad Hitlerum)" it would be interesting to see, if the system attends for example just to the word "Hitler", which can also be used in a different context or to the whole context like the questionmark and the exclamationmark:

"Do you know who else was doing that? Hitler !"
"Do you know who else was doing that ? Hitler!"[1]

The example shows two heatmaps that could be visualize the attentionweights of the system. Though both system could come to the same conclusion the second one should be preferred, because it also taskes the context into account.

References

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5636—-5646.

[1]The example was taken from the annotation guidelines