# Offensive Text Detection Across Languages and Datasets Using Rule-based and Hybrid Methods

Kinga Gémes[1,2], Ádám Kovács[1,2], Gábor Recski[1]

[1]TUWien, Vienna, Austria

[2]Budapest University of Technology and Economics, Budapest, Hungary

## Summary

We want to detect offensive text. But **what is** offensive?

Shared tasks don't seem to agree on one definition, especially if the datasets are in different languages.

We present a simple hybrid system that is made up of two parts. The deep learning model can be trained on a data similar to the target, but in a different language. This can be then supplemented by graph pattern rules created using human-in-the-loop learning.

The main contributions of our paper are the following:

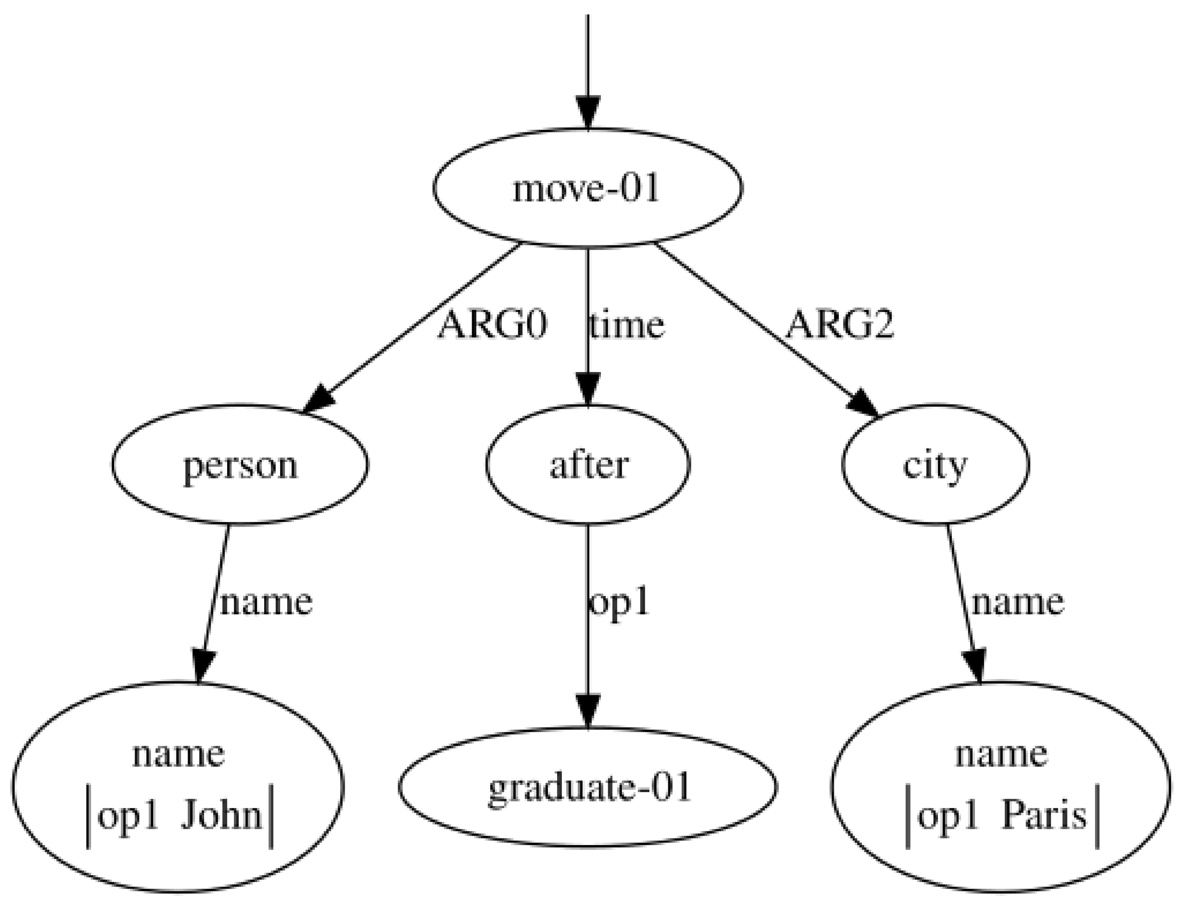
- A rule-based method for offensive text detection using semantic parsing and graph patterns
- 5 high-precision rule systems for English and German offensive text detection based on datasets from two shared tasks
- Quantitative evaluation of our rule systems, deep learning baselines, and their ensembles across 5 datasets, demonstrating that rule based and hybrid systems can outperform deep learning models in cross-dataset and cross-language settings.
- Detailed error analysis of each system on samples of 100 posts each from one English and one German dataset.

## Data

Both HASOC and GermEval (our chosen datasets) define a binary classification of social media texts (Tweets or Facebook comments) into the *offensive* and *non-offensive* classes, and a fine-grained classification of the offensive category into the subclasses *abusive*, *insulting*, and *profane*.

- GermEval [5, 6, 7]
  - Just German data
    - 2021 → "toxic" text
    - 2019 → "offensive" text
    - 2018 → "offensive" text
- HASOC [4, 2, 3]
  - English data
    - 2021 → "toxic" text
    - 2019 → "offensive" text
    - 2018 → "offensive" text
  - German data
    - 2020 → "offensive" text
    - 2019 → "offensive" text

We convert them to AMR format for our graph pattern learning framework POTATO .




## Our solution

**1.** A simple multilingual BERT based model trained on a different language dataset (English for German test, German for English test)

**2.** Human-in-the-loop learning using POTATO [1] to define semantic graph patterns that indicate offensive behavior.

Such rule might be

EN *kill* $\xrightarrow{ARG1}$ *person*

DE *(normal | eingerechnet | außer | müssen)* $\xrightarrow{polarity}$ *NEGATIVE*

We define the rules on the particular dataset's train section. Our goal is to achieve high precision with our rules.

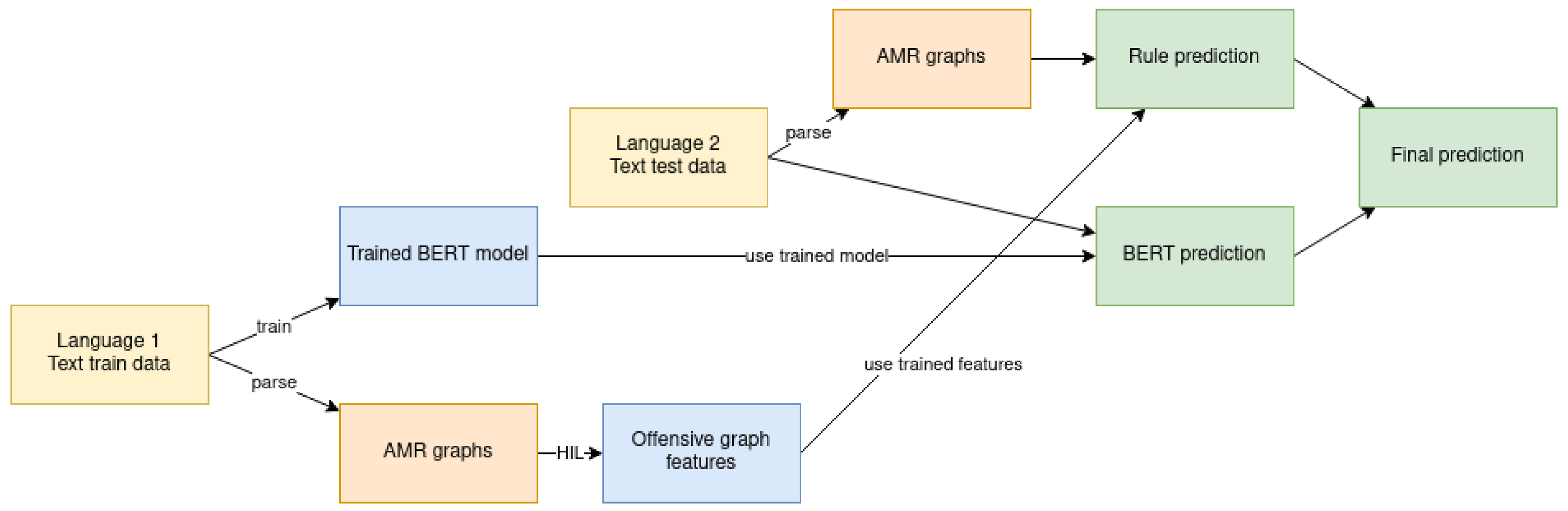



Figure 1. Our hybrid system uses both the rules and the multilingual BERT model

## Results

Our rule system almost always achieves the highest precision in the Offensive category, and does so with full interpretability.

The created rule set used together with the multilingual BERT model, that has been trained on the other language performs close to the language specific BERT model trained on the language.

| Test | System | Offensive P | Offensive R | Offensive F | Macro avg P | Macro avg R | Macro avg F |
|---|---|---|---|---|---|---|---|
| DE GermEval2021 | Rules | 65.4 | 9.7 | 16.9 | 65.0 | 53.3 | 58.6 |
| | DE BERT | **72.9** | **35.4** | **47.7** | **71.9** | **63.8** | **67.6** |
| | Multilingual EN BERT | 53.4 | 20.0 | 29.1 | 59.5 | 54.9 | 57.1 |
| | Multilingual EN BERT ∪ Rules | 54.9 | 27.4 | 36.6 | 60.9 | 57.1 | 58.9 |
| DE HASOC2020 | Rules | **92.4** | 28.3 | 43.4 | **84.7** | 63.7 | 72.7 |
| | DE BERT | 55.4 | **93.0** | **69.4** | 75.7 | **81.0** | **78.3** |
| | Multilingual EN BERT | 57.4 | 49.0 | 52.9 | 68.8 | 67.0 | 67.9 |
| | Multilingual EN BERT ∪ Rules | 62.1 | 61.7 | 61.9 | 73.2 | 73.1 | 73.1 |
| EN HASOC2021 | Rules | **87.2** | 45.1 | 59.5 | 68.4 | 67.1 | 67.7 |
| | EN BERT | 80.3 | **95.2** | **87.2** | **84.5** | **78.4** | **81.3** |
| | Multilingual DE BERT | 82.7 | 23.9 | 37.1 | 62.4 | 57.8 | 60.0 |
| | Multilingual DE BERT ∪ Rules | 84.1 | 53.9 | 65.7 | 68.2 | 68.6 | 68.4 |
| EN HASOC2020 | Rules | **95.3** | 74.6 | 83.7 | 86.9 | 85.4 | 86.2 |
| | EN BERT | 90.2 | **90.5** | **90.3** | **90.2** | **90.2** | **90.2** |
| | Multilingual DE BERT | 79.3 | 20.9 | 33.1 | 66.5 | 57.7 | 61.8 |
| | Multilingual DE BERT ∪ Rules | 89.8 | 78.7 | 83.9 | 85.2 | 84.8 | 85.0 |
| EN HASOC2019 | Rules | **73.2** | 35.1 | 47.4 | **77.4** | 65.4 | 70.9 |
| | EN BERT | 59.6 | **76.7** | **67.1** | 75.5 | **79.7** | **77.5** |
| | Multilingual DE BERT | 53.1 | 47.9 | 50.4 | 68.1 | 66.9 | 67.5 |
| | Multilingual DE BERT ∪ Rules | 55.0 | 63.5 | 58.9 | 71.1 | 73.1 | 72.1 |

## Errors

We analyzed 2 samples with size 100 and categorized the errors with human evaluation.

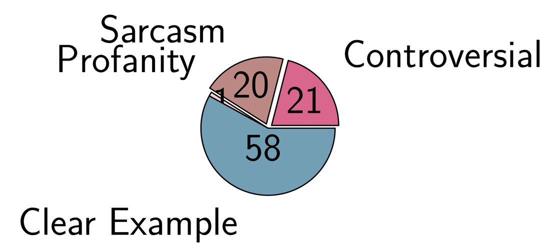



Figure 2. Frequency of error types in the German sample.

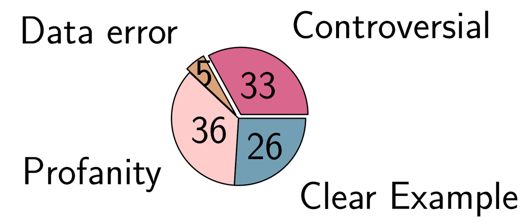



Figure 3. Frequency of error types in the English sample.

Controversial examples:

EN FN Sad reality of Indian news channels. A minute by minute coverage of elections while a common man struggles to find #covid treatment essentials. Useless News channels. #COVIDSecondWaveInIndia #CoronaPandemic #IndiaCovidCrisis #COVID19India #IndiaChoked #aajtak #zeenews #ABPnews

DE FP @USER. . . äh, Verzeihung! Fangen Sie doch einfach mal bei sich selbst, mit Ihren unnützen Motorrädern, an!

## References

[1] Á. Kovács, K. Gémes, E. Iklódi, and G. Recski. Potato: explainable information extraction framework. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, page 4897–4901. Association for Computing Machinery, 2022.

[2] T. Mandl, S. Modha, A. K. M, and B. R. Chakravarthi. Overview of the HASOC Track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA, 2020. Association for Computing Machinery.

[3] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, M. Chintak, and A. Patel. Overview of the HASOC Track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery.

[4] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, and A. K. Jaiswal. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, FIRE 2021, pages 1–3, New York, NY, USA, December 2021. Association for Computing Machinery.

[5] J. Risch, A. Stoll, L. Wilms, and M. Wiegand. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*, pages 1–12, Düsseldorf, Germany, 2021. Association for Computational Linguistics.

[6] J. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, and M. Klenner. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*, pages 352–363, München, Germany, 10 2019. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

[7] M. Wiegand, M. Siegel, and J. Ruppenhofer. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018*, pages 1–10, Vienna, Austria, 2018. Austrian Academy of Sciences.