SecHUMAN
FORSCHUNGSKOLLEG
SECURITY FOR HUMANS IN CYBERSPACE

Author profiling for hate speech spreaders

Nora Giljohann, M.A.
Faculty of Philology – German Linguistics
Ruhr-University Bochum, Germany
Supervisors: Prof. Dr. phil. Karin Pittner / Dr. phil. Kerstin Kucharczik-Kohrt

Wentao Yu, M.Sc.
Institute of Communication Acoustics
Ruhr-University Bochum, Germany
Supervisor: Prof. Dr.-Ing. Dorothea Kolossa

Motivation

Hate speech is a very complex phenomenon that has become increasingly common in our society in recent years (especially on social media). On many levels, hate speech poses a threat to our democratic values. At the same time, taking action against hate speech is not trivial. This research aims to improve the detection of hate speech by combining manual linguistic analysis with artificial intelligence, so that hate speech can be better identified and classified automatically. In addition, techniques related to author profiling of hate speech spreaders will be elaborated in this project to counteract anonymity on the Internet.

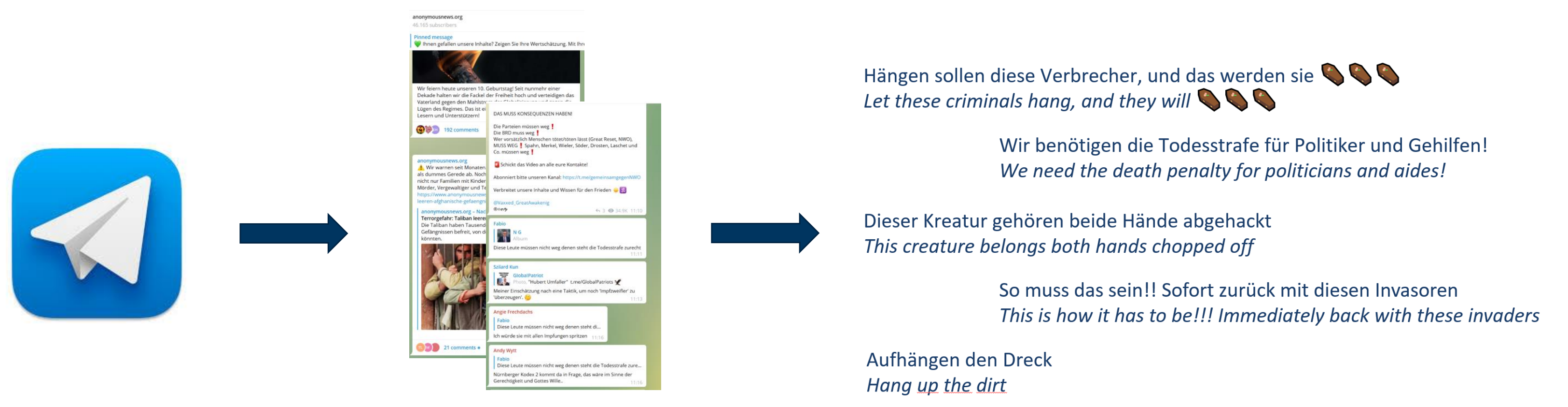
Project challenges

This research deals with the challenge of creating author profiles of hate speech spreaders by including artificial intelligence for better pre-filtering. The linguistic part focuses on the language used by the authors of hate speech and whether specific statements can be made about individuals or groups of authors. A major problem in combining hate speech and author profiling is that hate speech is characterized by short and spontaneous messages. In contrast, author profiling requires a lot of material to make valid statements about an author. Moreover, reference must be made to the existing problem of defining hate speech since the context and framework of a comment play a decisive role in determining whether it is hate speech or not[1].

From a technical perspective, developing a model with a deeper understanding of hate speech, especially sarcasm, remains challenging. Combining meaningful linguistic features with elaborately designed model structures can detect a more subtle form of hate speech.

Dataset

The self-generated dataset (in German) on which the analyses are based consists of data from the messenger service Telegram. Groups were selected according to certain criteria (e.g. topics and the possibility to create comments was important) and chat histories were downloaded.



Hängen sollen diese Verbrecher, und das werden sie ⚰⚰⚰
Let these criminals hang, and they will ⚰⚰⚰

Wir benötigen die Todesstrafe für Politiker und Gehilfen!
We need the death penalty for politicians and aides!

Dieser Kreatur gehören beide Hände abgehackt
This creature belongs both hands chopped off

So muss das sein!! Sofort zurück mit diesen Invasoren
This is how it has to be!!! Immediately back with these invaders

Aufhängen den Dreck
Hang up the dirt

Figure 1 Visualization of the collected Telegram dataset.

The dataset is manually annotated and individually evaluated according to whether it is hate speech or not. Specific annotation criteria (e.g., the definition of hate speech and explanations of the categories) have been established for this purpose, but the work on the material has once again revealed the complexity in this area. After all, some degree of subjectivity cannot be avoided in the assessment. In addition, the messages were divided into different categories, such as racism or anti-Semitism. In total, about 13.000 messages were annotated manually. This information serves as the ground truth for training the artificial intelligence (AI) models. From the data, the postings of authors who wrote the greatest number of hate speech messages were extracted in order to examine the language of the individual authors in more detail.

Automatic classification

- **Project framework**

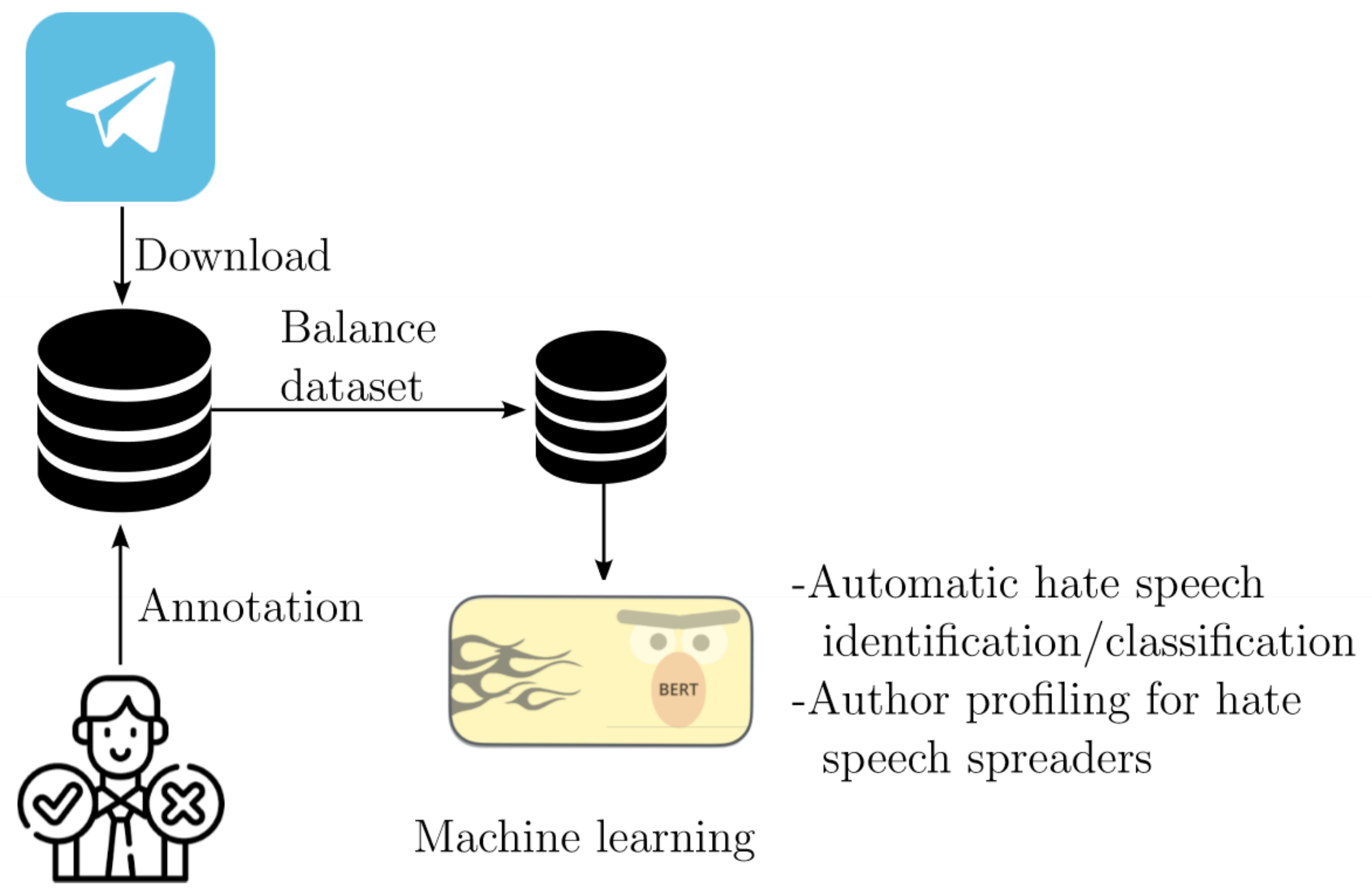


Figure 2 Project framework with BERT-based[2] model

- **Collected German Telegram dataset for training**

	Non-HS	HS						All
		Pol	***Soc***	***Rac***	***Anti-Sem***	***Med***	***Et al.***	
Samples	2500	1385	390	272	171	66	206	4990

Table 1 Hate speech distribution in collected dataset. ***HS***: hate speech; ***Pol***: politics; ***Soc***: society; ***Rac***: racism; ***Anti-Sem***: Anti-semitism; ***Med***: media; ***Et al.***: other categories.

- Accuracy of hate speech identification task: 77.71%
- Accuracy of hate speech classification task: 74.79%

- **Competitions**
  - PAN 2022: profiling irony and stereotype spreaders on Twitter with English text (1/64)
  - Memotion 3.0: sentiment and emotion analysis of English and Hinglish memes
    - Sentiment analysis (5/5)
    - Emotion classification of memes (1/5)
    - Classifying the intensity of meme emotions (1/5)

Conclusion and future work

The annotation of the comments from Telegram is very complex due to the variability of the language and the subjective assessment of hate speech. In addition, the analysis of language shows that the form of hate speech can differ greatly among different groupings and authors. Our work will further elaborate which individual, linguistic aspects of the authors are conspicuous and which linguistic structures are particularly interesting in analyzing and filtering hate speech. After that, our work will conduct several comparisons between the groups. The conclusions of this work should be able to help enhance AI performance. On the technical side, we will optimize model structures and training strategies for hate speech detection and classification. However, for the author profiling task, machine learning needs lots of data from one author and lots of author samples. We plan to construct a new dataset from our annotated Telegram data for the hate speech spreader identification task by gathering the posting from one author. The author is labeled as a hate speech spreader according to the amount of their hate speech. Our work aims to efficiently filter the hate speech spreaders (using German language) and identify the amount of hate speech of an author.

[1] Marx, K. (2020). Warum automatische Verfahren bei der Detektion von Hate Speech nur die halbe Miete sind. In: Rüdiger, T.-G. & Bayerl, P. S. (Hrsg.). Cyberkriminologie – Kriminologie für das digitale Zeitalter. Wiesbaden: Springer VS, 707-726.
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Funded by:
Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

Project partners:
Fachhochschule Dortmund
University of Applied Sciences and Arts
technische universität dortmund
RUB

Further partners:
Bundeskriminalamt