Detecting patterns of implicit offensive language in multilingual data

Ana Ostroški Anić
Institute of Croatian Language and Linguistics
Republike Austrije 16, 10 000 Zagreb, Hrvatska
aostrosk@ihjj.hr

Kristina Štrkalj Despot
Institute of Croatian Language and Linguistics
Republike Austrije 16, 10 000 Zagreb, Hrvatska
kdespot@ihjj.hr

Luka Terčon
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, 1000 Ljubljana
luka.tercon@fri.uni-lj.si

Relevant UniDive working groups: WG1, WG3

1 Introduction

Computational approaches to developing methods of automatic detection and classification of offensive language use various terms to denote socially inappropriate use of language that insults and offends others, varying from *incivility* (Stoll et al., 2020) and *toxic language* (Kunupudi et al., 2020) to *abusive language* (Caselli et al., 2020; Waseem et al., 2017; Wiegand et al., 2021b), *offensive language* (Zampieri et al., 2019) and *hate speech* (Gao et al., 2017; ElSherief et al., 2021; Schmidt and Wiegand, 2017). Although there are differences in terms used, there is a general consensus in classifying offensive language into explicit and implicit forms, and in identifying the target of the offense as an individual or group. Methods for the detection of explicit instances of offensive language have been well developed (Zampieri et al., 2019; Kumar et al., 2018; Gao et al., 2017), but detecting implicit offensive language remains a challenge (Waseem et al., 2017), partly due to a lack of rigorous linguistic analysis in existing typologies of offense. While hate speech datasets in languages other than English have also been developed (Beyhan et al., 2022; Ljubešić et al., 2021), the focus is still very much on creating English datasets, which can result in biased and limited training data.

Recently, implicit offense has attracted much attention from both linguistic and computational communities. However, one of the challenges in detecting implicit language lies in defining what constitutes implicit offense. Relying solely on applying classification tasks based on detecting explicitly offensive language, such as vulgarisms and slurs, overlooks many idiomatic expressions used to express offense. Given the wide range of conceptual and linguistic phenomena that make up implicit language, creating smaller datasets focused on specific subtypes of implicit offensive language may be a better solution ((Wiegand et al., 2021a).

2 Research aim

We propose to conduct research on developing multilingual datasets of implicit offensive language, which could be used to train language models and improve text classification and sentiment analysis for smaller and under-resourced languages. To do this, we will apply a newly proposed typology of implicitly offensive language based on an extensive linguistic analysis of a small English dataset of sentences annotated as implicitly offensive (Despot and Ostroški Anić, 2022). This typology differs between the content of offense and the linguistic devices used to express it. We categorize implicit offensive language as aggressive, insulting, and discrediting/condescending speech, as well as dehumanization, derogation, and stereotypes. Common linguistic devices used to convey offense include metaphor, metonymy, simile, irony, hyperbole, euphemisms, repetition, rhetorical questions, circumlocution, name-calling, generalizations, contrastive statements, and the use of graphic and non-verbal

devices.

3 Methodology

Our first task is to annotate comparable English, Slovene and Croatian datasets for implicit offensive language. The FRENK dataset, consisting of comments to Facebook posts of news articles of mainstream media outlets from Croatia, Great Britain, and Slovenia, on the topics of migrants and LGBT (Ljubešić, Fišer and Erjavec, 2021), will be used for this. Each dataset contains whole discussion threads, which have been annotated for the type of socially unacceptable discourse and its target. Training and testing data for each language are divided into separate discussion threads.

We will then identify common syntactic constructions used to express implicit offense, such as similes or comparisons (e.g. *looks like*, as in *Looks like reading and understanding is not your strongest point*, *Looks like you need to check your facts*), negative constructions containing positive sentiment adjectives (e.g. *not your brightest idea*, and rhetorical questions (e.g. *You think any of those women would look at you?*). Figurative comparisons are particularly significant as they convey sentiment, which is crucial in hate speech analysis. A compiled list of typical syntactic constructions can then be used to detect more examples in larger corpora.

The second task involves creating specific training datasets, such as datasets of comparisons, which can be automatically annotated with syntactic dependency annotations to identify constructions of implicit offense within them. Created datasets can be also used to investigate the role of metaphor in universal construction of offense, e.g. in expressing dehumanization as a type of offense (e.g. *You will never be anything more than a replaceable component to be put to work*). We hope to develop and describe a procedure that can be applied in detecting common syntactic patterns used in expressing implicit offense, which not only leads to further improving the detection of offensive language, but also to better understanding universal features of implicitness.

References

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish hate speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202.

Kristina Despot and Ana Ostroški Anić. 2022. Overview of approaches to implicitness. Presentation at the workshop Taxonomy and annotation of offensive language: implicitness. Nexus Workshops Days in Jerusalem, May 23-24.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Deepti Kunupudi, Shantanu Godbole, Pankaj Kumar, and Suhas Pai. 2020. Toxic language detection using robust filters.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2021. Offensive language dataset of croatian, english and slovenian comments frenk 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1433.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.