▶ Implicit Offensive Language Taxonomy and Its Application for Automatic Extraction and Ontology

Anna Bączkowska,
University of Gdansk, Poland, University of Luxembourg, Luxembourg, e-mail: anna.baczkowska@ug.edu.pl
Barbara Lewandowska-Tomaszczyk,
State University of Applied Sciences in Konin, Poland, e-mail: barbara.lewandowska.tomaszczyk@gmail.com
Slavko Žitnik,
University of Ljubljana, Slovenia, e-mail: Slavko.Zitnik@fri.uni-lj.si
Giedre Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Lithuania, e-mail: gentrygiedre@gmail.com
Chaya Liebeskind,
Jerusalem College of Technology, Israel, e-mail: liebchaya@gmail.com
Marcin Trojszczak,
University of Bialystok, University of Applied Sciences in Konin, e-mail: marcintrk@gmail.com

Purpose: In this current study, we intend to explore varying forms of implicit (mostly figurative) offensiveness (e.g., irony, metaphor, hyperbole, etc.) in order to propose a linguistic taxonomy of implicit offensiveness (and how it permeates explicit forms), and an ontology of offensive terms readily applicable to fine-tuned, pre-trained language models (word and phrase embedding). Offensive language has recently attracted great attention from computational scientists (e.g., Zampieri et al., 2019) and linguists alike (e.g., Haugh & Sinkeviciute, 2019). While in NLP scholars focus on ways of automatic extraction of what is generally and most often referred to as toxic language, in linguistics the concept of hate speech is frequently explored. Implicit offensive language, however, as opposed to explicit offence, has received little scholarly attention which so far has focused solely on single and unrelated concepts/terms. This paper aims at proposing an overarching model where varying subtypes of implicitness used in the context of offensive language are conceptually linked (Bączkowska et al., 2022).

Design/methodology/approach: The linguistic model of implicit offensive language results from a thorough literature review on implicit language seen from three perspectives: Gricean, post-Gricean, and neo-Gricean. Resulting from the analysis of existing typologies and definitions, a new model embedded mostly in a neo-Gricean approach to implicitness has been proposed. Offensiveness, on the other hand, is anchored in current approaches to offensive as well as impolite language (Culpeper, 2011, 2021; Haugh & Sinkeviciute, 2019). This taxonomy is further validated by computational methods (word and phrase embedding) aiming at finding an algorithm to cluster sim-

ilar concepts/terms and show existing dependencies among select word clusters. This study is conducted in line with the focus and approach adopted within the framework of COST ACTION CA 18209, European Network for Web-centred Linguistic Data Science (NexusLinguarum).

Findings: The validation of the linguistic model of implicit offensive language conducted by means of computational approaches to language analysis based on neural networks generally supports the linguistic taxonomy proposed in the preliminary model (Bączkowska et al., 2022). Linguistically, the research proves that the implicit model of implicitness and the explicit forms of offensiveness we proposed in our earlier model (Lewandowska-Tomaszczyk et al., 2021; Bączkowska, 2021; Lewandowska-Tomaszczyk et al., 2022; Žytnik et al., in press) are intertwined.

Research limitations/implications: The research results are dependent to a large extent on the choice and size of datasets used for the word and phrase embedding as well as the methods used for embedding (FastText, word2vec, Glove, ELMo, BERT), which are taken into account in our analysis. The implications of the study are easily transferable to offensive language annotation practice and to Linguistic Linked Data.

Practical implications: The taxonomy proposed here can be readily applied to other languages as well as to real linguistic data in order to implement automatic detection of offensive language in discourse, in particular online discourse (Twitter, Facebook, etc.). The taxonomy has already been used for linguistic data annotation with the aid of a semantic annotation tool INCEpTION (https://github.com/inception-project/inception) as part of Cost Action WG 4.1.1. Incivility in Media and Social Media.

Originality/Value: Even though the topic of offensiveness has received some attention both in the realm of linguistics and computer science, the terms ascribed to offensiveness are not well-defined and the relations among them (such as abusive, bullying, profane, obscene, insulting, etc.) rarely go beyond ad-hoc typologies and “non-systematic lexicography” (Goddard, 2018, p. 498). Our study marshals the terms that refer to various forms of offensiveness, shows relations held among them, and validates the proposed taxonomy by resorting to computational methods of language analysis. The study is thus original in the choice of methods used and the depth and breadth of concepts/terms involved in building the model of implicit offensiveness.

Keywords: *implicitness, offensiveness, embeddings, NLP.*

Research type: Research paper.

References

Bach, K. (1994). Conversational impliciture. *Mind and Language*, *9*(2), 124–162.

Bączkowska, A. (2021). “You’re too thick to change the station” – Impoliteness, insults and responses to insults on Twitter. *Topics in Linguistics*, *22*(2), 62–84.

Bączkowska, A., Lewandowska-Tomaszczyk, B., Valunaite-Oleškevičiene, G., Žitnik, B., Liebeskind, C. (2022). *Jerusalem workshop: A taxonomy of implicit language*.

Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence*. Cambridge: CUP.

Culpeper, J. (2021). Impoliteness and hate speech: Compare and contrast. *Journal of Pragmatics*, *179*, 4–11.

Goddard, K. (2018). “Joking, kidding, teasing”: Slippery categories for cross-cultural comparison but key words for understanding Anglo conversational humor. *Intercultural Pragmatics*, *15*(4), 487–514.

Haugh, M., & Sinkeviciute, V. (2019). Offence and conflict talk. In: M. Evans, L. Jeffries and J. O’Driscoll (Eds.), *The Routledge handbook of language in conflict* (pp. 196–214). Abingdon, Oxon, UK: Routledge.

Lewandowska-Tomaszczyk, B. et al. (2021). LOD-connected offensive language ontology and tag set enrichment. In S. Carvalho & R. Rocha Souza (Eds.), *LDK Workshops and Tutorials 2021* (Vol. 3064).

Lewandowska-Tomaszczyk, B., Liebeskind, C., Žitnik, B., Bączkowska, A., Valunaite-Oleškevičiene, G., & M. Trojszczak (2022). *An offensive language taxonomy and a webcorpus discourse analysis for automatic offensive language identification*. Presentation at 3rd International Conference: Approaches to Digital Discourse Analysis (ADDA 3). St Petersburg, Florida May 13–15, 2022.

Žitnik, B., Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite-Oleškevičiene, G., & Mitrovic, J. (in press). *Detecting Offensive Language: A new approach to offensive language data preparation*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (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)* (pp. 1415–1420).