

The State of Profanity Obfuscation in Natural Language Processing Scientific Publications

Debora Nozza, Dirk Hovy

Bocconi University

Milan, Italy

{debora.nozza,dirk.hovy}@unibocconi.it

Abstract

Work on hate speech has made considering rude and harmful examples in scientific publications inevitable. This situation raises various problems, such as whether or not to obscure profanities. While science must accurately disclose what it does, the unwarranted spread of hate speech can harm readers and increases its internet frequency. While maintaining publications' professional appearance, obfuscating profanities makes it challenging to evaluate the content, especially for non-native speakers. Surveying 150 ACL papers, we discovered that obfuscation is usually used for English but not other languages, and even then, quite unevenly. We discuss the problems with obfuscation and suggest a multilingual community resource called PROF with a Python module to standardize profanity obfuscation processes. We believe PROF can help scientific publication policies to make hate speech work accessible and comparable, irrespective of language.

Warning: *this paper contains unobfuscated examples some readers may find offensive.*

1 Introduction

A major downside of unsavory research subjects is that they still need to be investigated and reported, especially if we want to improve matters. Hateful language poses this challenge in natural language processing. We first need to collect and annotate it to detect, classify, and mitigate it. Setting aside the ethical conundrum of subjecting annotators to hateful language (Kennedy et al., 2018; Vidgen et al., 2019), reporting on it presents researchers with various challenges.

On the one hand, science should unflinchingly report on its subject matter, no matter how unpleasant (Jane, 2014). On the other hand, if that subject matter is language, then reporting on it is almost equivalent to producing it. This issue presents two problems: 1) proliferation and 2) audience framing.

A political h*mo? I am not listening to a fairy gay f*ggot [...]	(Zhu and Bhat, 2021)
suck a pig d*ck c*nt	(Botelho et al., 2021)
Bruh im tired of n*ggas [...]	(Shvets et al., 2021)
Someone should r*pe her	(Guest et al., 2021)

Table 1: Examples originally reported as unobfuscated in research papers. Here we obfuscate them.

Context should disambiguate whether a word is used in its intended meaning or as a *meta-function* of talking *about* the word (Jakobson, 2010) without using its original intent. However, written content that is freely accessible online might appear in unexpected contexts. E.g., as training data for language models (Brown et al., 2020).

Open access also means it is unclear who will read a given text. Scientific readers will likely assume the meta-function and discount hateful language. However, there is no guarantee as to who other readers may be. Disclaimers can help frame this problem and give the reader a choice. However, **readers who do not want to read unpleasant examples should not be excluded from conducting research in hate speech detection.** Disregarding their personal feelings about an offensive term seems cruel and insensitive at best, especially if they are members of targeted groups or abuse victims (see Table 1).

A compromise solution is **obfuscation**, where one or more letters in an offensive term are replaced with stars or other symbols. This approach preserves the word shape and allows interested readers to reconstruct the original word without allowing it to proliferate or forcing it upon readers.

However, based on our survey of 150 NLP papers, there are several issues with obfuscation:

- **Obfuscated words are often not discernible**, especially for non-native (English) speakers. Profanities are not taught in school, and we

cannot expect people learning English to know and recognize them when characters are hidden (Dewaele, 2004). This kind of language changes more readily than more formal language - even older native speakers might not recognize novel slurs. Moreover, it is basically impossible to search for obfuscated words without guessing their meaning (not to mention the impact on the search history).¹

- Authors have **many choices when obfuscating** words, e.g., obfuscate only vowels, keep or obfuscate only the first letter, etc. As a result, profanities can be obfuscated in a wide variety of ways, making their interpretation even harder. E.g., *cunt*, *c*nt*, *c**t*, *c****.
- There is **no clear definition** of what is a profanity and what should be obfuscated, especially if a word has other, more neutral, meanings. E.g., *retarded* or *r*tarded*.
- Profanities in **languages other than English** tend not to be obfuscated.

This prompts one simple question: *How can we use profanity obfuscation in scientific publication?* Among the 150 *ACL publications from 2021 reporting profanities, a number of solutions emerge. Even if standardization has been made in specific venues², these solutions are not consistent.

As outlined above, obfuscation leads to several unintended problems, predominantly for non-native speakers. As the NLP community grows, an increasing number of readers face this conundrum. *N**ger* is easy enough to guess, but *p*cker* is difficult without advanced knowledge. Or, if you are a native speaker of English, consider Danish *p*rker*, German *F*tz**, or Italian *bo***ino*. Now try googling them. Moreover, many slurs and insults are culture-specific. For example, without knowledge of the history of racism in America, it is almost impossible to even guess at the meaning of *c**n*. This issue is related to the bias towards work on English in NLP (Bender and Friedman, 2018).

Contributions We surveyed profanity reporting in 150 scientific publications. Based on our findings, we propose PROF (Profanity ObFuscation), a multi-lingual resource to help researchers converge

¹None of the authors of this paper are native speakers of English and all have faced this issue.

²<https://www.workshoponlineabuse.com/resources-and-policies/reporting-examples>

Proceedings	# obfuscated profanities
ACL 2021	67
EACL 2021	15
EMNLP 2021	11
WOAH 2021	57

Table 2: Statistics of papers using obfuscation.

on common procedures for profanity obfuscation. PROF will permit researchers to report profanities in scientific publications while ensuring formal appearance, readability, and accessibility.

2 Do we need a framework for profanity obfuscation?

Research in hate speech inevitably needs to face the use of profanities in language. These taboo words are known to be perceived negatively (Johnson, 2012; Coyne et al., 2012), leading to heightened states of emotional arousal (Jay et al., 2008), and potentially causing vicarious trauma (Vidgen et al., 2019). As scholars publishing on open access platforms, we need on the one hand to protect readers from this content and, on the other hand, to report the message in its unexpurgated entirety (Jane, 2014) because euphemisms and generic descriptors cannot convey the hostility. Obfuscation, if not ideal, is the best compromise to deal with the conundrum of hateful language in scientific literature: **obfuscated words should be discernible to allow accessibility and replicability**. How can a researcher test the same examples if it is not possible to discern the text? At the same time, while the content deciphering is left to the reader, the conveyed emotion is still negative (Stout, 2015).

3 Methodology and Results

We surveyed the ACL Anthology for proceedings of *ACL conferences that took place in 2021 and a workshop specifically focused on abuse detection, the Workshop on Online Abuse and Harms (WOAH). We searched this data for occurrences of “*” and “#”, used to obfuscate profanities.³ For each paper that included one or more profanities, we noted whether or not the authors notify the use of offensive language and which languages the authors considered in their offensive examples. Each conference’s profanity count is listed in Table 2.

³When the proceedings did not include all papers in a single file, we searched in each paper the mention of *abuse/hate/offensive/toxic* in the title or abstract.

3.1 Current practice

Several approaches can be used for obfuscation. To minimize the possibility of offending readers, it is possible to completely obfuscate the word or maintain only the first letter. E.g. *fuck* would result in “****” or “f***” or “f*”. However, this practice makes the words (almost) impossible to decipher.

The most common practice is to obfuscate vowels. For example, *fucking* would become “f*cking” or “fuck*ng” or “f*ck*ng”. This hypothetically makes the meaning intelligible by suppressing the fewest number of characters.

Summarizing the scientific publications in the *ACL community, the current practice is: (1) Obfuscation is always performed via the "*" symbol, (2) there is no shared practice of which letters to obfuscate, and (3) there are different sensibility levels when choosing which words to obfuscate, especially in languages other than English.

3.2 Considerations regarding obfuscation

Lack of a uniform profanity obfuscation in scientific articles affects readability and accessibility.

Lack of * use consistency Obfuscation is highly **inconsistent across different papers**. Some authors remove the first vowel (Xu et al., 2021a; Chuang et al., 2021; Luccioni and Viviano, 2021; Xu et al., 2021b; ElSherief et al., 2021), others obfuscate two letters (e.g., f**king) (Qian et al., 2021; Sheng et al., 2021b; Turcan et al., 2021; Bhat et al., 2021), or obfuscate the first letter (e.g., *ucking) (Kang and Hovy, 2021), or other customized choices (Ousidhoum et al., 2021; Gros et al., 2021; Sawhney et al., 2021; Mishra et al., 2021; Baheti et al., 2021). Some choices may lead to sentences that are not understandable, e.g., “All you n* and s*” (Du et al., 2021). A more important issue is the **lack of consistency within the same paper**, further compounding the confusion around profanity obfuscation practices. For example, in Sheng et al. (2021a); Mostafazadeh Davani et al. (2021), the authors obfuscate almost all letters for some words but few for others (e.g., f*** and a**hole), and Salawu et al. (2021) use both p*ssy and pu**y. If the same word is obfuscated differently, though, readers may think they are actually different words, maybe unknown (e.g., putty).

Word obfuscation choices Another problem is the choice of *whether* to obfuscate a word. Some authors choose to also obfuscate words that are

not vulgar per se, such as *dumb* or *queer* (Caselli et al., 2021; Röttger et al., 2021), but that may be offensive in a specific context, i.e., when used as an insult. Again, we found a **lack of consistency in obfuscation choices in the same paper**. This means that some authors decide to obfuscate some words (e.g., ni***r) but not others (e.g., *whore*) (Cheng et al., 2021; Vidgen et al., 2021; Bagga et al., 2021; Laugier et al., 2021; Zhou et al., 2021).

Typos We also observed typos in obfuscated words. This can generate confusion for readers who might misinterpret these mistakes as unknown profanities. For example, we found b*itch (Bertaglia et al., 2021) and wh*ore (Kirk et al., 2021).

No obfuscation A number of papers reported profanities without any form of obfuscation (Shvets et al., 2021; Fortuna et al., 2021; Hahn et al., 2021; Nozza, 2021; Zampieri et al., 2020; Sen et al., 2021; Chiril et al., 2021; An et al., 2021; Cercas Curry et al., 2021; Xie et al., 2021; Dale et al., 2021; Mehrabi et al., 2021; Leonardelli et al., 2021; Zhu and Bhat, 2021; Botelho et al., 2021; Guest et al., 2021; Hede et al., 2021). Some of these works study other languages in addition to English. Since the scientific research is English-centric, the authors potentially found the profanities in other languages less hurtful (Gonzalez-Reigosa, 1972; Harris et al., 2003; Christianson et al., 2017).

A possible solution for not using obfuscation in hate speech detection is to select examples that do not contain profanities (Niraula et al., 2021). However, we argue that scholars are responsible for reporting hate speech as severe as it is, no matter how unpleasant (Jane, 2014). Note that offensive language can occur in other non-hateful contexts as well (Malmasi and Zampieri, 2018).

Multimodality A challenging issue is profanities in images containing text, such as memes or artifact figures. While the same procedures outlined above could be applied, a solution is that (1) images created by the authors should conform to the standards, while (2) they can report images from the internet in their original form, but with a disclaimer on the paper’s first page. We observe this procedure in several publications (Zia et al., 2021; Kougia and Pavlopoulos, 2021; Qian et al., 2021; ElSherief et al., 2021; Baheti et al., 2021). There are still exceptions where artefact figures report unobfuscated profanities (An et al., 2021; Bucur et al., 2021; Zhou et al., 2021).

3.3 Considerations regarding disclaimers

Less than 20% of NLP papers use disclaimers of offensive content. However, the community needs to reach a behavioral standard. Knowing where and how disclaimers should be placed is important to ensure every reader is aware of the use of offensive examples in the paper. The papers including disclaimers applied very different practices. Disclaimers are placed (1) before the abstract (Xu et al., 2021b; Mehrabi et al., 2021), (2) after the abstract (Cercas Curry et al., 2021), (3) as a footnote on the first page (Nozza, 2021; Zampieri et al., 2020), or (4) under the table of offensive examples (Kang and Hovy, 2021; ElSherief et al., 2021). Most papers warning users of offensive language do not use any form of obfuscation in the paper. **We recommend authors add an italicized disclaimer at the end of the abstract to signal that a paper includes offensive terms.** This practice should be implemented even when profanities are obfuscated.

4 PROF

We propose PROF, a multi-lingual community resource for the obfuscation of profanities in scientific publications. It allows for the uninterrupted reading of papers with profanities while allowing non-native speakers to look up words and definitions if they desire. For the definition, we use the multi-lingual BabelNet (Navigli and Ponzetto, 2012). PROF consists of a table reporting:

- 1) the unobfuscated profanity (e.g., *fuck*)
- 2) first-vowel obfuscation (e.g., *f*ck*)
- 3) the language (e.g., *English*)
- 4) the part-of-speech (POS) tag (e.g., *NOUN*)
- 5) the BabelNet multi-lingual synset (e.g., <https://babelnet.org/synset?id=bn:00006453n&lang=EN>) or other resources if the synset does not exist.

We suggest the obfuscation practice of removing the first vowel. For compound words, we obfuscate the first vowel of the element with an offensive meaning (e.g., *femin*zis*).

We extend PROF to other languages with the help of native speakers, reaching a total of 203 profanities: 50 in English, 44 in French, 19 in German, 42 in Italian, and 48 in Spanish. Details about PROF construction are given in Appendix 4.

We understand that our work is currently limited to the profanities of the languages we speak and the set of profanities we cover. However, we currently cover all the profanities reported in ~3000

published papers. As we advance, we hope that PROF will be used as a community-based research tool that evolves in conjunction with the research conducted on it. Therefore, we aim for a crowdsourcing strategy. We recommend that researchers submit their entries to the project repository, which the authors of this paper will maintain. This method ensures the repository is comprehensive and up-to-date while also facilitating access.

PROF construction PROF construction starts with the list of English profanities surveyed from recent proceedings of *ACL conferences (see Section 3). This starting list comprises 50 entries, of which 37 unique terms and 4 unique POS tags (ADJ, ADV, NOUN, VERB). Note that profanities can be associated with different POS tags (e.g., *f*ck* can be a noun, a verb, and an adverb). Table 3 lists the most common English profanities and their associated obfuscated version.

We used these 50 English profanities as a seed for creating German, Italian, French, and Spanish PROF. Using BabelNet, we retrieve all associated concepts in other languages for each profanity. Note that each language is characterized by a different number of profanities that can be associated with a target group (e.g., women). Using BabelNet instead of a translation tool enables us to retrieve all these terms instead of just one exact translation.

The limitation of this approach is that the number of retrieved concepts starting from one term is very high and not all relevant. For example, some terms can be used to refer to profane acts in some contexts, but their main meaning is non-profane (e.g., *avvitare (screw)* is a word that can also refer to the act of having sexual intercourse). In other cases, the profanities related concepts in BabelNet are still in English or are literally translated, resulting in nonsense terms in the target language (e.g., *piece of tail*⁴ is literally translated in Italian with the non-existing idiom *pezzo di coda*). For this reason, we use a hurtful lexicon to filter retrieved related concepts (Bassignana et al., 2018).

Finally, we asked native speakers to validate the resulting filtered list of terms by removing terms that were not unambiguous profanities. We also permit native speakers to include additional profanities if they felt some popular ones were missing, on average they added 4 profanities.

⁴<https://www.urbandictionary.com/define.php?term=Piece%20of%20Tail>

obfuscated word	count
f*ck	20
n*gga	14
b*tch	13
f*cking	9
f*g	8
n*gger	8
sh*t	8
sl*t	7

Table 3: Most common obfuscated profanities in 2021 *ACL proceedings with their counts.

Using PROF We release PROF as a Python package⁵ and web application. The package automatically obfuscates profanities starting from a string or a text file, and can reveal profanities from their obfuscated versions (see Appendix B).

5 Related Work

Profanities have been investigated in NLP for discovering how to automatically filter them or how to prevent their obfuscation. These issues can be solved straightforwardly with a forbidden word list. However, preparing this list is difficult, as people are constantly creating new forms to avoid filtering via dictionary lookups, such as *\$h!t*, *sh!t*, or *s.h.i.t*. I.e., introducing spacing or punctuation between letters, swapping or removing characters, and 0–9 substitutions. Approaches to automatically filter variations of vulgar words are based on string matching techniques (Yoon et al., 2010; Ghauth and Sukhur, 2015). The research on de-obfuscation of profanities is much larger. This is due to NLP tools’ need to access the content of a sentence. Several studies (Mishra et al., 2018a,b; Eger et al., 2019; Mehdad and Tetreault, 2016) showed that obfuscated words are often ignored or treated as out-of-vocabulary impacting tasks like sentiment analysis or hate speech detection. Methods range from sequence alignment algorithms used in genomics (Rojas-Galeano, 2017) to word embeddings (Lee et al., 2018; Renwick and Barbosa, 2021). Our work differs from this literature in that we focus on scientific publications, not on social media. In this setting, the use of a dictionary is feasible.

6 Conclusion

Our work highlights the lack of obfuscation standards for reporting profanities in scientific publications. Prevailing practice allows for dangerous pro-

⁵<https://github.com/dnozza/profanity-obfuscation>

cedures and restricted access. We introduce PROF, a resource to standardize profanity obfuscation in scientific publications. PROF allows researchers to prevent offending readers while ensuring that information is readable and accessible. We plan to expand to more languages. As researchers add new words featured in their papers, PROF will grow along with the number of publications.

Limitations

We consider as profanities words that have highly offensive or vulgar connotations. We acknowledge that readers may have different sensibilities with respect to profanities. Obscene words depend on different factors, such as culture, social or religious background, and more (Hovy and Yang, 2021). Consequently, some words may be disturbing for a number of people, and should be obfuscated, while other readers may not have any issue with reading them. Moreover, we should consider that there is typically a hierarchy of offense, whereby some words are more severe than others; for example, *f*ck* is often socially accepted while the *n-word* usually is not (Sap et al., 2019).

Acknowledgements

This project has in part received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). Debora Nozza and Dirk Hovy are members of the MilanLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

- Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogan Jun, and Yong-Yeol Ahn. 2021. [Predicting anti-Asian hateful users on Twitter during COVID-19](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4655–4666, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sunyam Bagga, Andrew Piper, and Derek Ruths. 2021. [“are you kidding me?”: Detecting unpalatable questions on Reddit](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2083–2099, Online. Association for Computational Linguistics.

- Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. [Just say no: Analyzing the stance of neural dialogue generation in offensive contexts](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. [Hurtlex: A multilingual lexicon of words to hurt](#). In *Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Thales Bertaglia, Andreea Grigoriu, Michel Dumontier, and Gijs van Dijck. 2021. [Abusive language on social media through the legal looking glass](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 191–200, Online. Association for Computational Linguistics.
- Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. 2021. [Say ‘YES’ to positivity: Detecting toxic language in workplace communications](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Austin Botelho, Scott Hale, and Bertie Vidgen. 2021. [Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1896–1907, Online. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. [An exploratory analysis of the relation between offensive language and mental health](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3600–3606, Online. Association for Computational Linguistics.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. [HateBERT: Retraining BERT for abusive language detection in English](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.
- Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. [ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. [Mitigating bias in session-based cyberbullying detection: A non-compromising approach](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2158–2168, Online. Association for Computational Linguistics.
- Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. [“be nice to your wife! the restaurants are closed”: Can gender stereotype detection improve sexism classification?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kiel Christianson, Peiyun Zhou, Cassie Palmer, and Adina Raizen. 2017. [Effects of context and individual differences on the processing of taboo words](#). *Acta Psychologica*, 178:73–86.
- Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. 2021. [Mitigating biases in toxic language detection through invariant rationalization](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 114–120, Online. Association for Computational Linguistics.
- Sarah M. Coyne, Mark Callister, Laura A. Stockdale, David A. Nelson, and Brian M. Wells. 2012. [“a hel-luva read”: Profanity in adolescent literature](#). *Mass Communication and Society*, 15(3):360–383.
- David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. [Text detoxification using large pre-trained neural models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jean-Marc Dewaele. 2004. [Blistering barnacles! what language do multilinguals swear in?!](#) *Sociolinguistic Studies*, 5(1):83–105.
- Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. [Self-training improves](#)

- pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.
- Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. **Text processing like humans do: Visually attacking and shielding NLP systems**. 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 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Paula Fortuna, Vanessa Cortez, Miguel Sozinho Ramalho, and Laura Pérez-Mayos. 2021. **MIN_PT: An European Portuguese lexicon for minorities related terms**. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 76–80, Online. Association for Computational Linguistics.
- Khairil Imran Ghauth and Muhammad Shurazi Sukhur. 2015. **Text censoring system for filtering malicious content using approximate string matching and bayesian filtering**. In *Computational Intelligence in Information Systems*, pages 149–158, Cham. Springer International Publishing.
- Fernando Gonzalez-Reigosa. 1972. *The anxiety-arousing effect of taboo words in bilinguals*. The Florida State University.
- David Gros, Yu Li, and Zhou Yu. 2021. **The R-U-a-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6999–7013, Online. Association for Computational Linguistics.
- Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. **An expert annotated dataset for the detection of online misogyny**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.
- Vanessa Hahn, Dana Ruiter, Thomas Kleinbauer, and Dietrich Klakow. 2021. **Modeling profanity and hate speech in social media with semantic subspaces**. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 6–16, Online. Association for Computational Linguistics.
- Catherine L Harris, Ayşe Ayçiçeği, and Jean Berko Gleason. 2003. **Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language**. *Applied Psycholinguistics*, 24(4):561–579.
- Anushree Hede, Oshin Agarwal, Linda Lu, Diana C. Mutz, and Ani Nenkova. 2021. **From toxicity in online comments to incivility in American news: Proceed with caution**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2620–2630, Online. Association for Computational Linguistics.
- Dirk Hovy and Diyi Yang. 2021. **The importance of modeling social factors of language: Theory and practice**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.
- Roman Jakobson. 2010. *Metalinguage as a Linguistic Problem*, pages 113–121. De Gruyter Mouton.
- Emma Alice Jane. 2014. **‘back to the kitchen, cunt’: speaking the unspeakable about online misogyny**. *Continuum*, 28(4):558–570.
- Timothy Jay, Catherine Caldwell-Harris, and Krista King. 2008. **Recalling taboo and nontaboo words**. *The American journal of psychology*, 121(1):83–103.
- Danette Ifert Johnson. 2012. **Swearing by peers in the work setting: Expectancy violation valence, perceptions of message, and perceptions of speaker**. *Communication Studies*, 63(2):136–151.
- Dongyeop Kang and Eduard Hovy. 2021. **Style is NOT a single variable: Case studies for cross-stylistic language understanding**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387, Online. Association for Computational Linguistics.
- Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and et al. 2018. **The gab hate corpus: A collection of 27k posts annotated for hate speech**.
- Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. **Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset**. In *Proceedings of the 5th Workshop on Online Abuse*

- and Harms (WOAH 2021), pages 26–35, Online. Association for Computational Linguistics.
- Vasiliki Kougia and John Pavlopoulos. 2021. [Multi-modal or text? retrieval or BERT? benchmarking classifiers for the shared task on hateful memes](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 220–225, Online. Association for Computational Linguistics.
- Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. [Civil rephrases of toxic texts with self-supervised transformers](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.
- Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. [An abusive text detection system based on enhanced abusive and non-abusive word lists](#). *Decision Support Systems*, 113:22–31.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. [Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alexandra Luccioni and Joseph Viviano. 2021. [What’s in the box? an analysis of undesirable content in the Common Crawl corpus](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.
- Shervin Malmasi and Marcos Zampieri. 2018. [Challenges in discriminating profanity from hate speech](#). *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.
- Yashar Mehdad and Joel Tetreault. 2016. [Do characters abuse more than words?](#) In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles. Association for Computational Linguistics.
- Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. [Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5033, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018a. [Author profiling for abuse detection](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2018b. [Neural character-based composition models for abuse detection](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.
- Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2021. [Modeling users and online communities for abuse detection: A position on ethics and explainability](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3374–3385, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. [Improving counterfactual generation for fair hate speech detection](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. [Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network](#). *Artificial Intelligence*, 193:217–250.
- Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. [Offensive language detection in Nepali social media](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75, Online. Association for Computational Linguistics.
- Debora Nozza. 2021. [Exposing the limits of zero-shot cross-lingual hate speech detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.
- Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. [Probing toxic content in large pre-trained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.
- Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. [Lifelong learning of hate speech classification on social media](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314, Online. Association for Computational Linguistics.

- Tobias Renwick and Denilson Barbosa. 2021. [Detection and identification of obfuscated obscene language with character level transformers](#). *Proceedings of the Canadian Conference on Artificial Intelligence*. <https://caiac.pubpub.org/pub/5uqi2h7k>.
- Sergio Rojas-Galeano. 2017. [On obstructing obscenity obfuscation](#). *ACM Trans. Web*, 11(2).
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.
- Semiu Salawu, Jo Lumsden, and Yulan He. 2021. [A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 146–156, Online. Association for Computational Linguistics.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Ramit Sawhney, Puneet Mathur, Taru Jain, Akash Kumar Gautam, and Rajiv Ratn Shah. 2021. [Multitask learning for emotionally analyzing sexual abuse disclosures](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4881–4892, Online. Association for Computational Linguistics.
- Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. [How does counterfactually augmented data impact models for social computing constructs?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021a. [“nice try, kiddo”: Investigating ad hominem in dialogue responses](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021b. [Societal biases in language generation: Progress and challenges](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.
- Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. [Targets and aspects in social media hate speech](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.
- Jay Stout. 2015. *An Examination of Reader Responses to Grawlixes*. Ph.D. thesis, Doctoral dissertation, University of Hawaii at Manoa.
- Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. [Emotion-infused models for explainable psychological stress detection](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909, Online. Association for Computational Linguistics.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. [Learning from the worst: Dynamically generated datasets to improve online hate detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.
- Zhouhang Xie, Jonathan Brophy, Adam Noack, Wencong You, Kalyani Asthana, Carter Perkins, Sabrina Reis, Zayd Hammoudeh, Daniel Lowd, and Sameer Singh. 2021. [What models know about their attackers: Deriving attacker information from latent representations](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 69–78, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021a. [Detoxifying language models risks marginalizing minority voices](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021b. [Bot-adversarial dialogue for safe conversational agents](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Taijin Yoon, Sun-Young Park, and Hwan-Gue Cho. 2010. A smart filtering system for newly coined profanities by using approximate string alignment. In *2010 10th IEEE International Conference on Computer and Information Technology*, pages 643–650. IEEE.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. [SemEval-2020 task 12: Multilingual offensive language identification in social media \(OffensEval 2020\)](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. [Challenges in automated debiasing for toxic language detection](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. [Generate, prune, select: A pipeline for counterspeech generation against online hate speech](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. [Racist or sexist meme? classifying memes beyond hateful](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics.

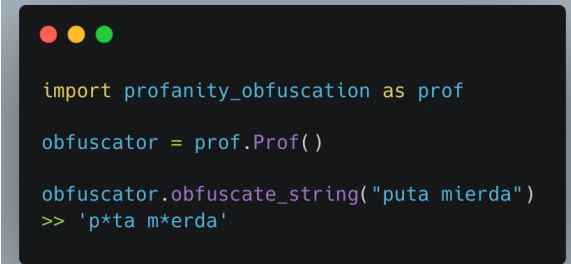
A Data Statement

We follow [Bender and Friedman \(2018\)](#) on providing a Data Statement for the proposed PROF resource.

Language-specific profanities have been validated by native speakers of each language (French, German, Italian, and Spanish). The annotators are in the age group of 25-35 and have experience in computational linguistics. Annotators were chosen from among colleagues and instructed on the research objective. The data we share is not sensitive to personal information, as it does not contain information about individuals.

B Python package

We released PROF as a Python package under the MIT license. We report some code snippets for demonstrating how the library can be used to obfuscate a profanity from a string (Figure 1) or from a text file, like a \LaTeX source (Figure 2). Figure 3 shows how our library can be used for revealing

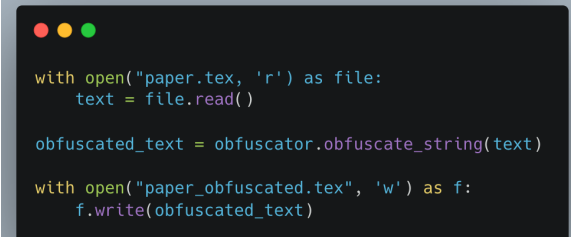


```
import profanity_obfuscation as prof

obfuscator = prof.Prof()

obfuscator.obfuscate_string("puta mierda")
>> 'p*ta m*erda'
```

Figure 1: Usage examples of the Python package for obfuscating text from a string.



```
with open("paper.tex", 'r') as file:
    text = file.read()

obfuscated_text = obfuscator.obfuscate_string(text)

with open("paper_obfuscated.tex", 'w') as f:
    f.write(obfuscated_text)
```

Figure 2: Usage examples of the Python package for obfuscating text from a file.



```
obfuscator.reveal_profanity("m*erda")
>> 'mierda'
```

Figure 3: Usage examples of the Python package for revealing obfuscated profanities.

a profanity from its obfuscated version. Finally, Figure 4 demonstrates the use of PROF as a web application for obfuscating and de-obfuscating profanities.

Profanity Obfuscation

Obfuscate

A political homo? I am not listening to a fairy gay faggot for anyone.

The obfuscated sentence is: A political h*mo? I am not listening to a fairy gay f*ggot for anyone.

Deobfuscate

A political h*mo? I am not listening to a fairy gay f*ggot for anyone.

The obfuscated sentence is: A political homo? I am not listening to a fairy gay faggot for anyone.

Figure 4: Usage examples of the web-app for obfuscating and revealing obfuscated profanities using the example in Table 1.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Limitations section
- A2. Did you discuss any potential risks of your work?
Limitations section
- A3. Do the abstract and introduction summarize the paper’s main claims?
Abstract and Introduction
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

4

- B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
GitHub webpage
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix B
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C Did you run computational experiments?

Left blank.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

No response.

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

No response.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

No response.

D **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Appendix A

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Annotators were asked to validate a limited set of profanities, and instructions were communicated via short in-person discussion.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Appendix A

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

Appendix A

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Not applicable. Left blank.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

Appendix A