Corpus-derived inoffensive vocabulary for blacklists

Keywords: impoliteness, offensiveness, blacklist, corpus linguistics, computational linguistics

Context: Filtering offensive user comments is one of the oldest NLP tasks (since Spertus 1997). It is still such an urgent research problem that this task has now become a standard one in programming competitions, e.g. CodaLab (2019). Recent publications focus on collecting and preprocessing data (e.g. Wulczyn et al. 2017), designing more accurate models (e.g. Park & Fung 2017) and dealing with problematic sentences, for instance impolite sentences without swear words (Klenner 2018).

Linguistically, this constitutes research on the pragmatic category of impoliteness (cf. Culpeper 2013, Brown & Levinson 1987), conducted within many linguistic disciplines. We approach the topic from the perspective of corpus linguistics, also applied to impoliteness research by e.g. Dewaele (2015), McEnery (2005).

Data: Our corpus consists of 73.6 k internet user comments from six public datasets: Davidson et al. (2017), Waseem & Hovy (2016), Waseem (2016), Impermium (2012), Wulczyn et al. (2017), and Cachola et al. (2018). The variety thus achieved leads to the assumption of representativeness.

14 k sentences were annotated for offensiveness by linguists, while the remainder was assessed automatically against a word blacklist. The corpus is balanced, with 40 k offensive sentences and 36 k non-offensive sentences.

Methods: For each word in the corpus, we calculated its frequency in the offensive subcorpus and the non-offensive subcorpus, normalized by the size of each subcorpus. Next, we calculated its relative offensive frequency by dividing its frequency in the offensive subcorpus by the sum of its frequencies in the offensive and non-offensive subcorpora. Finally, we performed analogous calculations for each word's rank and relative rank.

Words selected for analysis satisfied four conditions simultaneously: (1) highest frequency; (2) highest relative frequency; (3) highest rank; (4) highest relative rank.

Bigrams selected for analysis had the highest frequency in the offensive subcorpus, and were absent from the set of most frequent bigrams in the non-offensive subcorpus.

Observations: As expected, the majority of the selected expressions are swear words, slurs, etc. However, the list also contains inoffensive structures belonging to the following categories:

- lexemes for body parts (*penis*, *throat*) and bodily functions (*pee*, *swallowed*);
- lexemes for negative opinions and feelings (*hate, worthless*);
- substandard forms (*u*, *ain't*, *wanna*);
- pronominal structures (*give me*, *you can't*).

Interpretation: Lexemes for body parts and bodily functions are closely related to bodily taboos; they can be semantically shifted into dysphemisms of taboo activities (Allan & Burridge 2005). Negative opinions and feelings – expressed directly – may breach the tact maxim (Leech 1983), violating social norms and creating a face attack (Locher & Watts 2008).

Substandard forms are typical in computer-mediated communication (Al-Sa'di & Hamdan 2005, Shaw 2009) and predict impoliteness. Finally, pronominal structures constitute a building block in face-threatening acts and in linguistic insult models (Culpeper 2013).

Conclusions: Offensive user comments differ from non-offensive but comparably informal ones by referring to body parts, bodily functions, negative opinions, as well as by using substandard forms and certain pronominal structures. These observations can be used in blacklists for improved offensive sentence detectors, together with the commonly included swear words.

Bibliography

Allan, K., & K. Burridge. 2005. *Forbidden words*. Cambridge: CUP.

Al-Sa'di et al. 2005. "*Synchronous online chat* English: Computer-mediated communication". *World Englishes* 24.4. 409-424.

Brown, P. & S.C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: CUP.

Cachola, I. et al. 2018. "Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2927-2938.

CodaLab. 2019. *OffensEval: Identifying and categorizing offensive language in social media*. https://competitions.codalab.org/competitions/20011

Culpeper, J. 2013. "Impoliteness: Questions and answers". In: *Aspects of linguistic impoliteness*. 2-15.

Davidson, T. et al. 2017. "Automated hate speech detection and the problem of offensive language." In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. 512-515.

Dewaele, J.-M. 2015. "British *bollocks* versus American *jerk*: Do native British English speakers swear more – or differently – compared to American English speakers?". *Applied Linguistic Review* 6.3. 309-339.

Impermium. 2012. *Detecting insults in social commentary*. https://www.kaggle.com/c/detecting-insults-in-social-commentary/data

Klenner, M. 2018. "Offensive language without offensive words (OLWOW)". In: Ruppenhofer, J. et al. (eds.), *Proceedings of the GermEval 2018 Workshop*. 11-15.

Leech, G. 1983. "Principles of pragmatics". London: Longman.

Locher, M. A., & Watts, R. J. 2008. "Relational work and impoliteness: Negotiating norms of linguistic behavior". In : *Impoliteness in language: Studies on its interplay with power in theory*. 77-99.

Mateo, J., & F. Yus. 2013. "Towards a cross-cultural pragmatic taxonomy of insults". *Journal of Language Aggression and Conflict* 1.1. 87-114.

McEnery, T. 2005. *Swearing in English. Bad language, purity and power from 1586 to the present*. London: Routledge.

Park, J.H. & P. Fung. 2017. "One-step and two-step classification for abusive language detection on Twitter". In: *ALW1: 1st Workshop on Abusive Language Online*. Preprint.

Shaw, P. 2009. "L8r or l8a? Rhoticity variation in computer-mediated communication". In: *Corpora and discourse – and stuff*. 267-276.

Spertus, E. 1997. "Smokey: Automatic recognition of hostile messages". In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*. 1058-1065.

Waseem, Z. 2016. "Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter". In: *Proceedings of the First Workshop on NLP and Computational Social Science*. 138-142.

Waseem, Z. & D. Hovy. 2016. "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter". In: *Proceedings of the NAACL Student Research Workshop*. 88-93.

Wulczyn, E. et al. 2017. "Ex machina: Personal attacks seen at scale". In: *WWW'17. Proceedings of the 26th International Conference on World Wide Web*. 1391-1399.

Word count: 500