Smatgrisene at SemEval-2020 Task 12:

Offense detection by AI – with a pinch of real I

Peter Juel Henrichsen
Danish Language Council
pjh@dsn.dk

Marianne Rathje
Danish Language Council
mr@dsn.dk

Abstract

This paper discusses how ML based classifiers can be enhanced disproportionately by adding small amounts of qualitative linguistic knowledge. As an example we present the Danish classifier Smatgrisene, our contribution to the recent OffensEval Challenge 2020. The classifier was trained on 3000 social media posts annotated for offensiveness, supplemented by rules extracted from the reference work on Danish offensive language (Rathje 2014b). Smatgrisene did surprisingly well in the competition in spite of its extremely simple design, showing an interesting trade-off between technological muscle and linguistic intelligence. Finally, we comment on the perspectives in combining qualitative and quantitative methods for NLP.[1]

1 Introduction

Offense in the social media (SoMe) is a nuisance, not only for the targeted victim, but for editors, consumers, and stakeholders - even damaging the reputation of the internet as a whole. Could we just detect and eliminate offensive SoMe posts with the same security as we can syntax errors in python programs or even in human texts, much would be gained. The need for automatic offense detection motivated the recent OffensEval Challenge 2020 (Sigurbergsson et al. 2020, https://arxiv.org/pdf/2006.07235.pdf). In this paper we present our Danish classifier (called Smatgrisene) and report on our participation in OffensEval.

After introducing OffensEval we discuss the annotation principles behind the shared training materials and comment on a possible consistency problem. We suggest, as a remedy, a classifier design based on a mixture of simple surface rules and synthetic linguistic knowledge. We then present our OffensEval results, and conclude with a discussion of knowledge-enhanced NLP for offense detection and beyond.

2 OffensEval 2020 - Danish chapter

The Danish training material consisted of 3290 SoMe posts each annotated with labels OFF ('offensive') or NOT ('not offensive').

index	*text*	*label*
#443	jeg spørger lige dumt: hvad er det for et spil?	NOT
#507	Var det ham der sad i en lort?	OFF
#1327	Rigtig god bedring til vores Dronning ...	NOT
#2504	Holy fuck, fik lige en orgasme takket være dette billede.	OFF

Figure 1. Samples from the OffensEval training material (*"asking a stupid question: what game is it?", "was it him sitting in a shit?", "Wishing a quick recovery for our queen", "Holy fuck, I just had an orgasm from this picture")*

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2140–2145
Barcelona, Spain (Online), December 12, 2020.

Parts of the annotations were kept secret from the participants (the test corpus) while the largest part was available as training data (2961 posts). Each participating team trained an offense classifier, applied it to the test corpus, and submitted the output as their contribution. The Danish chapter of OffensEval 2020 had 39 contributing teams including Smatgrisene, the authors of this paper.

2.1 The meaning of OFF

As we gather, no instruction manual in Danish was made available to the OffensEval annotators[2]. This may be one of the reasons for the many inconsistencies in the annotations. To get an impression of the annotators' dilemmas, we translated the central term 'offend' to Danish. At least four different translations are offered in the major Danish dictionaries (Akselsen, 2000; Schwarz, 2009; Pedersen, 2007; Pedersen, 2008).

a)	(to) offend	>ED>	fornærme	>DE>	(to) offend; insult; affront
b)	(to) offend	>ED>	støde	>DE>	(to) offend; jar on; hurt
c)	(to) offend	>ED>	krænke	>DE>	(to) offend; abuse; affront
d)	(to) offend	>ED>	forarge	>DE>	(to) offend; outrage; shock

Figure 2. Danish translations of 'offend'

Fig. 2 shows the translations of 'offend' into Danish (>ED>) and back (>DE>) found in every one of the reference dictionaries. No other translation appears in all of them. Thus, in a practical sense the table provides a closed semantic field. We maintain that the Danish lexemes 'fornærme', 'støde', 'krænke', and 'forarge' are clearly semantically distinct. To substantiate this claim we had a group of native speakers of Danish (4 linguists and 4 professional writers) evaluate the semantics of the four translations of *offend*. The subjects were asked to sort several bundles of carrier sentences by likelihood ("Which utterance is more likely to appear in an ordinary conversation?"). The details are published by Dansk Sprognævn (https://dsn.dk/smatgrisene). The survey left no doubt: The four Danish lexemes are clearly semantically distinct, separated by intentionality (was the triggering action *intended* as an offense or not?) and more. In short, Danish translations of *offend* are highly ambiguous, as are related terms like *abuse*, *insult*, *violate*, and so on. It is not surprising, then, that the Danish annotation data are prone to inconsistency, the annotators having no way of resolving which translation of the English key words to prefer.

Of course, most annotation projects have to cope with fuzzy critera, and even if this does not necessarily lead to useless training data, the general versatility is at stake. Even a classifier having learned to perfection an irrelevant concoction of annotations will fail when facing the real world. In the data at hand, 'fuck' provides a telling example. The word is quite frequent in current Danish vernacular and appears in some form in 74 OffensEval posts, all labeled OFF[3]. As shown in Rathje (2014b) the actual impact of offensive words vary markedly by the recepient's age, social background, role in the interchange, relation to the utterer (2nd or 3rd person), and more. Some readers take offense at the simple appearence of 'fuck' no matter the intended meaning while others hardly notice it (#2504 in fig. 1 is an example). Similar effects are seen for words like 'lort' (#507), 'bøsse', 'perker', 'kælling', 'mullah', 'smatso', and so on (Rathje 2014b). An offense classifier not knowing such things will never surpass the IQ of a Pavlovian dog.

2.2 Regaining robustness

Faced with inconsistent training data and no context information we fight back by including some carefully crafted pieces of linguistic knowledge in the training procedure. This can be expected to enhance the classifier's performance, both in the narrow sense (more competitive in the OffensEval

[2] Instructions are non-language-specific and sometimes vague: "In our annotation, we label a post as offensive (OFF) if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct." (official OffensEval materials (Sigurbergsson et al., 2020); Zampieri et al., 2019a; Zampieri et al., 2019b)

[3] The 3 exceptions are all extreme, one (#1416) extending over 5403 characters (worthless for ML purposes?)

setting) and in a broader perspective (better prepared to face the world). To substantiate this claim we need to present a classifier either (*i*) superior to its competitors trained on the same data, or (*ii*) fully competitive even when using ML principles much simpler than its competitors while compensated by linguistic knowledge. For practical reasons we had to opt for the second option (alas), and so we adopted the simplest possible ML scheme based on unweighted regular expressions with a single-word scope (unigrams). Training was thus restricted to frequency analyses of the training data, leaving the lion's share to the linguist.

3 Danish groundwork

The data that form the basis for the classification of offensive language in this study, is collected in a sociolinguistic work on swear words and abusive terms (Rathje 2014a, 2014b).

3.1 The survey

The list of swear words used in this study is generated from a research project on language differences in three generations (Rathje 2014b). The data was harvested from conversations between informants consisting of 24 Danish women in three different generations (young, middle-aged and elderly): 8 young girls (16-18 years old), 8 middle-aged women (37-46 years old) and 8 elderly women (68-78 years old). None of the informants knew each other beforehand. The informants' task was to talk in pairs for 30 minutes in a cafe. The dialogues were recorded and subsequently transcribed. Half of the dialogues were *intra*-generational, i.e. with two participants from each their respective generation, and the other half of the dialogues were *inter*-generational, i.e. with two participants form their respective generation. In this way, it was possible to investigate how the informants communicated with someone from another generation as compared to someone from their own generation and whether the generations spoke differently: The purpose was to identify generational language.

One of the investigated language characteristics was swear words as defined in Rathje 2014b and Rathje 2017, which also draws on earlier definitions of swear words by Andersson and Trudgill, 1990; Stenström, 1991; Stroh-Wollin, 2008; Allan and Burridge, 2006; Montagu, 1967; Ljung, 2011.

> "Swear words are words that refer to something that is taboo in the culture the language is used in, they must not be taken literally, and they are used to express emotions and attitudes, but they are not used for (other) people." (Rathje 2017)

On the basis of this definition, a quantitative study was undertaken, and it was possible to determine the frequency and types of swear words in each generation. In the study it was found that the amount of swearing was the same in all three generations, but the types of swear words diverged generationally. The young participants primarily swore using English swear words (*shit*) and swear words stemming from the taboo area of 'the body's lower functions', i.e. sexuality and faeces (*fuck*, *pis*), while the middle-aged and elderly generations more frequently used religious swear words (*for fanden* and *du godeste*).

3.2 Swear words and abusive terms

Consequently, in the present study, our basis is a list of swear words used in authentic Danish speech in and between three generations of women. A later study of attitudes to Danish swearing (Rathje 2014a) reveals that it was most often the diabolic religious swear words (*for fanden, for satan*) and the English (*shit*), sexual (*fuck*) of faecal (*skide*) swear words that were perceived as coarse, i.e. offensive (Culpeper, 2011), and these types of swear words are the ones that the young generation use in Rathje 2014b. This is the rationale for using these swear words in this study:

(hvad) fanden, fuck, pisse, sgu, gud, (ikke en) skid, fandme, fucked up, (ad) helvedes (til), hulens, skide, sur røv, shit, holy shit, eddermaneme, (hvordan) fanden, røv-, søreme.

In many ways, abusive terms are similar to swear words because they too express a feeling or an attitude, and they are also associated with taboos (Rathje 2014a). For example, *luder* (whore) has a connection with the taboo of 'prostitution', and other taboos can be 'homosexuality' (e.g. *bøsse* (gay)) and 'mental illness' (e.g. *psykopat* (psychopath)). Like swear words, abusive terms should not be understood literally: i.e. *luder* (whore) as an abusive term refers not to an actual prostitute, but it expresses an opinion about a particular person who is not a prostitute in the literal meaning. And precisely the fact that there is a person at whom the word has been directed separates swear words and abusive terms from one another: Abusive terms are used about people, while swear words are not. It is, therefore, abusive terms in particular that are experienced as being offensive.

The abusive terms used in the present study stem from a survey in which two generations expressed their conscious attitudes to Danish swear words (Rathje 2014a). This data consisted of 844 completed questionnaires about young people's and elderly people's attitudes toward swearing. Of those, 63% were answered by young people 13–14 years old, and 37% were answered by elderly people 65–93 years old. Among the younger participants, half were boys and the other half were girls, while the elderly respondents consisted of 18% men and 82% women.

For the question in the survey "What are the coarsest swear words you know?", terms of abuse constituted as much as 68% of the words that young people mentioned and 17% of the words from the elderly. Many of the informants clearly did not distinguish between swear words and terms of abuse, but these answers have provided the list below: the most often-mentioned terms of abuse by young and old informants.

3.3 Exporting data to NLP

These are the ones we have used in our offense detection.

The young people's top 10 terms of abuse are (Rathje 2014a):

1. luder (*whore*)
2. kælling (*bitch*)
3. bøsse (*queer*)
4. so (*slut*)
5. bitch
6. fuck dig (*fuck you*)
7. idiot
8. perker (*paki*)
9. smatso (*cunt*)
10. pikslikker; svin (*cock sucker*; *pig*)

The top 5 terms of abuse among the elderly are (Rathje 2014a):

1. luder (*whore*)
2. idiot
3. svin (*pig*)
4. fuck you
5. mula (*mulatto*); fanden tage dig (*may the devil take you*); motherfucker; laban (*lout*)

4 Smatgrisene, the classifier

As the technical details are without real interest (and beside the main purpose of the paper), we keep this section at a minimum. Our training algorithm hardly qualifies as such: frequency-based extraction of regular expression templates from the OFF and NOT lists, thus immediately reproducible from the code (cf. https://dsn.dk/smatgrisene).

4.1 Script

These program lines (perl) are at the core of the classifier.

```
$regexp = join '|',(@trained_triggers,@Dan_epi,@Dan_swear);
 (iterating over OffensEval posts:)
print "$outputline\t".($inputline=~/$regexp/? 'OFF': 'NOT')."\n";
```

`$regexp` holds a regular expression with the disjunction of (*i*) all triggers extracted from the training data `@trained_triggers,` (*ii*) a set of Danish epithets `@Dan_epi,` and (*iii*) a set of Danish swear words `@Dan_swear`. The two latter sets provide the 'pinch of real I'. All tokens were lowercased, but otherwise unchanged. As shown, all three sets were used indiscriminately not utilizing the rich metadata available in Rathje (2014b). Even this blunt use of linguistic quality data boosted the performance significantly.

4.2 Evaluation

Is Smatgrisene competitive? This depends on how 'competitiveness' is determined. We found these criteria to be natural: Firstly, the classifier must meet or surpass the state-of-the-art prior to OffensEval 2020 (Sigurbergsson et al. 2020). Secondly, it must compare favourably with its competitors, being

- among the best-scoring half
- above the average score for all participants
- above the median score for all participants

According to the official scoreboard (https://arxiv.org/pdf/2006.07235.pdf). Smatgrisene reached F1=0.759 earning a shared 14th place, meeting our success criteria comfortably. While Smatgrisene may not be the champion of OffensEval Challenge 2020, it *is* demonstrably competitive. Moreover, as the table shows, Smatgrisene clearly owes its competitiveness to the combination of AI and real I.

'OFF' triggering configuration	**`@Tr,@D1,@D2`**	`@Tr,@D1`	`@Tr,@D2`	`@D1,@D2`	`@Tr`
F1	**0.759**	0.743	0.719	0.736	0.700
*n*th place of 39 participants	**#14 (shared)**	#19	#23	#20	#25
≥ State-of-the-art (F1=0.73)	***yes***	*barely*	*no*	*barely*	*no*
Above average ($\mathrm{F1}_{Mean}$=0.707)	***yes***	*yes*	*barely*	*yes*	*no*
Above median ($\mathrm{F1}_{Med}$=0.727)	***yes***	*yes*	*no*	*barely*	*no*

Table 1: Smatgrisene performance. Figures from the OffensEval scoreboard are in **bold**; col. 3-6 are versions deprived of one or more OFF triggers (`@Tr`=trained, `@D1`=epithets, `@D2`=swear words, cf. 4.1).

Consider the various classifier modes in table 1 (col. 2-6). In mode `@Tr` ('no linguistics') the classifier performs much better on the training data than on the test data: F_{train}=0.846, F_{test}=0.700. This is hardly surprising, the triggers being derived from the former. In mode `@D1,$D2` ('linguistics only'), in contrast, the figures are far more equal: F_{train}=0.731, F_{test}=0.736, a clear sign of the robustness towards unseen data types that Smatgrisene acquired from the expert knowledge.

5 Conclusion

We have presented the classifier Smatgrisene (the name being a rarely, if ever, used lexical variant of 'smatso', one of the most abusive epithets in Danish), trained to detect offensive language in SoMe posts. The training algorithm combined simple surface rules (derived from the OffensEval training data) and deep linguistic knowledge (distilled from a comprehensive academic study of Danish offensive language, Rathje (2014b)). Despite its extreme formal simplicity, Smatgrisene did unexpectedly well in the OffensEval Challenge 2020 and came out in the top third of participants.

A real achievement, however, would be to utilize the rich metadata provided in Rathje (2014b) in a future classifier for offense detection. Rathje (2014b), like others in the same tradition, analyses the perception of potentially offensive language as a function of the interlocutors' age, social standing and societal background, and of the purpose and intentions of the interchange. We surmise that such synthetic knowledge could sharpen the focus of classifiers monitoring the communication channels in chat rooms and social media. Offensive communication is a complex social game. At the end of the day there is no such thing as an intrinsically offensive word.

"More data will solve any problem", "acribia is for sissies", "fire your linguists!". Such fresh attitudes are currently shared by many manufacturers of language technology. We invite the serious developer to rediscover the power of linguistic insight.

References

Keith Allan and Kate Burridge. 2006. *Forbidden words.* Cambridge University Press, Cambridge.

Lars-Gunnar Andersson and Peter Trudgill. 1990. *Bad Language.* Basil Blackwell, Oxford, United Kingdom.

Jens Axelsen (ed). 2000. *Dansk-Engelsk Ordbog.* Gyldendal, København, Denmark.

Jonathan Culpeper. 2011. *Impoliteness: Using Language to Cause Offence.* Cambridge University Press, Cambridge, United Kingdom.

Magnus Ljung. 2011. *Swearing. A Cross-Cultural Linguistic Study*. Palgrave Macmillan, New York, United States of America.

Ashley Montagu. 1967. *The Anatomy of Swearing*. The Macmillan Company, New York, United States of America.

Viggo Hjørnager Pedersen (ed). 2007. *Engelsk-Dansk Ordbog Kjærulff Nielsen.* Gyldendal, København, Denmark.

Viggo Hjørnager Pedersen (ed). 2008. *Dansk-Engelsk Ordbog Vinterberg & Bodelsen.* Gyldendal, København, Denmark.

Marianne Rathje. 2014a. Attitudes to Danish swear words and abusive terms in two generations. In Marianne Rathje (ed.): *Swearing in the Nordic Countries*. Dansk Sprognævns konferenceserie 2, Dansk Sprognævn, København, Denmark: 37-61.

Marianne Rathje. 2014b. Swearing in the speech of young girls, middle-aged women and elderly ladies. In Helga Kotthoff and Christine Mertzlufft (eds). *Jugendsprachen. Stilisierungen, Identitäten, mediale Ressourcen.* Peter Lang, Frankfurt am Main, Germany 347-372.

Marianne Rathje. 2017. Swearing in Danish Children's Television Series. In Kristy Beers Fägersten and Karyn Stapleton (eds). *Advances in Swearing Research: New languages and new contexts.* John Benjamins Publishing Company. Pragmatics and Beyond New Series, volume 282: 17-42 https://doi.org/10.1075/pbns.282.02rat

H. L. Schwarz et al. (eds). 2009. *Engelsk-Dansk Ordbog,* Munksgaard, København, Denmark.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. *Offensive Language and Hate Speech Detection for Danish*; proceed. 12th Language Resources and Evaluation Conference.

Anna-Brita Stenström. 1991. Expletives in the London-Lund Corpus. In: K. Aijmer & B. Altenberg (eds): *English Corpus Linguistics*, Longman, New York, United States of America: 239-253.

Ulla Stroh-Wollin. 2008. Dramernas svordomar – en lexikal och grammatisk studie i 300 års svensk dramatik. Svensk dramadialog 10, FUMS Rapport nr 224. Institutionen för nordiska språk vid Uppsala Universitet, Uppsala, Sweden.

M. Zampieri, Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. 2019a. *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, Association for Computational Linguistics, Minneapolis, Minnesota, June, United States of America.

M. Zampieri, Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. 2019b. *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*. In Proceedings of SemEval.