Are People on the Internet More Rude at Night? — Computational Literacy course project

Heljä Räisänen

December 21, 2023

Research questions

The question that is attempted to answer here are is, do people on Internet forums use more crude, vulgar, and insulting language at night? ”Crude”, ”vulgar”, and ”insulting” here mean any swearing, overt sexual references, and insults (whether widely considered swearing or not).

Swearing can be defined as ”words that ‘have the potential to be offensive, inappropriate, objectionable, or unacceptable in any given social context’”, as stated by Stapleton et al. in their 2022 publication ’What is Swearing?’ (Stapleton, Beers Fägersten, Stephens, & Loveday, 2022). This is a broad category, and for the purposes of this paper, insults are also considered swearing, as these are viewed as rude and offensive. Vulgar terms referring to body parts and sexual acts are also included, as these are words that are widely deemed inappropriate and offensive to say, and the tabooness and rudeness of the words used on the forum are what is specifically of interest here.

The assumption is, that people are more rude to one another on the Internet during night, when they are not actively partaking in social institutions, such as work or family. Anonymous and pseudonymous people are probably more rude than people using their own name, so a forum such as Suomi24 probably contains abundant rudeness, but does the context of being at work or with family during the day hours affect the rudeness of posts made at that time? Or do people who tend to be rude post during the night, with little activity during the day? These questions are left for further research, but the initial question will be answered.

Data

The data used in this project comes from the Finnish forum Suomi24. Suomi24 might be one of the biggest discussion forum of Finland, but whether the behaviour of its userbase reflects Internet forums at large cannot be answered here, and the results of the project hold verifiably true only for Suomi24. Suomi24 is technically a group of forums with different topics, and the differences between these ”sub-forums” are considered.

This project uses a Suomi24 corpus from the Korp web interface (Aller Media ltd., 2014). The corpora includes a fraction of the posts from the forum from between 2001-2014, and belongs to Kielipankki (Language Bank of Finland), which is coordinated by FIN-CLARIN (Kielipankki, 2021). The corpus is freely available under the CC-BY licence.

Body & sexuality	references to body parts, sexual acts
Censored	the word is fully or partially censored by omitting letters, replacing letters with symbols
Common	the most common swear words (subjective & biased)
Insults	insults, whether including swearing or not
Insults to men	insults aimed traditionally at men, such as emasculating insults
Insults to women	insults aimed traditionally at women, such as attacks on sexual behaviour
Religion (Christianity)	swears that have a religious double meaning, such as “devil”
Slurs	slurs aimed at minorities, such as sexual minorities or people with disabilities (excluding ethnic minorities)
Weak	words that you can say in the workplace or around children (this is highly subjective and biased)

Table 1: Categorization of swear words.

Swear words

The word list of crude language that is used was decided as follows. The Turku word2vec language model (Turku NLP group, n.d.) was used to harvest words semantically similar to a hand-picked list of common swears at ’data/common_swears.txt’. The 100 most similar words of each in the list were collected to a list of words to query from the API. This list was cleaned using OpenRefine (Delpeuch et al., 2023), removing duplicate words and capitalizations. Then, the words were manually cleaned of irrelevant words. This isn’t entirely reproducible, as what counts as a swear is somewhat subjective. The words were then categorized into 1-3 of a total of 9 categories, which is also subjective, and not reproducible. The categories are listed in the table 1 and the total list of swears used can be found at ’data/swears-csv.csv’.

This categorization is arbitrary, but the categories ’Religion’ and ’Body & sexuality’ can be found in other literature also (Stapleton et al., 2022), as these subjects are ubiquitously taboo in western societies and thus good material for swearing. Opposed to Stapleton et al, here is included the category ’Common’, which has words of various different roots and semantic meanings. This is useful, as this serves as an attempt to preserve the context in which these words were used, and their meaning is of lesser concern. The ’Common’ category holds swears and insults that a Finnish person might hear the most during their life, though this is a biased supposition, and words that hold no specific object to be targeted at or a context other than simply swearing. Conversely, the ’Weak’ category holds swears that are aimed towards a recipient or audience that the swearer doesn’t want to insult, such as your grandma, and the ’Insults to men’ and ’women’, and also ’Slurs’, are aimed at a recipient belonging, or insinuated to belong, to a specific group. The words in the ’Body & sexuality’ category can be used in honest yet crude discussion of the named matters, not only as plain swears. The ’Censored’ category is interesting because the swearer is, in my interpretation, showing an effort to not insult others with their crudeness, and words in the ’Religion’ category can also denote a mere discussion of religious concepts, instead of vulgarity. This is my motivation for the categorization. Note that words in the ’Common’ category often belong to other categories, too, and most words in all categories have an additional category.

Methods

The data processing pipeline was as following:

The special characters of the censored words were manually "escaped" because this was surprisingly difficult to do with Python and I gave up.

The Korp API was queried to get all instances of crude language posted on the forums. The corpus was accessed through its public API using the 'korp' Python package (Hämäläinen, 2019). The code used for obtaining the instances of swearing ('concordances' in Korp's terms) can be found in 'code/get_concordances.py'. This Python script queries the API and saves the results. The script fetches extraneous information because it was interesting to look at. The fetched concordances are pruned of extra information and saved in a .json file.

In addition to the instances of swearing, the total amount of words posted needs to be considered. The statistics on all posts made on all the forums is queried in a Web browser, because the korp library doesn't support this and I didn't want to figure out how to write this functionality myself. The query sent to the API is the first url in 'data/example_query.txt'.

The collected JSON data is then processed and visualized with another script, 'code/process_frequencies.py'. Here the MatPlotlib Python package (Hunter, 2007) was used to draw bar plots. In this script, the frequencies of crude language per each hour of the day across all years spanned by the corpus are proportioned to the amounts of total 'tokens', or words, posted on the forums each hour of the day. This way the fraction of crudeness across all posts is obtained separately for each hour. This is done across all forums in one plot, and also showing the subforums individually in another plot. At this point, some subforums are excluded from the process for the latter plot, because they contain two or less instances of swearing.

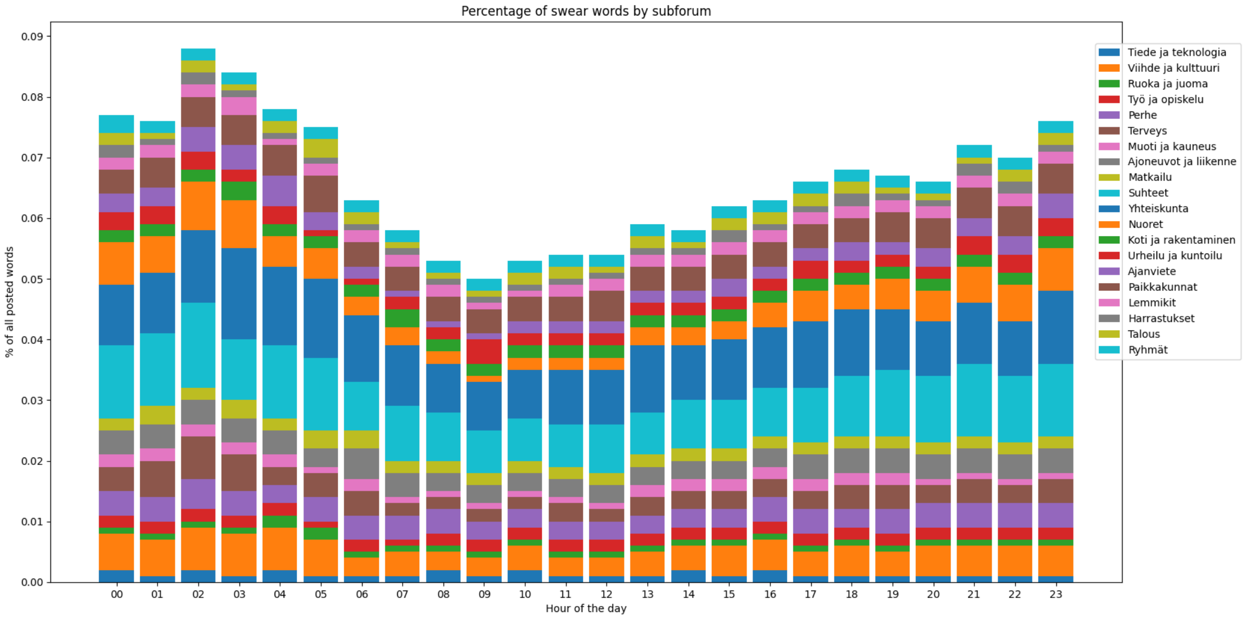
More visualization is also made of the different categories of swearing for the whole forum, and also separately for each subforum excluding the ones with too little data. The code for it can be found in 'code/process_categories.py' Here the amounts of words belonging to each of the given categories are counted and visualized as piecharts. The relative amounts of words belonging to the different categories is also visualized in a bar plot, where each category is divided by the total amount of instances from all categories posted by hour.

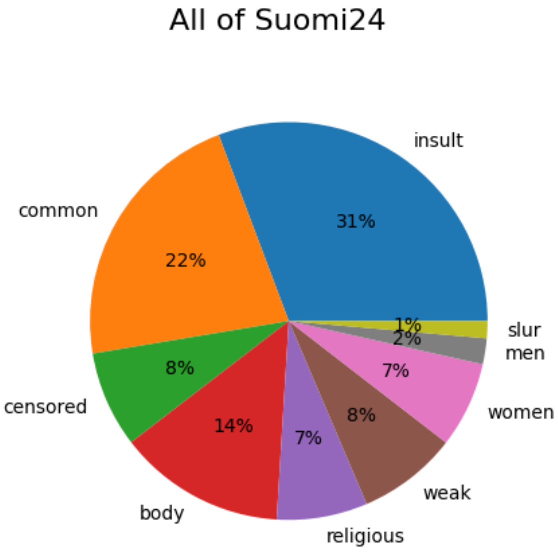
Note that the percentages and amounts concerning categories do not denote instances of words, but instances of a word belonging to a category, and a single word can yield at most three hits to different categories.

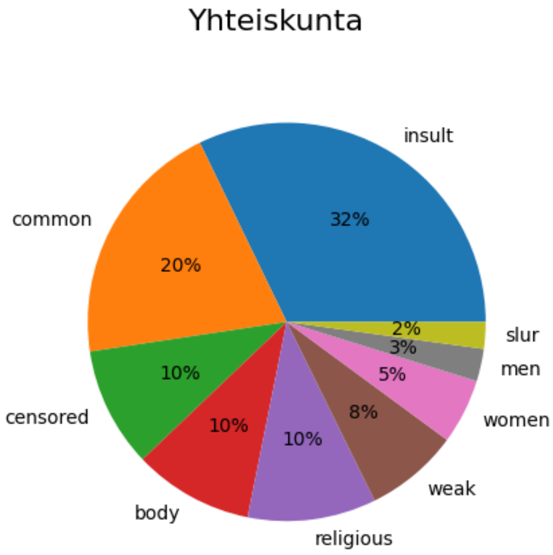
Analysis

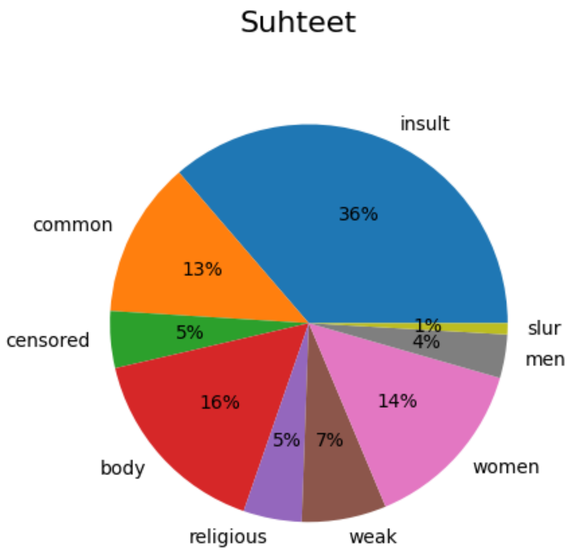
Once the occurrence of swear words has been proportioned to the total amounts of words, we can see that there is a drop in the frequency of occurrences around 7-12 in the morning, as can be seen in the bar plot on the next page. Note that the colors in the legend are ordered and in reverse order relative to the plot. The peak of swearing occurs at 2 in the morning, with high activity during the next three hours. Some subforums saw no change in swearing frequency by the hour, but for some, the change was clearly visible, like 'Nuoret' ("Youth") and 'Ajanviete' ("Entertainment"). It is also notable that 'Suhteet' ("Relationships") and 'Yhteiskunta' ("Society") had the most swearing, each over double the amount as the biggest category after them ('Viihde ja kulttuuri', "Entertainment and culture").

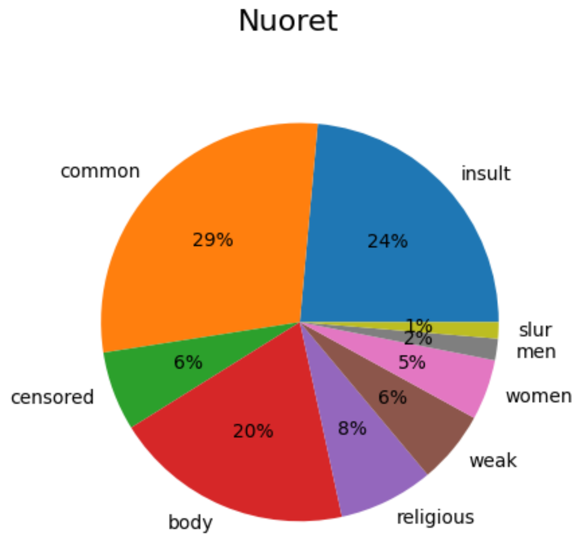
It could be said, that the hour of the day has an influence on how rude people are on Suomi24, but this isn't a very significant difference. The difference of the peak at 2 and the lowest point at 9 is only roughly one third, and the numbers are very small: at the peak of swearing, swear words make up just under 0.1% of all words posted. As the dataset was a sample, randomness may have influenced the result.

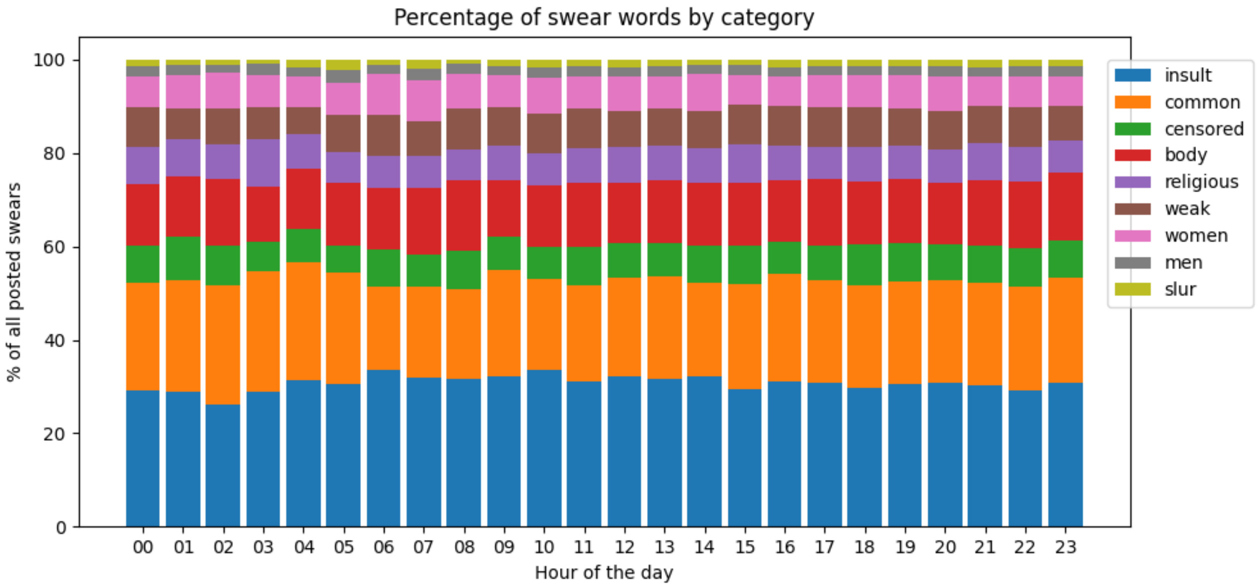












In addition to the amount of swearing, the project looked at how popular the different categories of swearing were. The three forums with approximately the most swearing were chosen for the piecharts. From the piecharts it can be seen that the majority of all swears on Suomi24 were insults, which is the biggest category for the subforums also excluding 'Nuoret'. On 'Suhteet', insults targeted at men and women are twice as popular as on all of the forums, highlighting a focus on gender and sexual relationships. On 'Nuoret', common swear words were the most popular category, and words of the Body & sexuality category were also relatively popular, which makes sense, as this subject is probably of especial interest to young people.

The results seem to show a correlation between late night and more swearing, though the reason for this cannot be definitely answered. Some literature suggests, that swearing is a sign of a higher IQ (Jay & Jay, 2015), and that people with higher IQs tend to stay awake late into the night (Kanazawa & Perina, 2009), and this could be an explanation: late at night there's very little social activities one can do, except post on the Internet, so perhaps the demographic using forums such as Suomi24 at night are people who tend to swear more often. This is, however, pure speculation.

The results also show that there isn't a lot of change in the categories the swears belong to based on the hour of the day, as can be seen in the last bar plot. There's a minimal increase in the percentage of 'Common' swears from 2 to 5 in the morning, but otherwise the proportions of the categories stay the same throughout the day.

Limitations, Ethics

My time and lack of expertise sets limits on this study.

My plan was to use all the Suomi24 data available from 2001-2020 (City Digital Group, 2021), but since I built my pipeline around the sample corpus from 2001-2014 (Aller Media ltd., 2014), the results I got with the larger dataset were so surprising and different, that I believe there might be some error in my pipeline, and I don't have the time to verify that results using that dataset are correct, unfortunately. If I have the time and energy I'll propably try to make the more complete dataset work, but I cannot make it before the course deadline.

The dataset used in this project is a sample, containing some posts from the Suomi24 forum. Information on how this sample was chosen isn't available, and its small size gives room for randomness to affect the results. This makes the results quite unreliable. The dataset is also from 2001-2014, and since the Internet evolves so rapidly, the state of Internet discussions in 2014 probably doesn't reflect the reality nowadays.

Other limitations include that the query method can yield duplicates, if swear words in Korp have been lemmatized. It appears that duplicates are not a concern with the querying method used here, but as I'm not entirely sure the query works in the same way in the api vs. the web Korp interface, this is still included. The methods used also do not yield every single instance of swearing, because the list of query words is limited, and possibly biased, excluding some words completely. It excludes certain inflected forms of the included words, for example. The results would be improved by large by using Korp's intelligent query language that makes use of lemmatization and regular expressions, but this would require much manual work with processing and pruning the enormous query word list.

The swear word list for querying might overrepresent some categories of words, and seriously underrepresent others. For example, the 'Insults to men' category contains only a few words, while 'Insults to women' is much larger. This is a result of how the 'common swears' were chosen: by my native understanding of what is a common swear in Finnish. This in turn reflects that there are not very many gendered, common swears or insults towards men, or they're less used than those towards women, at least in my understanding.

Whether the methods chosen to visualize the data are meaningful is also not certain, and aren't motivated by any specific reasons.

The research could be improved with more context from earlier research. Contrasting the results from the Suomi24 data with results from some other forum data, for example Korp's Ylilauta corpus, could yield information on whether the swearing habits of the userbases are similar or different, and could lead to speculations about the etiquette of the forums, and of the internet at large.

I don't think this project has any ethical concerns. The usernames associated with the forum posts that contain crude language are not collected as a part of the research, so one cannot easily identify who has made crude remarks on the forums based on my data. The corpus is freely and openly available and doesn't contain any sensitive information, even the usernames are frequently pseudonyms. I don't think this project can negatively impact the reputation of anyone, including the Suomi24 forum itself.

References

Aller Media ltd. (2014). *The Suomi 24 2001-2014 (Sample) Corpus, Helsinki Korp Version* [data set]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016050901

City Digital Group. (2021). *The Suomi 24 Corpus 2001-2020, VRT version* [data set]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2021101527

Delpeuch, A., Morris, T., Huynh, D., (bot), W., Mazzocchi, S., Jacky, ... Chandra, L. (2023, November). *Openrefine/openrefine: Openrefine v3.7.7.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.10220116 doi: 10.5281/zenodo.10220116

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Hämäläinen, M. (2019). *Python korp library (version v1).* Retrieved 15.12.2023, from http://doi.org/10.5281/zenodo.1143374

Jay, K. L., & Jay, T. B. (2015). Taboo word fluency and knowledge of slurs and general pejoratives: deconstructing the poverty-of-vocabulary myth. *Language Sciences*, *52*, 251-259. Retrieved from https://www.sciencedirect.com/science/article/pii/S038800011400151X (Slurs) doi: https://doi.org/10.1016/j.langsci.2014.12.003

Kanazawa, S., & Perina, K. (2009). Why night owls are more intelligent. *Personality and Individual Differences*, *47*(7), 685-690. Retrieved from https://www.sciencedirect.com/science/article/pii/S0191886909002177 doi: https://doi.org/10.1016/j.paid.2009.05.021

Kielipankki. (2021). *Kielipankki.* Author. Retrieved 20.12.2023, from https://www.kielipankki.fi/language-bank/

Stapleton, K., Beers Fägersten, K., Stephens, R., & Loveday, C. (2022). The power of swearing: What we know and what we don't. *Lingua*, *277*, 103406. Retrieved from https://www.sciencedirect.com/science/article/pii/S002438412200170X doi: https://doi.org/10.1016/j.lingua.2022.103406

Turku NLP group. (n.d.). *word2vec* [language model demo]. Retrieved 15.12.2023, from http://epsilon-it.utu.fi/wv_demo/