1.2.1 Anonymization

General privacy

People who write WhatsApp chats can be recognized either by the stories they tell or by the names and places they mention. While we had no way to address the former, we addressed the latter by means of computational linguistics. If you happen to recognize informants based on the remaining information, we ask you to comply with common research ethics and keep that knowledge to yourself.

First names

All first names found are based on freely available reference lists in the respective language. It was then decided to not actually remove first names, but to rotate them, meaning that the name Peter in a chat would not get replaced by e.g. [FirstName], but by e.g. Ferdinand. This procedure has several advantages:

1. The text remains easy to read.
2. Because Peter is always replaced with Ferdinand, all occurrences of the same name remain the same. Conversations can therefore be more easier followed.

Tests showed, that more than 95% of all first names were found and rotated in this way.

We tried to assign the same sex to the rotated names as to the original one, such as to keep the text readable. Very often, this resulted in good replacements, but some names are gender-neutral. *Andrea* e.g. is a female name in German but a male one in Italian and Romansh. *Alex* can be Alexander or Alexandra etc. So you might come across a message, where the male name *Andrea* was replaced by *Olivia*. This will result in weird sentences like *I saw Olivia yesterday, he wasn't looking well*. If you keep in mind that the names were rotated, you will not have a problem realizing that this is a man with a wrongly assigned replacement name.

Last names

Only very few last names can in fact be found in the data. Because of this limitation, the same procedure as with first names could not be applied, because additionally some of the last names used are very rare if not unique. It was therefor decided to replace all last names with [LastName] instead. In a combined effort of manually analyzing and means of computer linguistics, more than 95% of all last names were removed.

Numbers

In an effort to remove information about phone numbers, bank accounts etc., all numbers with three and more digits where removed and each digit was replaced with one N. The phone number 079 987 65 43 would thus become NNN NNN 65 43, while 0799876543 would be NNNNNNNNNN. Reliability

here lies at 100%.

E-Mail addresses

All email adresses were removed and replaced with xxx@yyy.ch, while keeping the number of characters. info@uzh.ch would therefore become xxxx@yyy.ch, while admin@google.com would become xxxxx@yyyyyy.com.

Street addresses

Street addresses were removed and replaced by [StreetAddress].

WWW addresses

WWW addresses were kept since they contain information publicly available.

City names

Names of cities were kept because they cannot be considered as private information and because they may be important for the understanding of the text.

From:
https://whatsup.linguistik.uzh.ch/ -

Permanent link:
https://whatsup.linguistik.uzh.ch/01_corpus/02_preprocessing/01_anonymization?rev=1586879043

Last update: **2022/06/27 07:21**