

De-identification of Dutch Mental Health Data

Erik Tjong Kim Sang¹, Ben de Vries¹, Wouter Smink²,
Bernard Veldkamp², Gerben Westerhof², Anneke Sools²

¹Netherlands eScience Center

²Department of Psychology, Health and Technology, University of Twente

e.tjongkimsang@esciencecenter.nl

UNIVERSITY
OF TWENTE.

Introduction

Mental health data often contain information revealing the identity of patients. This information must be removed before the data can be used for scientific research. In this paper we describe our efforts to anonymize or de-identify mental health texts written in Dutch. Similar work has earlier been done by Menger et al. (2017) and Scheurwegs et al. (2013).

Data

Our mental health data consist of emails from online therapies in which patients discuss their personal situation. These data cannot be shared. In order to facilitate an open evaluation process, we use Wikipedia biographies, which after an egofication process have been converted to autobiographies which are similar to our mental health text.

Method

We remove all **person names**, **locations**, organizations, **numbers**, **dates**, mail addresses, urls and **miscellaneous** named entities from text. For identifying these entities, we rely on the tool Frog (Van den Bosch et al. 2007) and a collection of entity identifying rules.

Example of processed text

PER

Ik (**LOC**, **DATE** - **LOC**, **DATE**) was een **MISC** kunstschilder, etser en tekenaar. Ik geldt algemeen als een van de grootste schilders en etsers in de **MISC** kunst, en als de belangrijkste **MISC** meester van de **DATE**. Ik vervaardigde in totaal ongeveer **NUM** schilderijen, **NUM** etsen en **NUM** tekeningen. Mijn werk behoort tot de barok en is zichtbaar beïnvloed door het caravaggisme, alhoewel ik nooit in **LOC** is geweest.

Evaluation results

	UNL	AVG	PER	NUM
TKS2019	84%	55%	70%	70%
Menger2017	40%	31%	55%	1%

Our system outperforms previous work (Menger et al. 2017) with respect to F1 scores on all evaluated classes and class combinations

References

Menger V, Scheepers F, Van Wijk L, Spruit M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics*. 2017;35:727736.

Scheurwegs E, Luyckx K, Van der Schueren F, Van den Bulcke T. De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study. In: *Proceedings of the Workshop on NLP for Medicine and Biology*. RANLP, Hissar, Bulgaria; 2013. p. 1823.