GeenStijl.nl embeddings

Pierre Voué, Elizabeth Cappon & Tom De Smedt

FEBRUARY 2021

TEXTGAIN TECHNICAL REPORTS • TGTR4 • ISSN 2684-4842

AUTHORS

- **Pierre Voué** (pierre@textgain.com) has a master's degree in Artificial Intelligence and works as a data scientist at Textgain, studying online hate speech, radicalization and polarization.
- **Elizabeth Cappon** has a master's degree in computational linguistics, and works as a data scientist at Textgain, studying online hate speech, with a specific interest in online sexism.
- **Tom De Smedt** has a PhD in Arts and is CTO at Textgain, focusing on Natural Language Processing. He was awarded the Research Prize of the Auschwitz Foundation in 2019.

ABSTRACT

We collected over 8M messages from the controversial Dutch websites GeenStijl and Dumpert to train a word embedding model that captures the toxic language representations contained in the dataset. The trained word embeddings (±150MB) are released for free and may be useful for further study on toxic online discourse: https://textgain.com/geenstijl.

Keywords: *natural language processing, word embedding, abusive language, Dumpert, GeenStijl*

1 INTRODUCTION

The growing presence of online hate speech and accompanying offline threats is an established fact (Europol, 2020), and Dutch-speaking communities are no exception (Brandsma, 2017). Besides the well-known mainstream social media platforms, more localized and typically less moderated websites also contribute to this phenomenon (Flew, Martin & Suzor, 2019). In continuation of a previous study on language representation on fringe platforms 4chan and 8chan (Voué, De Smedt, De Pauw, 2020), we collected data from two controversial Dutch websites and used it to train word embeddings. The objects of our study are the news blog geenstijl.nl (*geen stijl* meaning *no class* or *bad taste* in Dutch) and the video and image hosting platform dumpert.nl, which is a GeenStijl spin-off (Holst, 2019).

The news blog GeenStijl was founded in 2003 as a right-wing counter-voice against the established media. The site is known for its controversial opinion pieces (usually published under a pseudonym) with an often inciting and derisive undertone. GeenStijl can be seen as a Dutch precursor to the many alternative news platforms that have appeared on social media and in blogs in recent times. Over the years, it has attracted an active and often toxic community of commenters who are known for their creative slang. The community refers to itself as *reaguurders*, a contraction of the verb *reageren* (to react) and the adjective *guur* (grim). Such neologisms have gained popularity to the point that they eventually were included in the leading dictionary of the Dutch language; the Van Dale.[1] GeenStijl and its toxic idioms have been the subject of frequent criticism. For example, the codeword *Finnen* (Finns) is used to make racist statements about Moroccans without directly referring to Moroccans.[2]

In 2017, over 100 female journalists called on advertisers to boycott GeenStijl because of the rampant sexist content in the comment section, in particular after the site ran a poll asking the commenters "if they would fuck" Dutch journalist Loes Reijmer.[3] This led to parent company Mediahuis divesting the provocative website.[4] GeenStijl has continued independently and continues to exist to this day.

2 METHODS AND MATERIALS

Using data collection algorithms written in Python,[5] we collected 1.1M comments from GeenStijl and 7M comments from Dumpert, ranging respectively from May 2003 until June 2020 and May 2014 until June 2020. The corpus can be reconstructed upon motivated request.

These comments were used as training material to create word embeddings. Word embeddings, and the algorithms developed to produce them, are a well-known topic of research in Machine Learning (ML), a field of Artificial Intelligence (AI). Word embeddings are numerical representations (vectors) of words, derived from the context (i.e., other surrounding words) in which these words are observed. Using large datasets, we can obtain word vectors by counting word frequencies and the frequency at which they co-occur with other words. Under the assumption that words occurring in a similar context have related meanings, such vectors can then be compared with each other, thereby comparing their underlying representations in a given dataset. In the context of hate speech, this approach allows for the discovery of previously unknown toxic expressions by querying the model with known words.

[1] https://www.geenstijl.nl/3956951/reaguurders_u_staat_eindelijk

[2] https://www.geenstijl.nl/2262001/finnen_waar_komen_ze_vandaan

[3] https://www.demorgen.be/nieuws/marktwaarde-van-seksisme-daalt~b13ed4f5

[4] https://www.tijd.be/ondernemen/media-marketing/mediahuis-stoot-geenstijl-af/10048275.html

[5] https://github.com/textgain/grasp

To train our model, we used the Python library gensim (Mikolov et al., 2013), and more specifically its two main algorithms: CBOW (Continuous Bag of Words) and skipgram. In brief, the difference between both is that CBOW will predict a word by its surrounding words, while skipgram predicts the context from a given word. CBOW is slightly more accurate for more frequent words, while skipgram performs well on rarer words. We trained both models, each generating word vectors of size 100 from a window of 7 context words, where each target word had to occur at least 15 times in the dataset to be included. The CBOW model is available for free.

3 RESULTS

The obtained representational models can be used to examine known trends or to discover new ones. For example, we can explore associations with the pejorative prefix *deug-* (e.g., *deugers*, *deugmensen*; *do-gooders*) by querying the model for the 10 most related words.

The top 10 words related to *deuger* are shown with their associated cosine similarity, which can be understood as a "relatedness score" from 0.0 (totally unrelated) to 1.0 (entirely similar). This exposes expressions like *sjw* (*social justice warrior*), *activist*, *politically correct*, and *dumb left-wing*, which might be telling of the ideological mindset of GeenStijl reaguurders:

	CBOW		**SKIP GRAM**
0.61	*sjw*	0.70	*gutmensch*
0.59	*policor*	0.70	*policor*
0.59	*linkse*	0.68	*linkse*
0.59	*linksche*	0.67	*politiek correcte*
0.59	*goedmensch*	0.67	*sjw*
0.57	*linksche gutmensch*	0.67	*linksche*
0.57	*gutmensch*	0.65	*deugen*
0.56	*politiek correcte*	0.64	*moralistische*
0.56	*extreem linkse*	0.63	*activistisch*
0.55	*domlinkse*	0.62	*deugmensen*

Words and word combinations discovered using the models can then be used to for example analyze other social media, expanding the corpus, iteratively discovering yet more words to train more robust models. To give an example, querying Twitter for *deugmensen* yields the following (removed) tweet:

"OK #deugeiser. 4e categorie der #deugmensen naast #deugpronker #deugdrammer en #deugdreiger of #deugeisen #deugdrammen #deugpronken #deugdreigen. Deugmensen deugen niet. #woke".

This message contains new, unknown *deug-* compounds that we can use as search queries, to monitor how toxic language evolves or to expand the words embeddings. Essentially, models could update themselves with weak human supervision to stay on top of polarizing trends.

4 ANALYSIS

Another way of exploring the embeddings is to visualize them as a network diagram, with words as nodes and their relations as edges. The diagram below shows the model centered on the entry *kutwijf*. (*cunt*). One striking feature is the rapid drift from fairly neutral words like *meisje* (*girl*) to quite belittling words such as *grietje* (*chick*), to *mokkel* (*broad*), *wijf* (*bitch*), *viswijf* (*hysterical bitch*) and *kutwijf*, and various related highly-offensive slurs. Such diagrams facilitate the visualization of many interrelated words at once and offer insights into language use and world representation typical of the dataset that was used to generate it. Possible use cases include comparing connections found herein with connections found in a network trained on data from neutral sources, such as regulated media outlets, which would help determine the degree of sexism present in unregulated platforms.

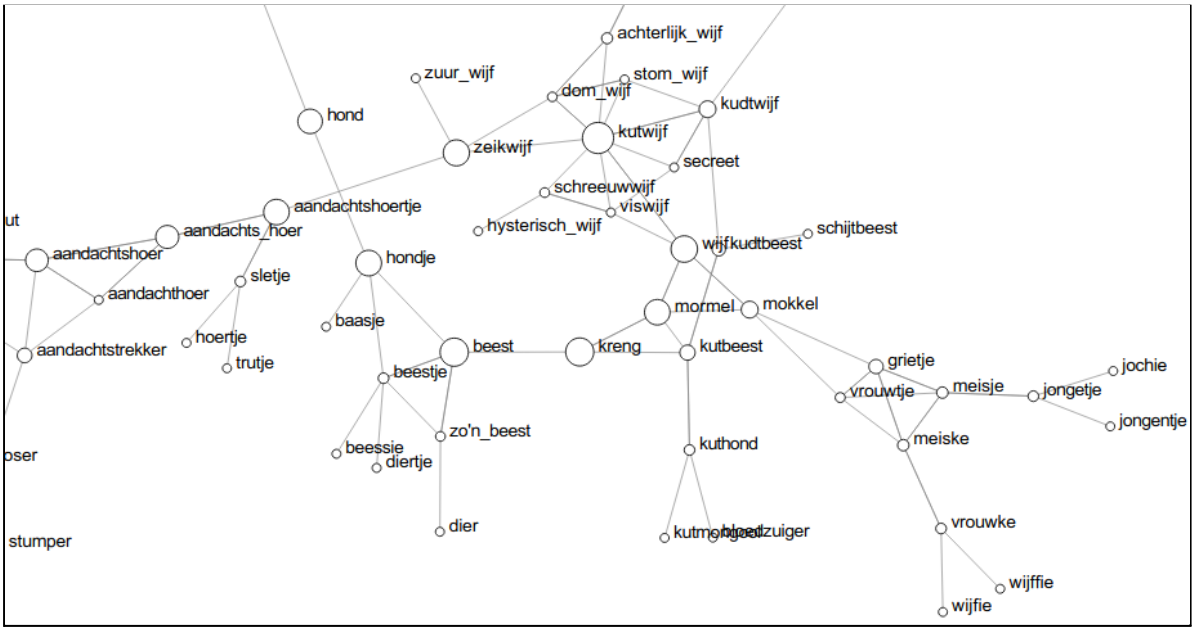

Figure 1. Related words in the model represented as a network diagram.

REFERENCES

Brandsma, B. (2017). Polarisatie: inzicht in de dynamiek van wij-zij denken. BB in Media.

Europol (2020). European Union Terrorism Situation and Trend report (TESAT 2020). Retrieved from https://www.europol.europa.eu/tesat-report

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy, 10*(1), 33-50.

Holst, D.E. van der (2019). Dumpert: een broeiplaats van vrouwonvriendelijk gedrag? Bachelor thesis. Faculty of Humanities of Utrecht University.

Řehůřek, R., & Sojka, P. (2011). Gensim – statistical semantics in python. Retrieved from https://gensim.org

Voué P., De Smedt T., & De Pauw G. (2019). 4chan & 8chan embeddings. *Textgain Technical Reports 1*.