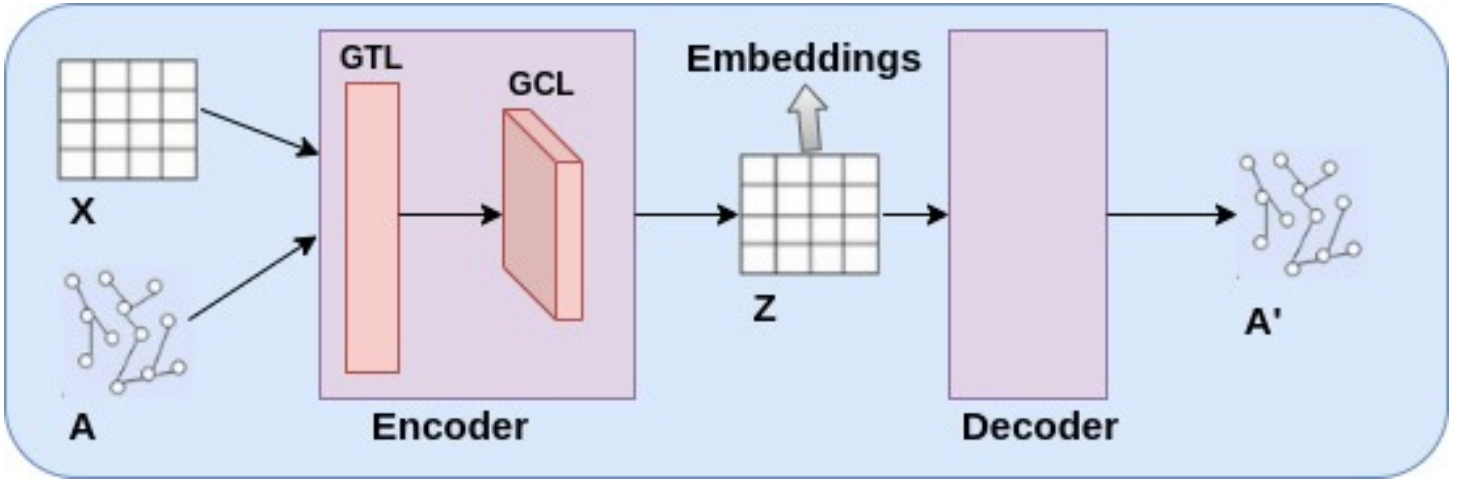
Unsupervised Embeddings with Graph Auto-Encoders for Multi-domain and Multilingual Hate Speech Detection

Gretel Liz De la Peña Sarracén[1,2]
Paolo Rosso[1]

[1]Universitat Politècnica de València
[2]Symanto Research

LREC, June 2022

Graph Auto-Encoders



Our model - GAE

XHate-999 dataset

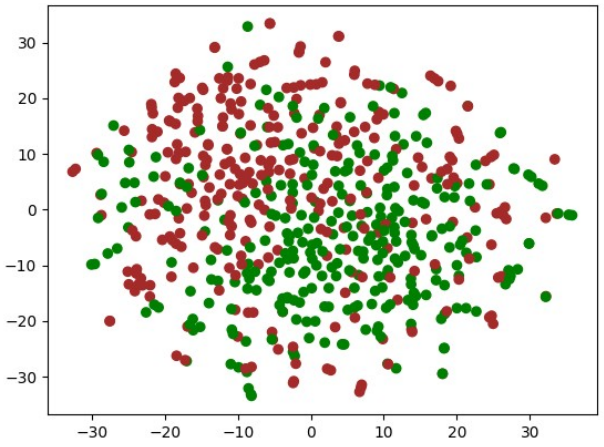
6 languages:

- English (EN)
- German (DE)
- Russian (RU)
- Turkish (TR)
- Croatian (HR)
- Albanian (SQ)

For each language there are 3 domains:

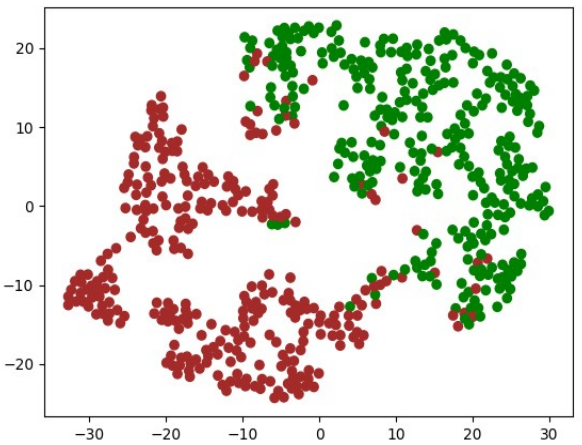
- Fox News (Gao) with 99 samples
- Twitter/Facebook (Trac) with 300 samples
- Wikipedia (Wul) with 600 samples

Analysis of Latent Representation - Embeddings

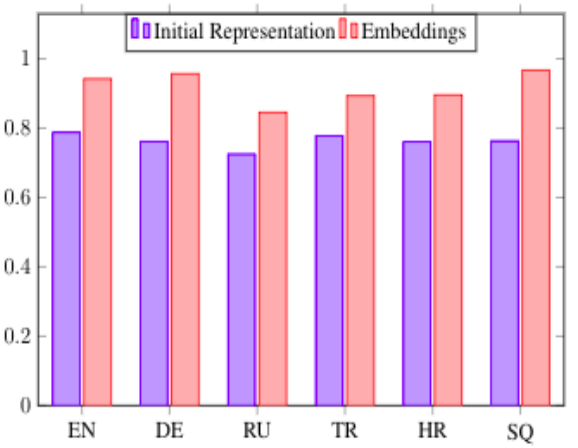


Initial representation

English - Wul

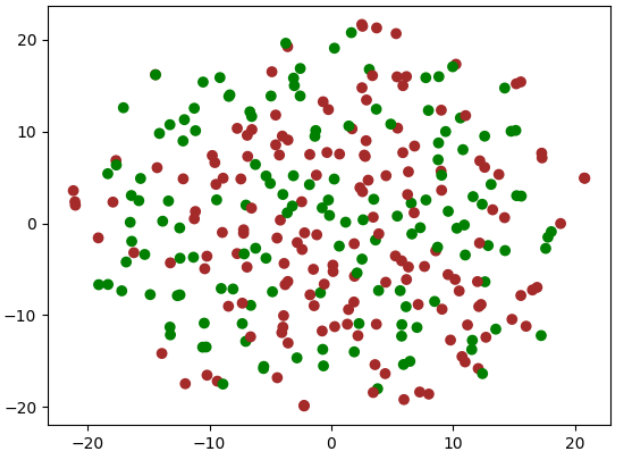


Embeddings



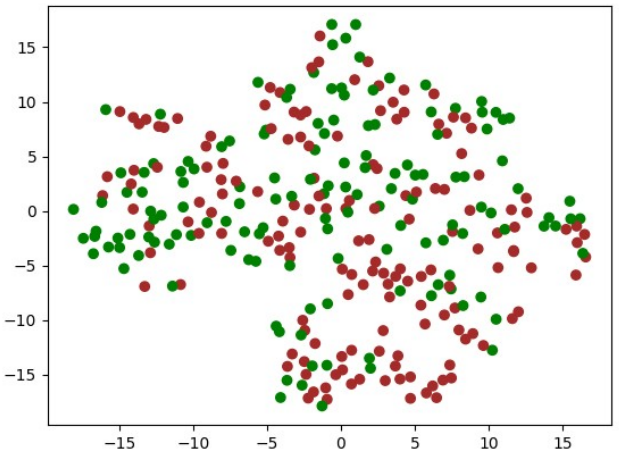
F1 in Wul.
Hate Speech Detection

Analysis of Latent Representation - Embeddings

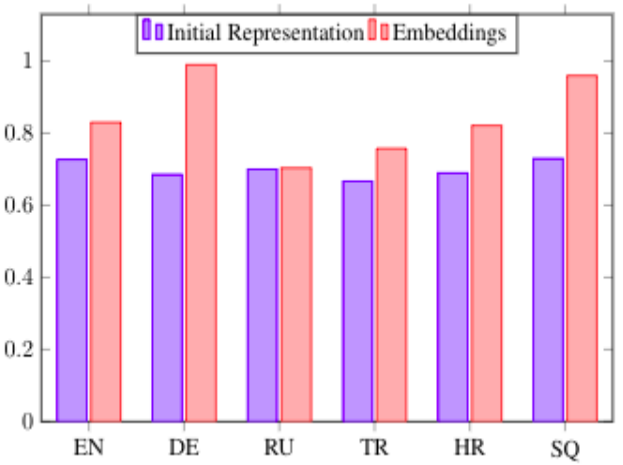


Initial representation

Russian - Trac



Embeddings



F1 in Trac.
Hate Speech Detection

Evaluation for Hate Speech Detection

	GAE		
	Gao	Trac	Wul
EN	$0.9714_{.018}$	$0.8301_{.025}$	$0.9424_{.007}$
DE	$0.9565_{.022}$	$0.9900_{.017}$	$0.9569_{.011}$
RU	$0.6667_{.014}$	$0.7238_{.010}$	$0.8453_{.014}$
TR	$0.7826_{.008}$	$0.7567_{.030}$	$0.8936_{.021}$
HR	$0.8799_{.030}$	$0.8214_{.023}$	$0.8958_{.017}$
SQ	$0.9565_{.011}$	$0.9600_{.011}$	$0.9673_{.026}$
	XLM-R		
	Gao	Trac	Wul
EN	$0.5909_{.041}$	$0.7245_{.022}$	$0.8538_{.034}$
DE	$0.6857_{.031}$	$0.7272_{.014}$	$0.8635_{.049}$
RU	$0.5754_{.009}$	$0.7070_{.005}$	$0.8384_{.041}$
TR	$0.5171_{.011}$	$0.7371_{.017}$	$0.8199_{.036}$
HR	$0.5050_{.047}$	$0.6377_{.041}$	$0.8384_{.047}$
SQ	$0.5642_{.041}$	$0.7148_{.016}$	$0.8231_{.041}$

F1 in Multi-domain Hate Speech Detection

	All		
	GAE	XLM-R	mBERT
EN	$\mathbf{0.8333}_{.010}$	$0.5642_{.047}$	$0.5111_{.053}$
DE	$\mathbf{0.9565}_{.014}$	$0.4545_{.038}$	$0.4850_{.045}$
RU	$\mathbf{0.8333}_{.001}$	$0.4923_{.061}$	$0.4527_{.031}$
TR	$\mathbf{0.8799}_{.004}$	$0.6192_{.043}$	$0.3864_{.057}$
HR	$\mathbf{0.8461}_{.012}$	$0.6459_{.039}$	$0.4545_{.044}$
SQ	$\mathbf{0.9565}_{.002}$	$0.4978_{.021}$	$0.4457_{.046}$

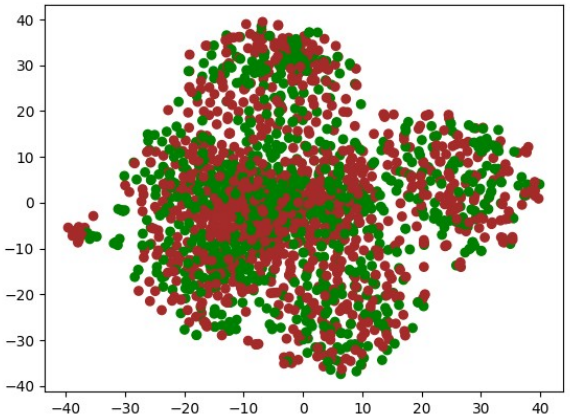
F1 in Multi-domain Hate Speech Detection.
(All domains to train)

Multilingual Evaluation

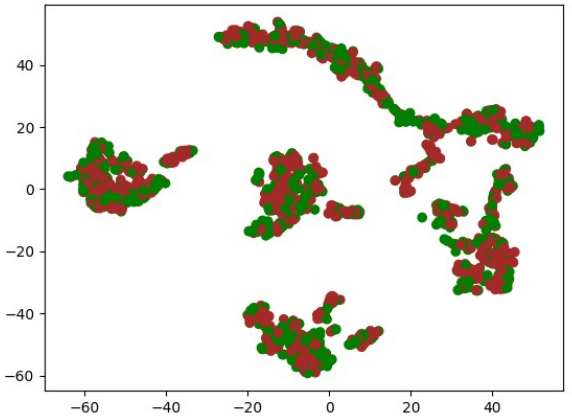
Domain	Gao	Trac	Wul
GAE	$0.3972_{.090}$	$0.6858_{.062}$	$0.6255_{.058}$
GAE-USE	$\mathbf{0.9308}_{.011}$	$\mathbf{0.9598}_{.005}$	$\mathbf{0.9491}_{.021}$
mBERT	$0.7047_{.054}$	$0.7952_{.080}$	$0.8939_{.072}$
XLM-R	$0.7349_{.015}$	$0.8585_{.012}$	$0.9303_{.008}$

F1 in Multilingual Hate Speech Detection

Multilingual Evaluation

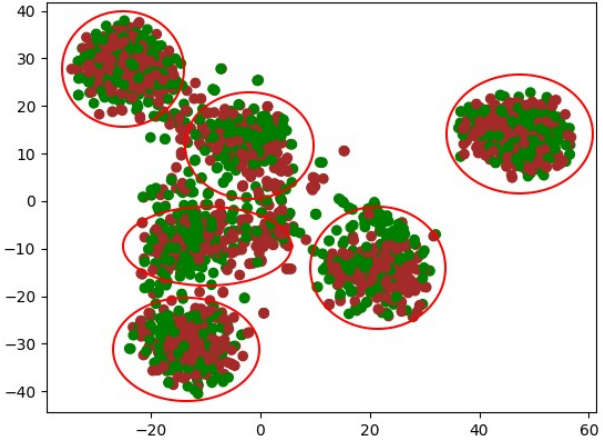


Initial representation with TFIDF

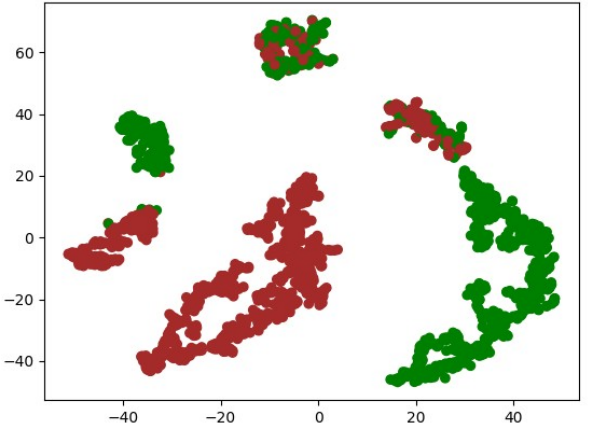


Embeddings

Multilingual Evaluation



Initial representation with Universal Sentences Encoder



Embeddings

Conclusion

- We generated promising embeddings to discriminate between hateful and non-hateful texts, by using the graph auto-encoder model.

- The model, based on the generated embeddings, achieved good results compared to mBERT and XLM-R.

Unsupervised Embeddings with Graph Auto-Encoders for Multi-domain and Multilingual Hate Speech Detection

Gretel Liz De la Peña Sarracén[1,2]
Paolo Rosso[1]

[1]Universitat Politècnica de València
[2]Symanto Research

LREC, June 2022