Comparative Study on Word Embeddings and Social NLP Tasks

Fatma Elsafoury[1], Steven R. Wilson[2], and Naeem Ramzan[1]

[1]University of the West of Scotland, [2]Oakland University

1.Research Problem

Figure 1: A crowdsourced definition of the word "Feminism" on Urban Dictionary

- **Grey social media platforms**, like Urban Dictionary and 4 & 8 Chan, are those with a loose moderation policy and hence they are rife with offenses.
- **Research problem:** Some word embeddings are pre-trained on data collected from grey social media platforms but they have not been investigated for the social related NLP tasks.
- **In this paper,** we carried out a comparative study between social-media-based and non-social-media-based word embeddings on two social NLP tasks: Detecting cyberbullying and Measuring social bias.

2. Word Embeddings

Informational-based word embeddings

Models pre-trained on informational data like Google News and Wikipedia articles.

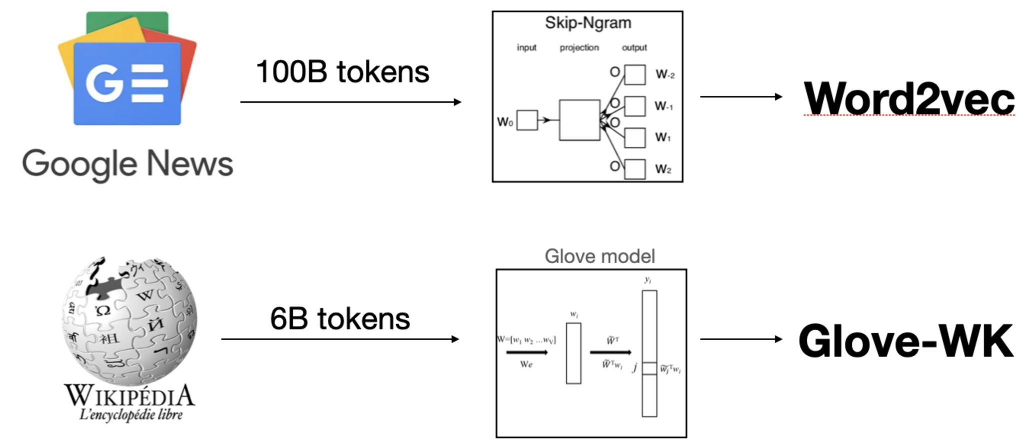

Figure 2: The used Informational-based word embeddings and how they are pre-trained.

Social-media-based word embeddings

Models pre-trained on informational data like Google News and Wikipedia articles.

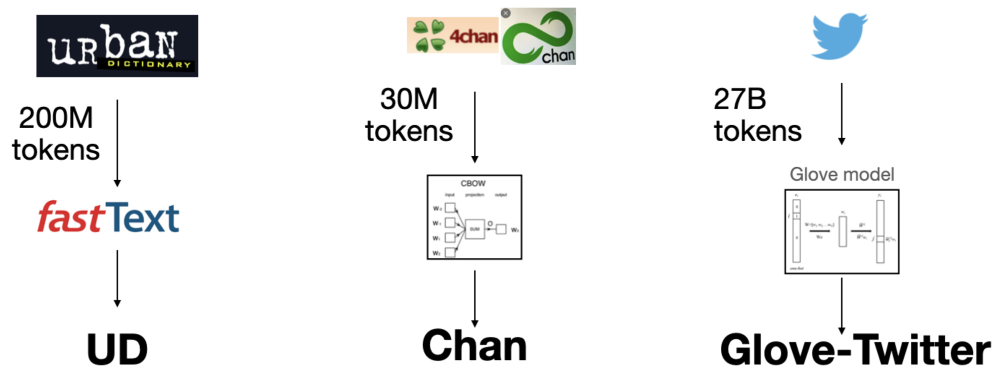

Figure 3: The used Social-media-based word embeddings and how they are pre-trained.

3. Cyberbullying Detection

Category	Description
PS	ethnic slurs
IS	words related to social and economic disadvantage
QAS	descriptive words with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
PR	words related to prostitution
OM	words related to homosexuality
ASF	female genitalia
ASM	male genitalia
DDP	cognitive disabilities
DDF	physical disabilities

Table1: Hurtlext 11 offenses categories

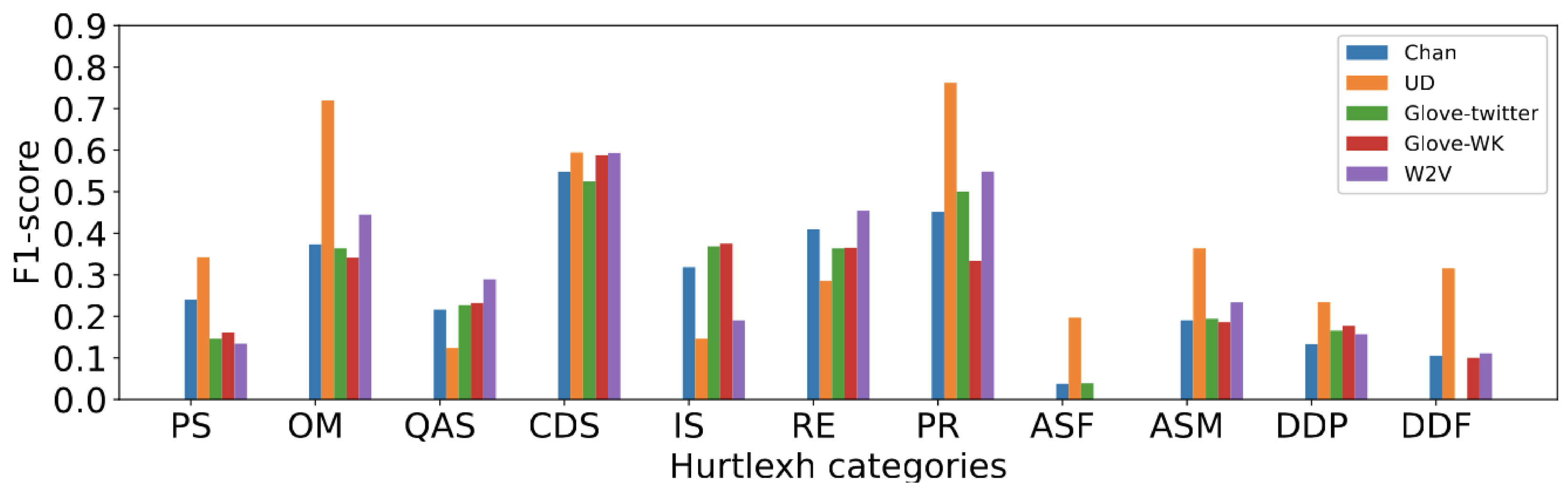

Figure 4: F1 scores of the KNN model with the different word embeddings on Hurtlext test set.

Table 2: The performance (F1 scores) of the BiLSTM model with each word embeddings On the different Hurtlex category within our cyberbullying datasets

	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
HateEval														
Chan	0.615	0.444	0.615	**0.666**	0.555	**0.647**	0.658	0.421	0.555	**0.857**	0.5	0.570	0.730	0.602
UD	**0.7**	0.444	0.571	0.603	0.533	0.562	0.678	0.4	0.603	0.571	0.375	0.508	0.734	0.560
Glove-Twitter	0.695	**0.5**	**0.736**	0.663	0.631	0.619	**0.711**	0.620	0.690	0.571	0.285	**0.605**	**0.738**	**0.620**
Glove-WK	0.583	0.222	0.571	0.616	**0.666**	0.515	0.614	**0.72**	0.691	**0.857**	0.333	0.535	0.699	0.586
W2V	0.315	**0.5**	0.666	0.648	0.631	0.514	0.614	0.714	**0.72**	0.571	**0.666**	0.593	0.705	0.604
Kaggle														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.380	**0.777**	**1**	**0.760**	0.571	**0.545**	0.571	**1**	**0.666**	0.916	**0.909**	0.571	0.783	**0.727**
UD	**0.72**	0.761	**1**	0.703	**0.75**	0.461	0.75	0.666	0.507	0.888	0.8	**0.611**	**0.813**	0.725
Glove-Twitter	0.454	0.727	0.444	0.627	0.727	0.285	**0.823**	0	0.520	**0.923**	0.8	0.513	0.790	0.587
Glove-WK	0.5	0.625	**1**	0.588	0.666	0.5	0.666	0.666	0.507	0.869	0.666	0.525	0.8	0.660
W2V	0.352	0.375	**1**	0.602	0.25	0.4	0.714	**1**	0.526	0.818	0.666	0.479	0.797	0.614
Twitter-sexism														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.666	0.829	0.421	0.523	0.695	0.4	0.45	0.6	0.510	0.666	0.56	0.561	0.586	0.574
UD	0.666	0.8	0.521	0.656	0.75	0.510	0.608	**0.923**	0.622	**0.75**	**0.687**	0.629	0.695	0.678
Glove-Twitter	0.666	**0.863**	0.380	0.640	**0.8**	0.5	0.693	**0.923**	**0.653**	0.571	0.645	0.631	0.702	0.667
Glove-WK	0.666	0.818	**0.608**	**0.686**	0.740	**0.655**	0.734	0.727	0.636	**0.75**	0.685	**0.675**	0.708	**0.699**
W2V	**0.727**	0.772	0.571	0.598	0.695	0.56	**0.769**	0.833	0.623	**0.75**	0.666	0.650	**0.730**	0.688
Twitter-racism														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	**0.76**	0.736	0.8	0.732	0.5	**0.809**	**0.4**	0	0.428	0.588	**1**	0.671	0.784	0.631
UD	0.754	**0.956**	**0.909**	0.762	**0.6**	0.8	0.333	0	0.571	0.583	0.909	0.658	0.783	**0.663**
Glove-Twitter	0.72	0.8	**0.909**	0.734	0.5	0.790	**0.4**	0	**0.666**	0.636	0.909	**0.694**	**0.813**	0.659
Glove-WK	0.703	0.8	0.833	**0.784**	0.5	0.793	0.333	0	0.615	**0.761**	0.769	0.688	0.800	0.644
W2V	0.680	0.588	0.75	0.622	0.571	0.767	0.333	0	0.545	0.631	0.8	0.654	0.748	0.591
Jigsaw-Toxicity														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.15	0.45	**0.461**	0.427	**0.5**	0.310	0.285	0.75	0.652	0.553	0.482	0.484	0.658	0.474
UD	0.303	**0.615**	0.387	**0.441**	0.333	0.274	0.285	0.666	0.653	0.461	0.538	0.449	0.666	0.467
Glove-Twitter	0.285	0.578	0.322	0.433	0.444	0.360	0.444	**0.888**	**0.693**	0.553	**0.571**	**0.493**	**0.687**	**0.519**
Glove-WK	0.166	0.514	0.428	0.362	0.428	**0.407**	0.25	0.75	0.615	0.558	0.363	0.454	0.661	0.458
W2V	**0.333**	0.437	0.230	0.421	0.333	0.350	**0.545**	0.571	0.543	**0.588**	0.518	0.448	0.678	0.461

4. Measuring social bias

	Gender Bias				Racial Bias			
Word embeddings	WEAT	RNSB	RND	ECT	WEAT	RNSB	RND	ECT
Word2vec	4 (0.778)	2 (0.033)	2 (0.087)	4 (0.752)	2 (0.179)	1 (0.095)	1 (0.151)	4 (0.786)
Glove-WK	**5 (0.893)**	4 (0.052)	4 (0.204)	2 (0.829)	**5 (0.439)**	2 (0.118)	4 (0.253)	1 (0.903)
Glove-Twitter	2 (0.407)	3 (0.041)	3 (0.127)	1 (0.935)	4 (0.275)	3 (0.122)	2 (0.179)	2 (0.898)
UD	1 (0.346)	1 (0.031)	1 (0.051)	**5 (0.652)**	1 (0.093)	4 (0.132)	3 (0.196)	**5 (0.726)**
Chan	3 (0.699)	**5 (0.059)**	**5 (1.666)**	3 (0.783)	3 (0.271)	**5 (0.299)**	**5 (2.572)**	3 (0.835)

Table3: The Bias scores using the different metrics of the different word embeddings.

5. Take Away Messages

1. Social-media-based word embeddings are better at cyberbullying detection and offenses categorization.
2. No certain word embeddings are better than others at detecting certain offensive categories.
3. Social-media-based word embeddings are **not** more socially biased than informational-based word embeddings.