

A comparative study on using word embeddings to detect different types of cyberbullying

Fatma Elsafoury, Steve R. Wilson, and Naeem Ramzan

Cyberbullying detection

- ▶ What is cyberbullying?
 - ▶ Spreading insults using an electronic medium.

Cyberbullying detection

- ▶ What is cyberbullying?
 - ▶ Spreading insults using an electronic medium.
- ▶ Why detect cyberbullying?
 - ▶ Support victims, warn/block bullies.

Cyberbullying detection

- ▶ What is cyberbullying?
 - ▶ Spreading insults using an electronic medium.
- ▶ Why detect cyberbullying?
 - ▶ Support victims, warn/block bullies.
- ▶ What are the types of cyberbullying?
 - ▶ The most common forms are: Racism, sexism, aggression, and toxicity.

Cyberbullying detection

- ▶ What is cyberbullying?
 - ▶ Spreading insults using an electronic medium.
- ▶ Why detect cyberbullying?
 - ▶ Support victims, warn/block bullies.
- ▶ What are the types of cyberbullying?
 - ▶ The most common forms are: Racism, sexism, aggression, and toxicity.
- ▶ Detecting cyberbullying
 - ▶ Mostly used Classic word embeddings: which are word embedding models that are pre-trained on formal text like news articles like word2vec or Wikipedia article like Glove.

Cyberbullying detection

- ▶ What is cyberbullying?
 - ▶ Spreading insults using an electronic medium.
- ▶ Why detect cyberbullying?
 - ▶ Support victims, warn/block bullies.
- ▶ What are the types of cyberbullying?
 - ▶ The most common forms are: Racism, sexism, aggression, and toxicity.
- ▶ Detecting cyberbullying
 - ▶ Mostly used Classic word embeddings: which are word embedding models that are pre-trained on formal text like news articles like word2vec or Wikipedia article like Glove.
 - ▶ Recently, there have been models that were trained on less formal text which we call here "Slang-based" word embeddings.

Slang-based word embeddings

- ▶ Slang-based word embeddings
 - ▶ Word embedding models pre-trained on text collected from social media platforms.

Slang-based word embeddings

- ▶ Slang-based word embeddings
 - ▶ Word embedding models pre-trained on text collected from social media platforms.
- ▶ Why use slang-based word embeddings?
 - ▶ No moderation social media platforms e.g. 4Chan , and Urban Dictionary
 - ▶ Abusive and Hateful content.

Slang-based word embeddings

- ▶ Slang-based word embeddings
 - ▶ Word embedding models pre-trained on text collected from social media platforms.
- ▶ Why use slang-based word embeddings?
 - ▶ No moderation social media platforms e.g. 4Chan , and Urban Dictionary
 - ▶ Abusive and Hateful content.
- ▶ Hypothesis:
 - ▶ Slang-based word embeddings perform better than classic word embeddings on the task of cyberbullying detection

Slang-based vs. Classic word embeddings

Binary F1-scores of Bi-LSTM using the different word embeddings on different datasets


	Slang-Based			Classic	
	Chan	UD	Glove-Twitter	Glove-Wikipedia	Word2Vec
HateEval (Hateful)	0.602	0.560	0.620	0.586	0.604
Kaggle (insults)	0.727	0.725	0.587	0.660	0.614
Twitter (racism)	0.631	0.663	0.659	0.644	0.591
Jigsaw (Toxicity)	0.474	0.467	0.519	0.458	0.461
Twitter (sexism)	0.574	0.678	0.667	0.699	0.688

Slang-based vs. Classic word embeddings

Binary F1-scores of Bi-LSTM using the different word embeddings on different datasets

	Chan	UD	Glove-Twitter	Glove-Wikipedia	Word2Vec
HateEval (Hateful)	0.602	0.560	0.620	0.586	0.604
Kaggle (insults)	0.727	0.725	0.587	0.660	0.614
Twitter (racism)	0.631	0.663	0.659	0.644	0.591
Jigsaw (Toxicity)	0.474	0.467	0.519	0.458	0.461
Twitter (sexism)	0.574	0.678	0.667	0.699	0.688

For 4 datasets, **slange-based** embeddings is the **best performing**



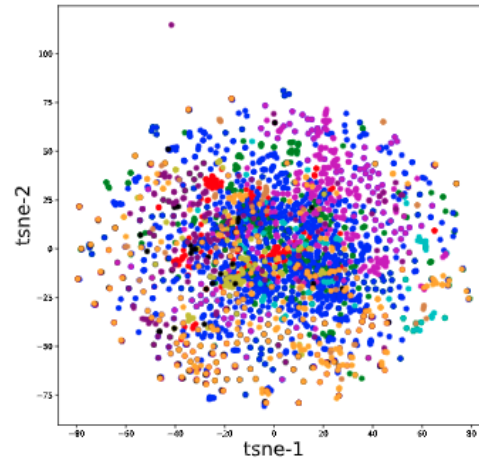
Detecting certain types of cyberbullying (Hurtlex)

- ▶ **Hurtlex lexicon** a multilingual lexicon containing 8228 offensive words and expressions, which are organized into 17 groups.
- ▶ We used only English lexicon and 11 groups.

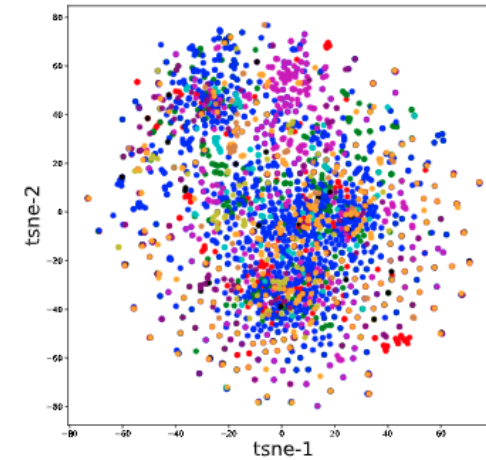
Label	Desc.	Label	Desc.
PS	negative stereotypes ethnic slurs	ASF	female genitalia
DDF	physical disabilities and diversity	PR	words related to prostitution
DDP	cognitive disabilities and diversity	OM	words related to homosexuality
IS	words related to social and economic disadvantage	QAS	with potential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

Detecting certain types of cyberbullying (Hurtlex)

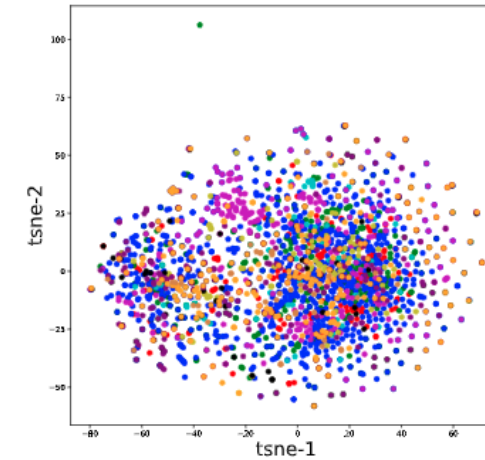
t-SNE of the different word embeddings of the words that belong to different groups in Hurtlex lexicon



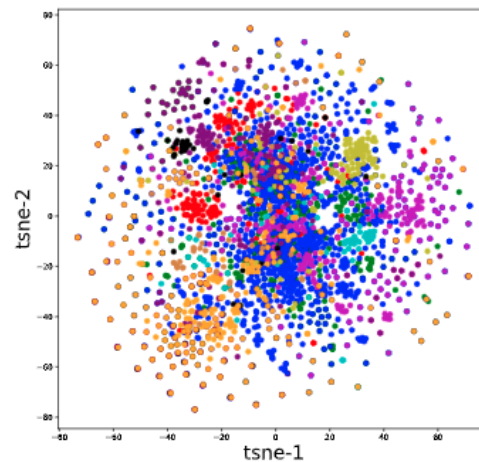
(a) W2v



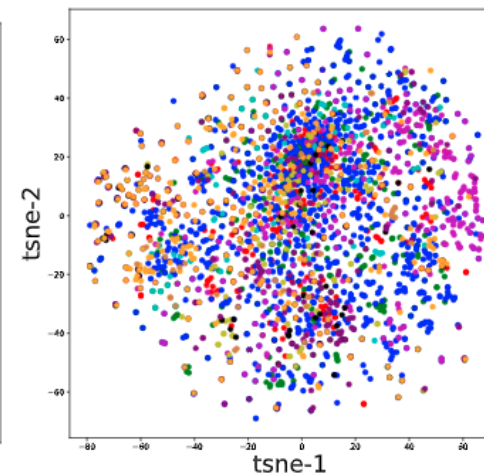
(b) Glove-WK



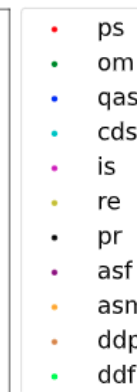
(c) Glove-Twtr



(d) UD

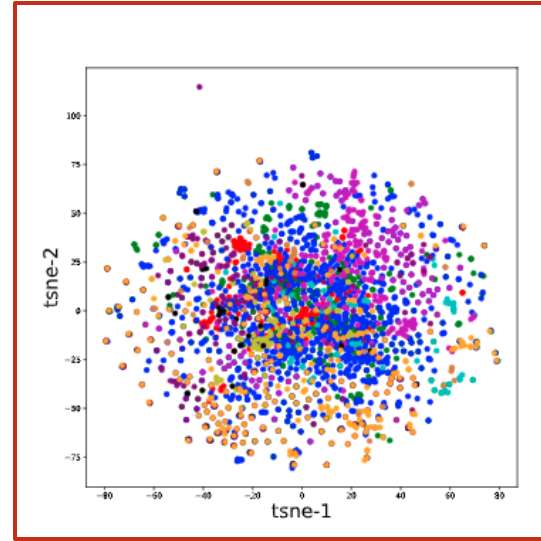


(e) Chan

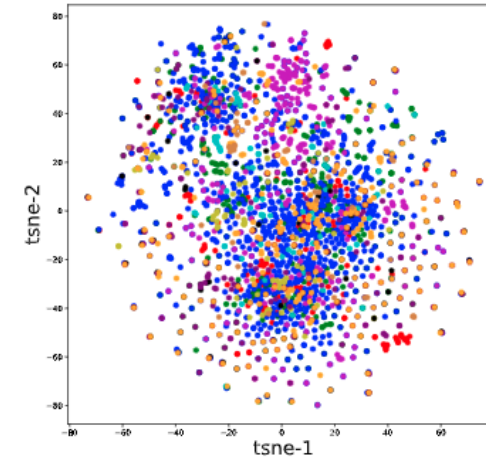


Detecting certain types of cyberbullying (Hurtlex)

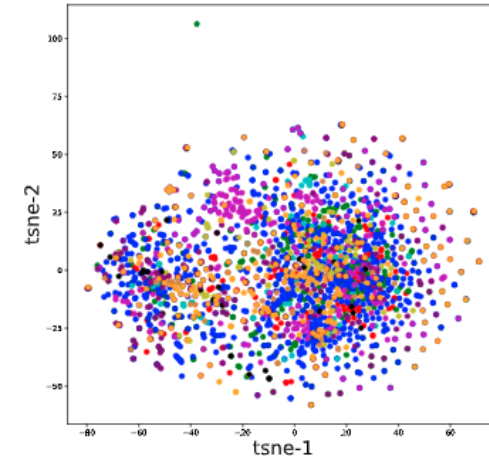
t-SNE of the different word embeddings of the words that belong to different groups in Hurtlex lexicon



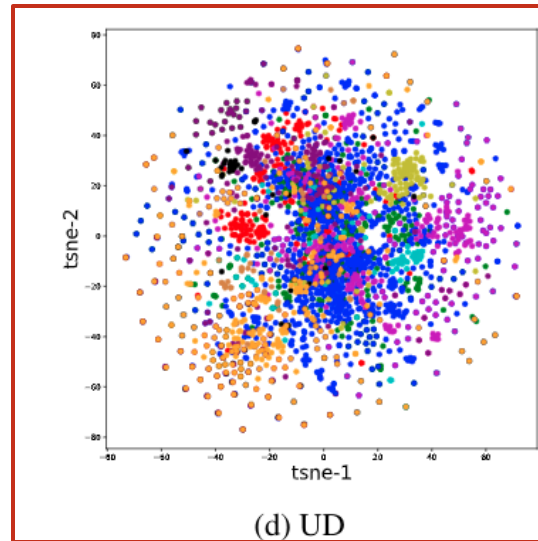
(a) W2v



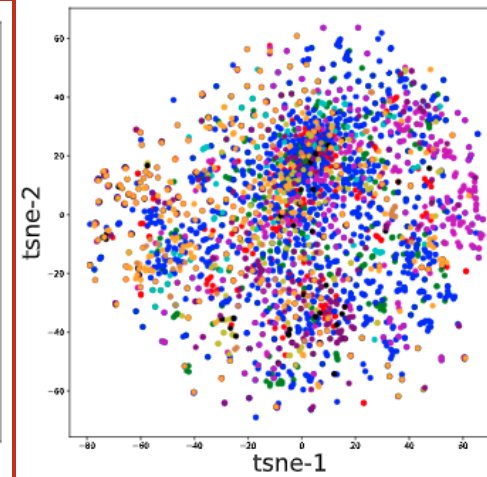
(b) Glove-WK



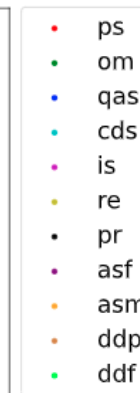
(c) Glove-Twtr



(d) UD

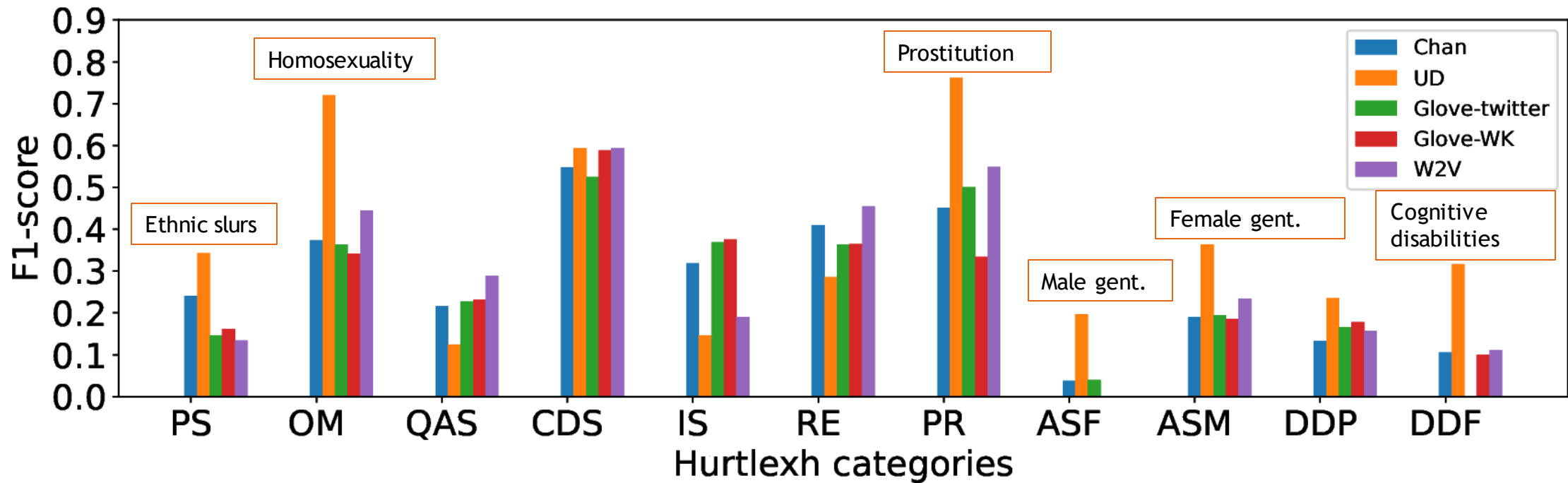


(e) Chan



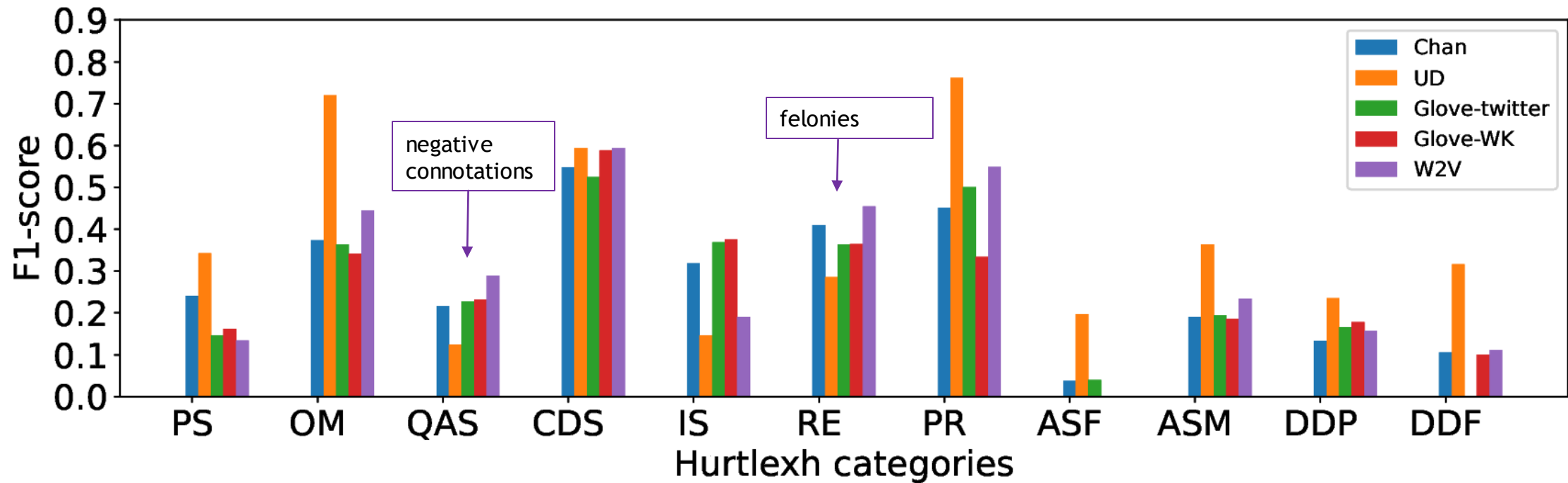
Detecting certain types of cyberbullying (Hurtlex)

F1 scores of the KNN model with the different word embeddings on Hurtlex test set



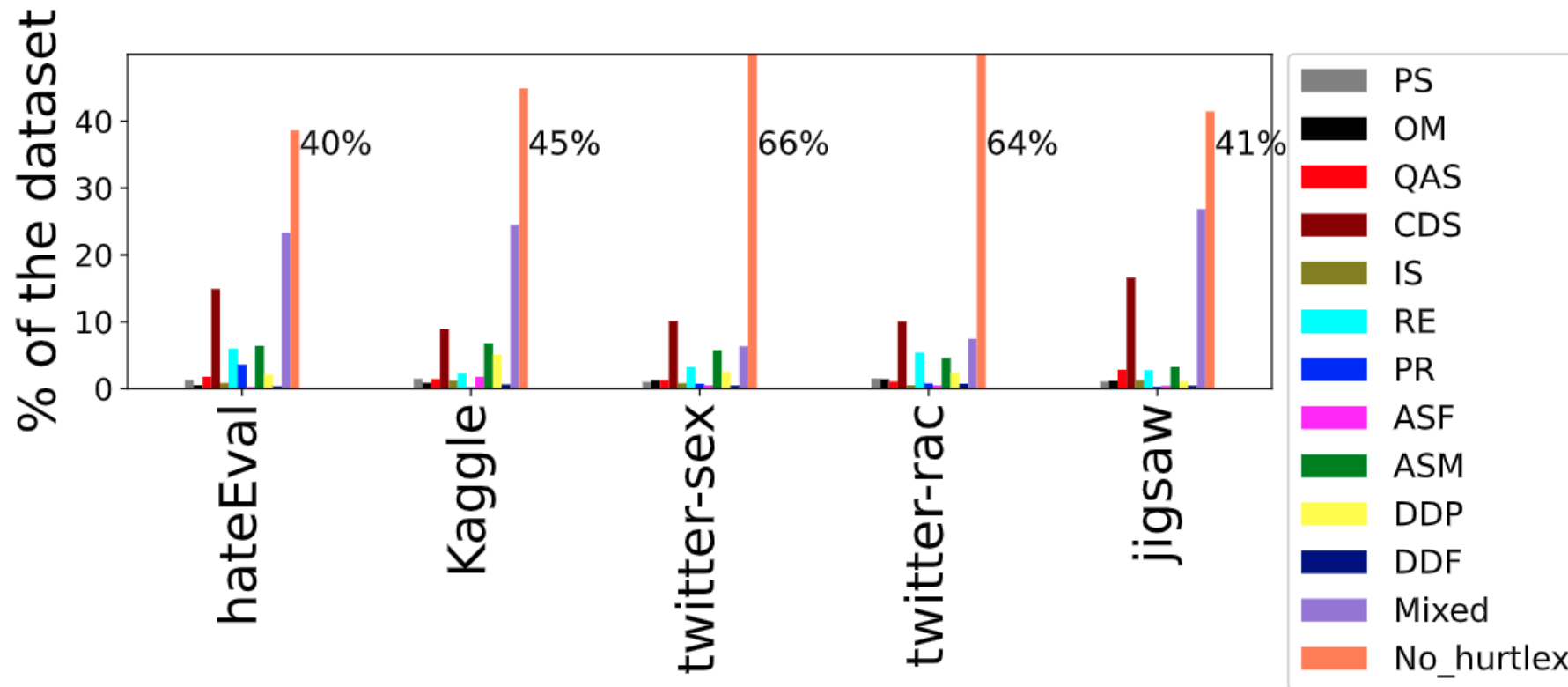
Detecting certain types of cyberbullying (Hurtlex)

F1 scores of the KNN model with the different word embeddings on Hurtlex test set



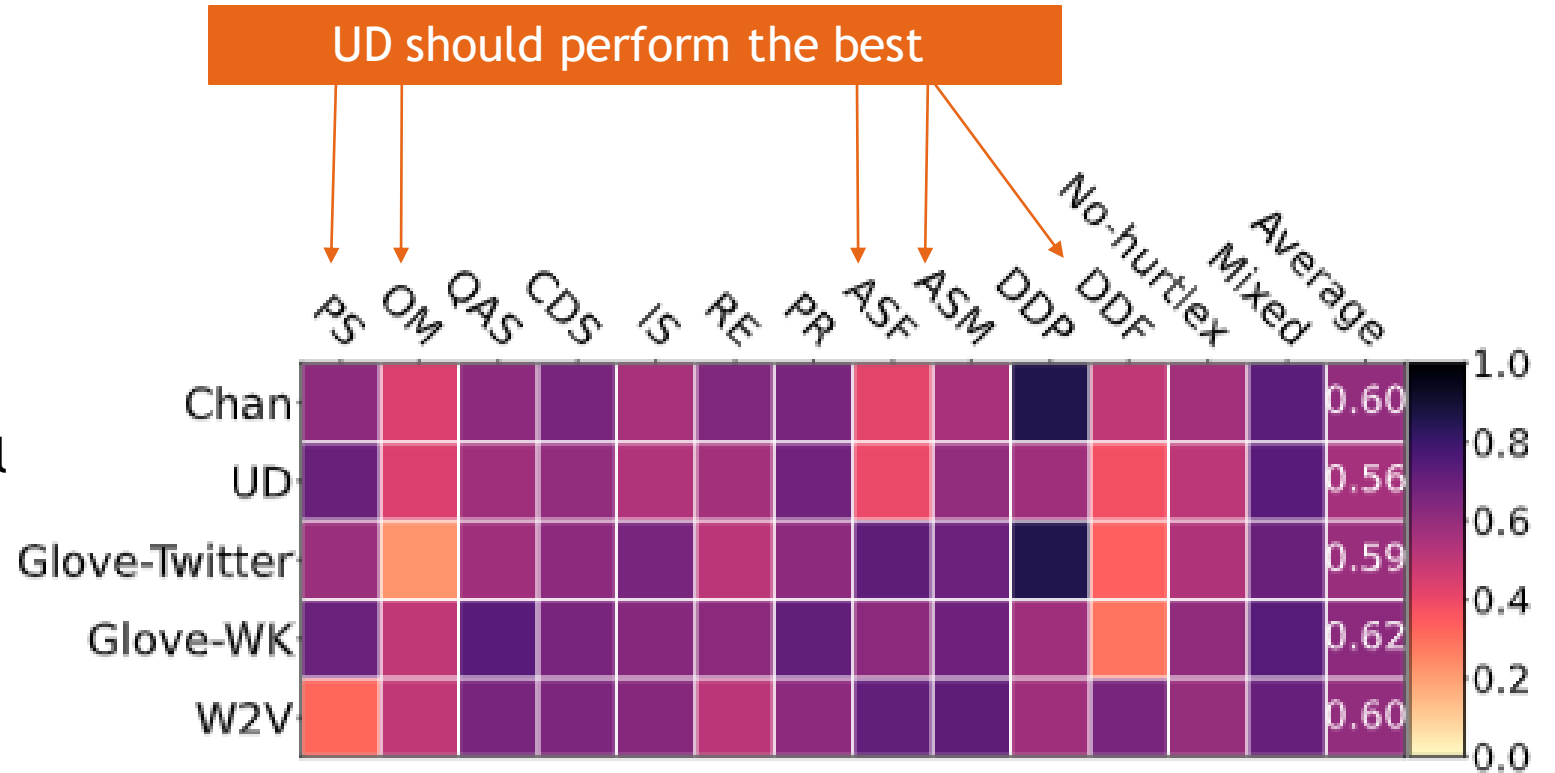
Detecting certain types of cyberbullying (Real-life datasets)

The percentage of the Hurltlex categories in each dataset



Detecting certain types of cyberbullying (Real-life datasets)

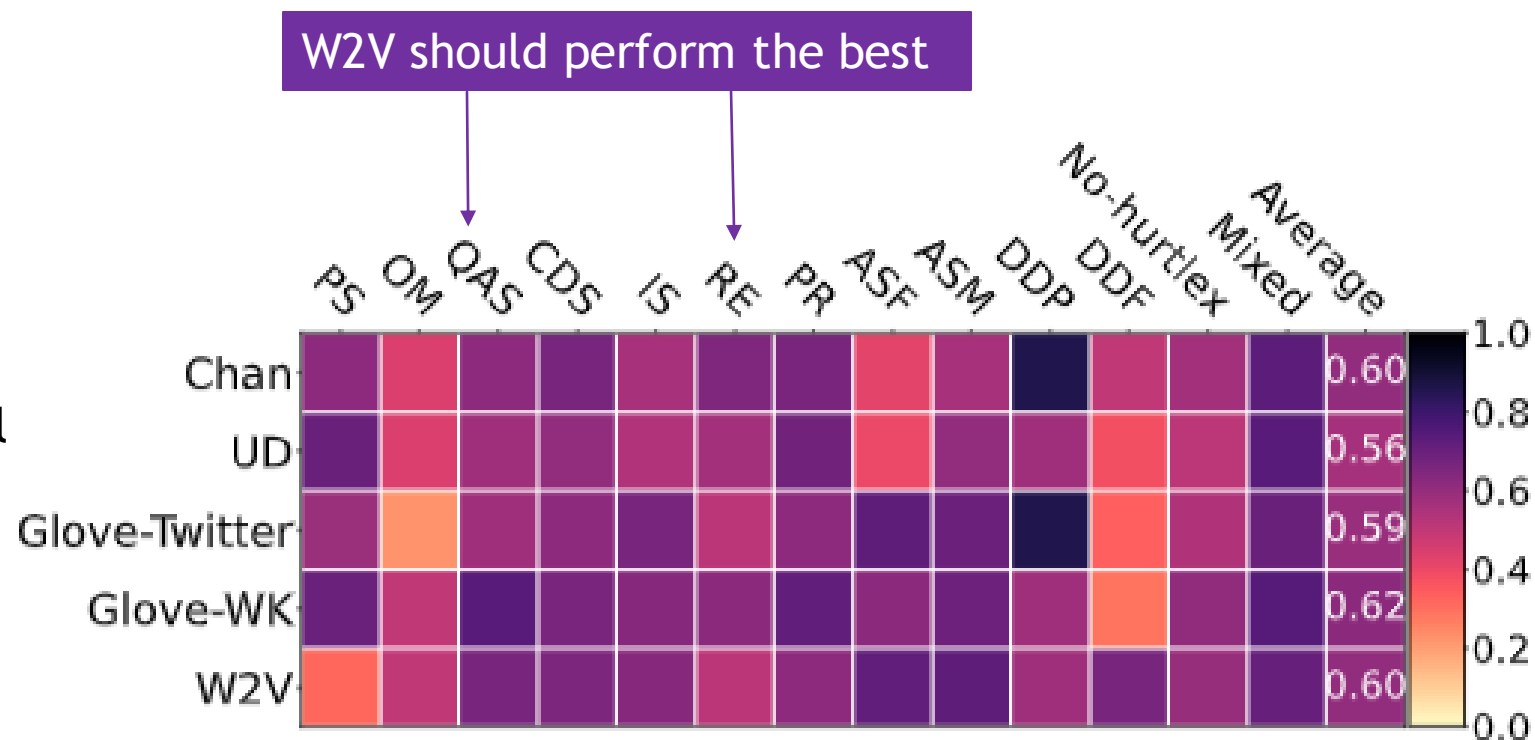
Binary F1-scores of the Bi-LSTM model with the different word embeddings on HateEval dataset.



(a) HateEval Dataset

Detecting certain types of cyberbullying (Real-life datasets)

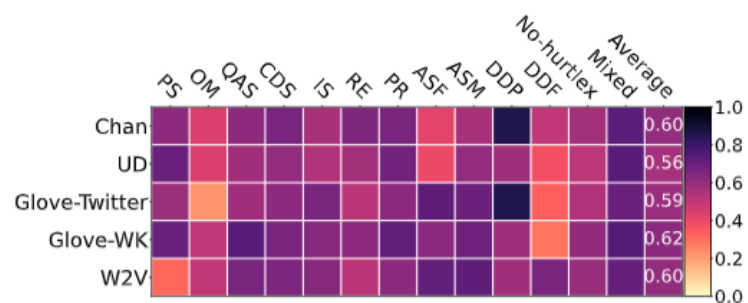
Binary F1-scores of the Bi-LSTM model with the different word embeddings on HateEval dataset.



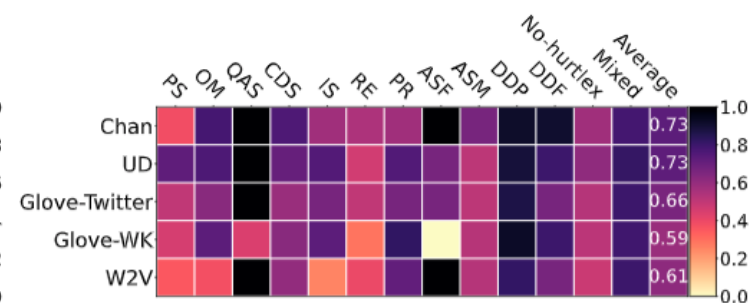
(a) HateEval Dataset

Detecting certain types of cyberbullying (Real-life datasets)

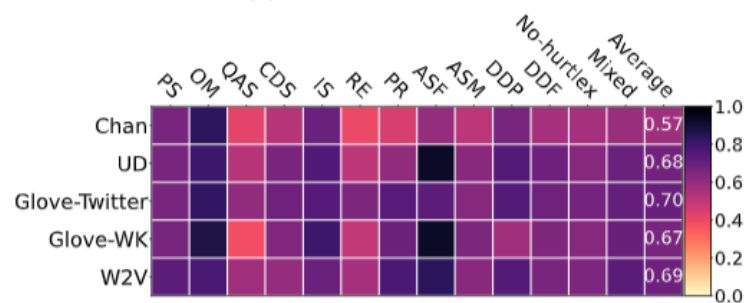
Binary F1-scores of the Bi-LSTM model with the different word embeddings on the different datasets



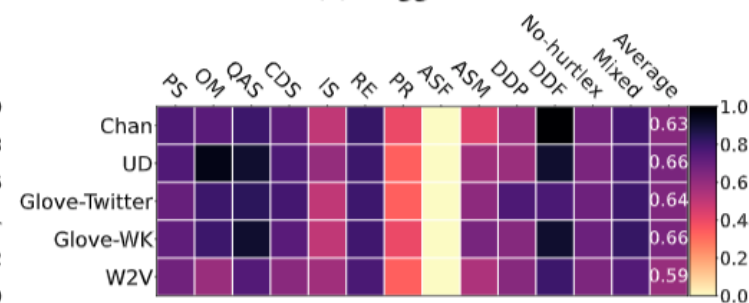
(a) HateEval Dataset



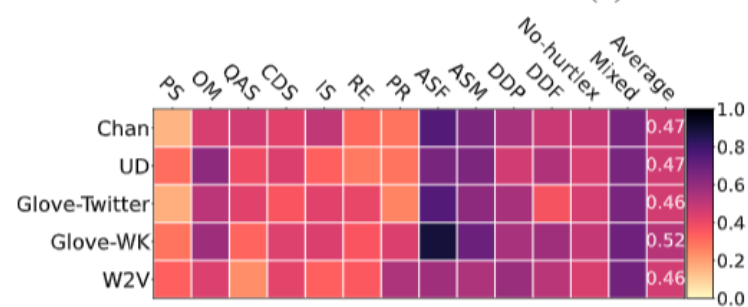
(b) Kaggle



(c) Twitter-sexism dataset



(d) Twitter-racism dataset



(e) Jigsaw-Toxicity dataset

Take away messages

- ▶ Slang-based word embeddings out-perform classic word embeddings on the task of cyberbullying detection.
- ▶ Some word embeddings are better at categorizing offensive words in the Hurltlex categories.
- ▶ However, these same embeddings do not necessarily perform the best on subsets of cyberbullying-related datasets that contain these types of words.

Thanks



fatma.elsafoury@uws.ac.uk

@fatmaelsafoury

Word embeddings

Word embedding models used in the paper

Word embeddings (WE)	Pre-training data	Type
Word2Vec	Google news	Classic
Glove-Wikipedia	Wikipedia articles	Classic
Glove-Twitter	Tweets	Slang-based
Chan	4&8Chan posts	Slang-based
Urban Dictionary (UD)	Head words & Definitions from Urban Dictionary website	Slang-based

Slang-based vs. Classic word embeddings

		Most similar 5 words to the word "queer"
Classic-word embeddings	Word2vec	genderqueer, LGBTQ, gay, LGBT, lesbian
	Glove-Wikipedia	transgender, lgbt, bisexual, lesbian, lgbtq

Slang-based vs. Classic word embeddings

Binary F1-scores of Bi-LSTM using the different word embeddings on different datasets

	Chan	UD	Glove-Twitter	Glove-Wikipedia	Word2Vec
HateEval (Hateful)	0.602	0.560	0.620	0.586	0.604
Kaggle (insults)	0.727	0.725	0.587	0.660	0.614
Twitter (racism)	0.631	0.663	0.659	0.644	0.591
Jigsaw (Toxicity)	0.474	0.467	0.519	0.458	0.461
Twitter (sexism)	0.574	0.678	0.667	0.699	0.688

For 4 datasets, **slange-based** embeddings is the **best performing**

1 dataset, **classic** embeddings is the **best performing**

Slang-based vs. Classic word embeddings

		Most similar 5 words to the word "queer"
Classic-word embeddings	Word2vec	genderqueer, LGBTQ, gay, LGBT, lesbian
	Glove-Wikipedia	transgender, lgbt, bisexual, lesbian, lgbtq
slang-based Word embeddings	Glove-Twitter	fag, faggot, feminist, gay, cunt
	UD	fag, homo, homosexual, bumblaster, buttyman
	Chan	faggot, metrosexual, fag, transvestite, homo

Abusive words to the gay community