Cross-lingual Hate Speech Detection in Social Media

Aiqi Jiang

Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
a.jiang@qmul.ac.uk

Disclaimer ⚠️

This presentation contains examples of hate speech;

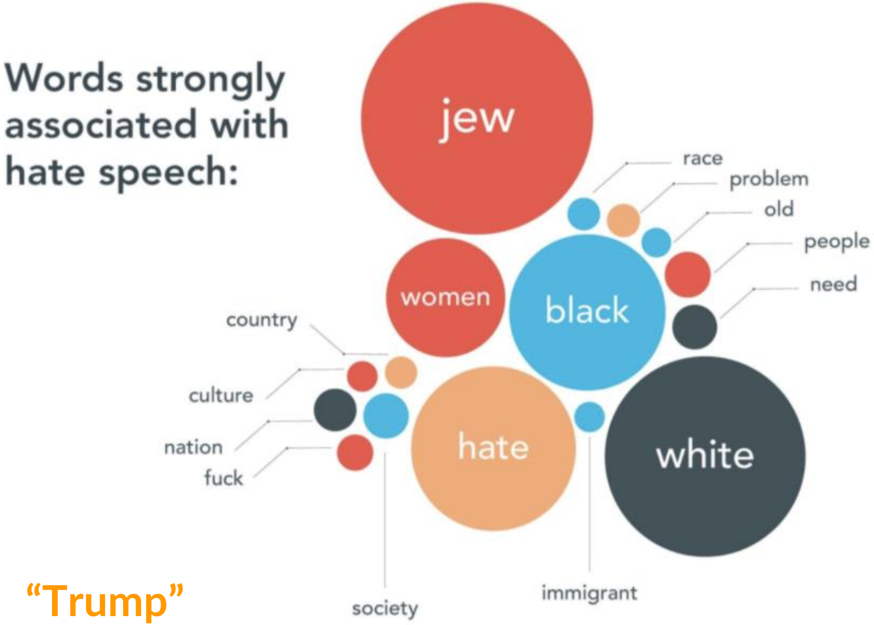
they do not represent the views of the author.

What & Why hate speech?

*"Speech that **attacks, insults or disparages** a person or group based on **specific characteristics**"*

e.g. gender, race, religion, sexual orientation, or nationality

What & Why hate speech?



“Trump”

- ❏ Prevalence of various social media platforms
- ❏ Anonymity and lack of moderation
- ❏ Increasing willingness to express

Why detect cross-lingual hate speech?

She is an ugly black hearted troll

You’re a sad little f*cking bitch

Trump kicks dem butt - its so fun

Why detect cross-lingual hate speech?

Trump patea el trasero
- es muy divertido

Ты грустная
маленькая гребаная
сука

She is an ugly
black hearted troll

-ترامب يركل بعقب
إنه ممتع للغاية

彼女は醜い黒い心
のトロールです

You're a sad little
f'cking bitch

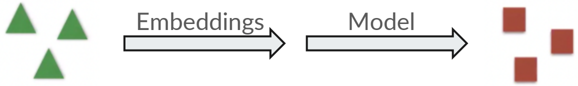
Trump kicks
dem butt - its so
fun

你是个可悲
的小婊子

È una brutta troll
dal cuore nero

Why detect cross-lingual hate speech?

*“Leverage NLP models to transfer the annotated hate speech data from one language (**source language** with more **resources**) to another language (**target language** with less **resources**)”* *(Pamungkas and Patti, 2019)*



Source language

Target language

Ты грустная маленькая гребаная сука

彼女は醜い黒い心のトロールです

You’re a sad little f*cking bitch

She is an ugly black hearted troll

Trump patea el trasero - es muy divertido

ترامب يركل بعقب- إنه ممتع للغاية

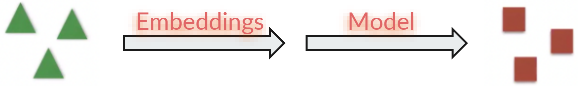
Trump kicks dem butt - its so fun

你是个可悲的小婊子

È una brutta troll dal cuore nero

Why detect cross-lingual hate speech?

*"Leverage NLP models to transfer the annotated hate speech data from one language (**source language with more resources**) to another language (**target language with less resources**)" (Pamungkas and Patti, 2019)*



Source language

Target language

Ты грустная маленькая гребаная сука

She is an ugly black hearted troll

Trump patea el trasero - es muy divertido

彼女は醜い黒い心のトロールです

ترامب يركل بعقب-
إنه ممتع للغاية

You're a sad little f*cking bitch

Trump kicks dem butt - its so fun

你是个可悲的小婊子

È una brutta troll dal cuore nero

Cross-lingual related work

Existing models

- Classic machine learning based -- LR, SVM *(Basile and Rubagotti, 2018; Pamungkas and Patti, 2021)*
- Neural network based -- LSTM, GRU *(Pamungkas and Patti, 2019; Corazza et al., 2020)*
- Transformer based -- Multilingual BERT, XLM, XLM-RoBERTa*(Dadu et al., 2020; Corazza et al., 2020; Ranasinghe and Zampieri, 2020)*

Cross-lingual related work

Existing models

- Classic machine learning based -- LR, SVM *(Basile and Rubagotti, 2018; Pamungkas and Patti, 2021)*
- Neural network based -- LSTM, GRU *(Pamungkas and Patti, 2019; Corazza et al., 2020)*
- Transformer based -- Multilingual BERT, XLM, XLM-RoBERTa*(Dadu et al., 2020; Corazza et al., 2020; Ranasinghe and Zampieri, 2020)*

Cross-lingual representation approaches

- Multilingual embedding model -- LASER, MUSE, Babylon *(Basile and Rubagotti, 2018; Ousidhoum et al., 2019; Pamungkas and Patti, 2021)*
- Multilingual pre-trained model -- Multilingual BERT, XLM, XLM-RoBERTa
- Monolingual embedding model or n-grams feature with machine translation *(Pamungkas and Patti, 2021)*

Challenges

- Limited resources and size of existing hate speech resources especially in non-English languages
  - English : Non-English ~ 14:1 *(Pamungkas et al., 2021)*
  - The size of most non-English datasets is less than 10k *(Pamungkas et al., 2021)*

Challenges

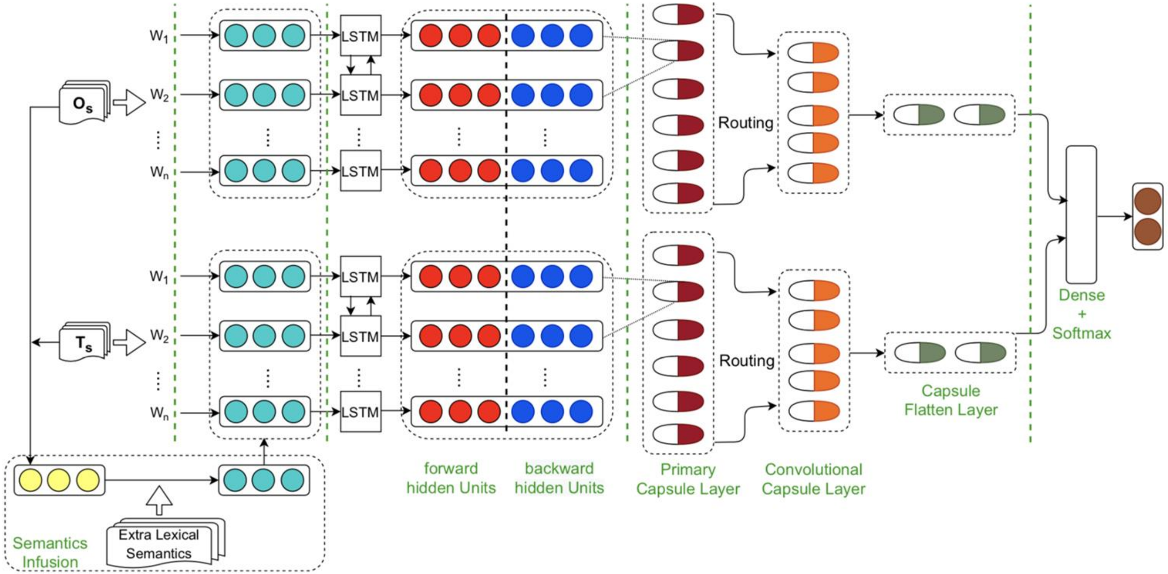
- Limited resources and size of existing hate speech resources especially in non-English languages
  - English : Non-English ~ 14:1 *(Pamungkas et al., 2021)*
  - The size of most non-English datasets is less than 10k *(Pamungkas et al., 2021)*
- Lack of integration with domain-specific knowledge
  - Limited hate-related resources -- e.g. Hate-specific lexicon (HurtLex), emotion and sentiment

Challenges

- ❑ Limited resources and size of existing hate speech resources especially in non-English languages
  - ❑ English : Non-English ~ 14:1 *(Pamungkas et al., 2021)*
  - ❑ The size of most non-English datasets is less than 10k *(Pamungkas et al., 2021)*
- ❑ Lack of integration with domain-specific knowledge
  - ❑ Limited hate-related resources -- e.g. Hate-specific lexicon (HurtLex), emotion and sentiment
- ❑ Lack of more comprehensive semantic features

CCNL-Ex: Cross-lingual Capsule Network Learning Model with Extra Lexical Semantics

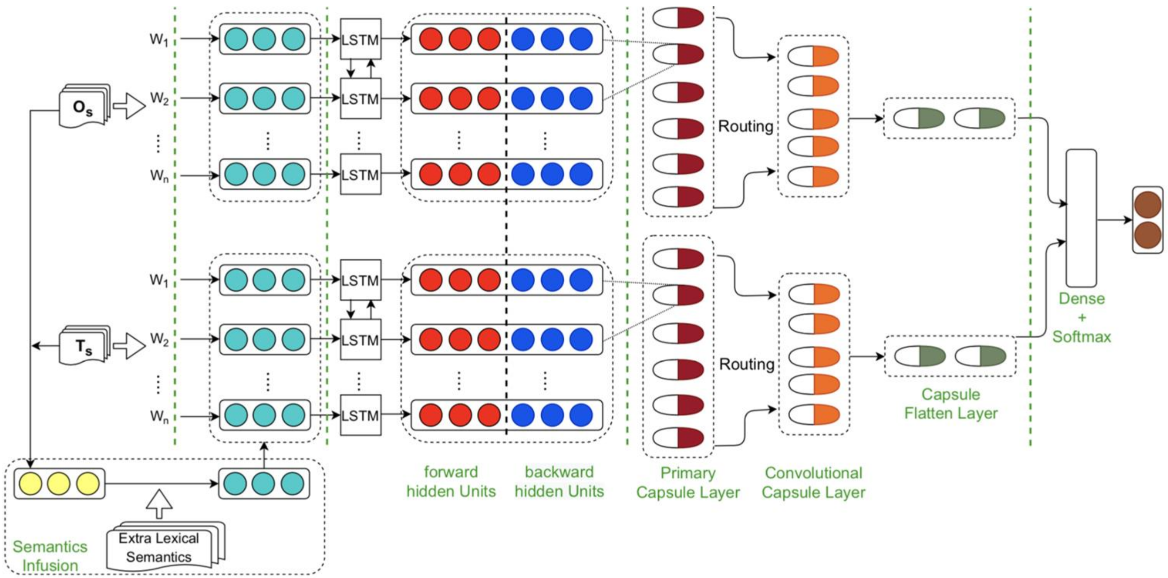
- ❑ A zero-shot joint framework with two parallel inputs
- ❑ Retrofitting embeddings by infusing domain-specific lexical semantics
- ❑ Weight-shared capsule network for spatial-level semantic features



I. Two parallel inputs

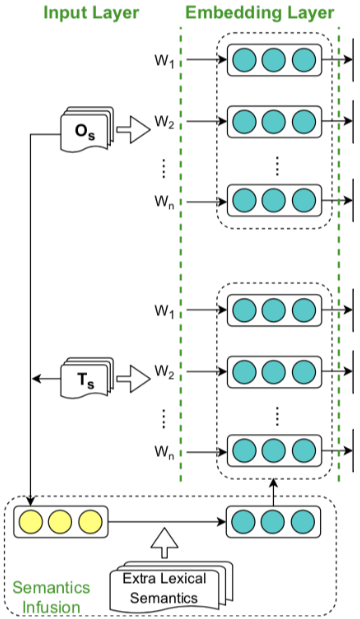
Bilingual data as two inputs fed to same capsule-based architectures

- Input 1: Source language Os
- Input 2: Target language Ts translated from Os
- Translation via Google Translate

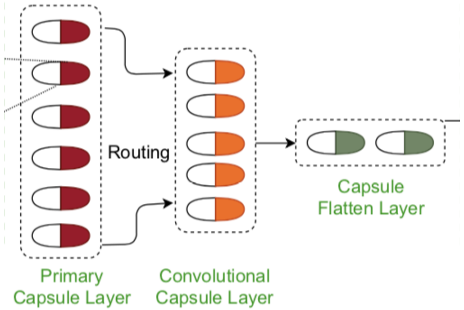


II. Domain semantic infusion in embeddings

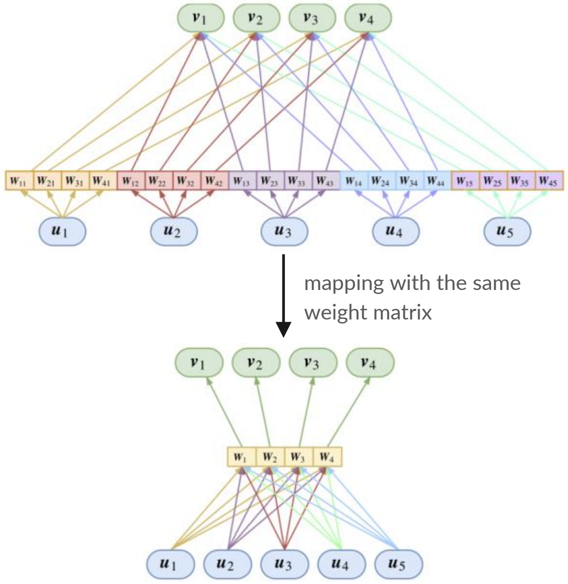
- Select multilingual domain-specific lexicon
  - HurtLex (Bassignana et al., 2018) -- hate-specific lexicon
  - Multilingual Sentiment Lexicon (Chen and Skiena, 2014)
- Select semantically related words for lexical words
  - SenticNet (Cambria et al., 2010) -- top five semantic words
- Retrofit pre-trained word embeddings by integrating lexicon-derived semantic information
  - minimise distances between lexical word and its semantic words (Faruqui et al., 2015)



Capsule Network



- Capsule network with shared weights obtain semantic compositionality
  - Vector-in/vector-out in each capsule
  - One capsule can be a group of neurons -- represent different characteristics of specific features
  - capture hierarchically spatial relationships between two-layer capsules
- Dynamic routing - determine the credit attribution between capsules in two layers



Dataset

- **Gender-based** hate speech datasets in English, Spanish, and Italian
- Binary labels -- misogynistic and non-misogynistic
- Evaluation -- macro-averaged F1 score

Language	English (EN)	Spanish (ES)	Italian (IT)
Train	3200	2646	3200
Validation	800	661	800
Test	1000	831	1000
MTR_{train} (%)	44.6	49.9	45.7
MTR_{test} (%)	46.0	49.9	50.9
Source	Evalita2018	IberEval2018	Evalita2018

MTR - misogynistic text rate

Comparison of CCNL and CCNL-Ex over baselines

Monolingual models

Multilingual representation models

Multilingual transformers

Joint learning model

(Pamungkas and Patti, 2021)

Model	ES→EN	EN→ES	IT→EN	EN→IT	ES→IT	IT→ES
Majority	0.351	0.334	0.351	0.329	0.329	0.334
SVM	0.620	0.561	0.588	0.227	0.643	0.525
CNN	0.598	0.613	0.592	0.275	0.636	0.607
BiLSTM	0.575	0.608	0.597	0.341	0.498	0.459
CapsNet	0.616	0.559	0.601	0.323	0.555	0.611
LASER	0.552	0.466	0.597	0.374	0.678	0.619
MUSE	0.592	0.491	0.618	0.400	0.717	0.666
mBERT	0.567	0.580	0.568	0.399	0.648	0.618
XLM-R	0.583	0.618	0.597	0.411	0.677	0.613
JL-HL	0.635	0.687	0.605	0.497	0.660	0.637
CCNL	0.624	0.719	0.628	**0.584**	**0.735**	**0.668**
CCNL-Ex	**0.651**	**0.729**	**0.629**	0.519	0.736	**0.670**

Best in bold & second underlined

Comparative experiments

Framework Ablation Analysis

- ❑ CCNL outperforms all ablated models, demonstrating the combined benefits of all CCNL components

Impact of Feature Extraction Layer

- ❑ Highlight the ability of the BiLSTM network to extract local contextual information

Model	**ES→EN**	**EN→ES**	**IT→EN**	**EN→IT**	**ES→IT**	**IT→ES**
Results for ablation experiments						
CCNL-non-parallel	0.522	0.558	0.570	0.513	0.626	0.624
CCNL-non-LSTM	0.373	0.609	0.565	0.406	0.685	0.623
CCNL-non-Caps	0.597	0.678	0.613	0.439	0.643	0.622
CCNL	**0.624**	**0.719**	**0.628**	**0.584**	**0.737**	**0.668**
Results for feature layers						
CCNL-non-FE	0.373	0.609	0.565	0.406	0.685	0.623
CCNL-CNN	0.521	0.592	0.577	0.439	0.633	0.622
CCNL-GRU	0.458	**0.722**	0.613	0.411	0.715	**0.671**
CCNL	**0.624**	0.719	**0.628**	**0.584**	**0.737**	0.668

Error Analysis

(a) Implicit hate

(b) Overuse of hateful words

(c) Lack of prior information

(d) Erroneous translation

Text	GT	P	ET
Analicemos esto: ¿Si te pones unos shorts así, en la calle, ¿qué esperas que te digan? ¿Acoso? ¿O Provocación... <u>Translation:</u> Let's analyse this: If you wear shorts like this, in the street, what do you expect them to say? Bullying? Or Provocation ...	1	0	a
tranquille ragazze, tranquilli gay, il Butturini c'ha una morosa che un pezzo di figa mostruosa! #TVOI <u>Translation:</u> quiet girls, quiet gays, Butturini has a girlfriend who is a piece of monstrous pussy! #TVOI	0	1	b
@user ben sasse is 100% correct. since 1973, all ive ever heard every two years for elections are hysterical women (all a leftist act) about back-alley abortions. this shit is getting old! i didn't hear one other protest issue being yelled about i	1	0	c
@user ma se la #culona #tedesca che predica #austerit mi sono perso qualcosa <u>Translation:</u> @user but if the #culona #german preaching #austerit I missed something	1	0	d

GT - Ground truth, **P** - Prediction, **ET** - Error Types, Labels are noted – **hateful (1)** and non-hateful (0)

Summary

- The **first approach** to cross-lingual hate speech detection that incorporates capsule networks
- **Integrate hate-related lexicons** into pre-trained word embeddings to investigate their potential to further boost performance
- CCNL-Ex model yields **SOTA performance** for all language pairs **compared with ten baselines**

Thanks for listening 🍀

Queen Mary
University of London

CIS centre for intelligent sensing