Generalisation in social media research

From fact verification to hate speech detection

Arkaitz Zubiaga

PAN @ CLEF 2021

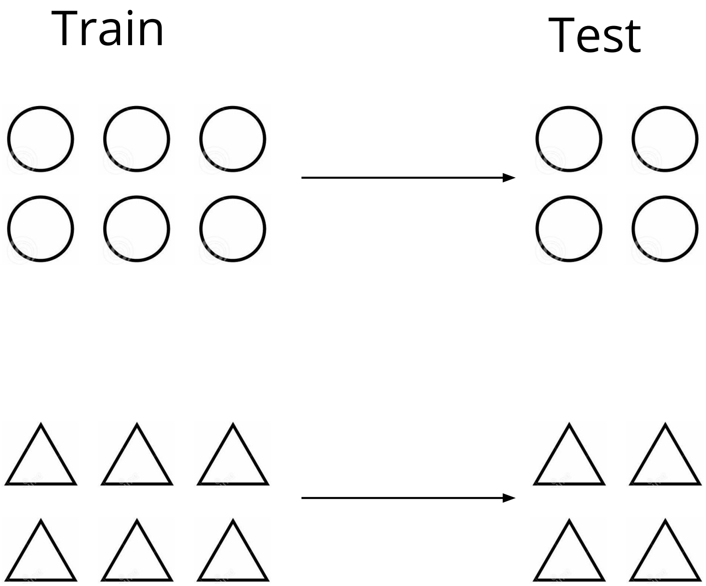
Who am I

- Lecturer, Queen Mary University of London.

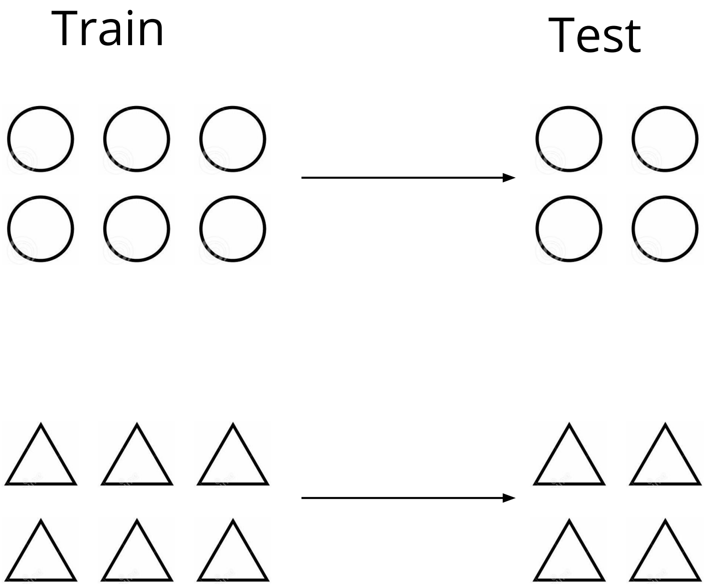
- Worked on social media & misinformation research for 10+ years.

- Currently focusing on a number of related areas:
    - Hate speech / abusive language detection.
    - Automated fact-checking.
    - Stance detection.

What do we mean by generalisation

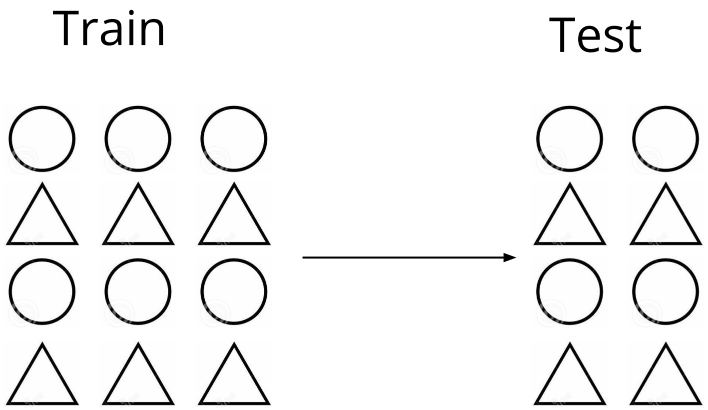


What do we mean by generalisation



Hardly generalisable

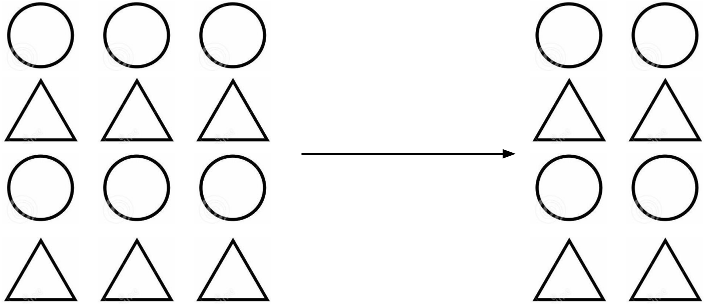
What do we mean by generalisation



What do we mean by generalisation

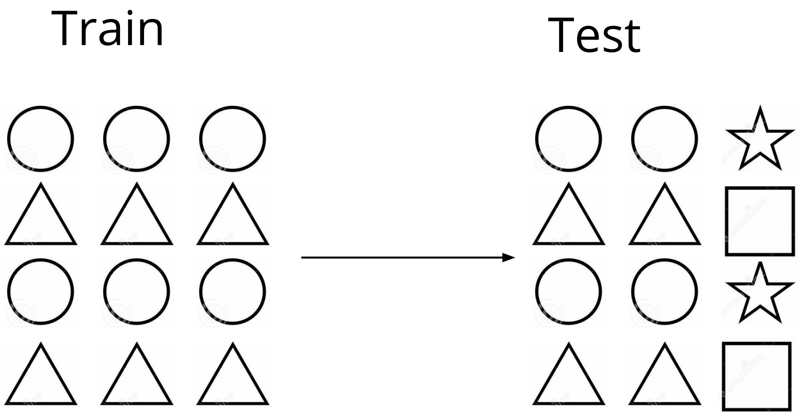
Train

Test



Some generalisability

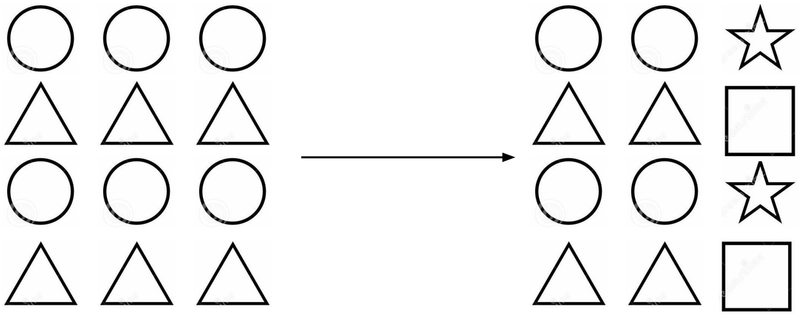
What do we mean by generalisation



What do we mean by generalisation

Train

Test



More generalisable

Why is generalisation important

Issues with artificial intelligence-based content moderation highlighted after world's most popular YouTube chess channel labelled 'harmful and dangerous'

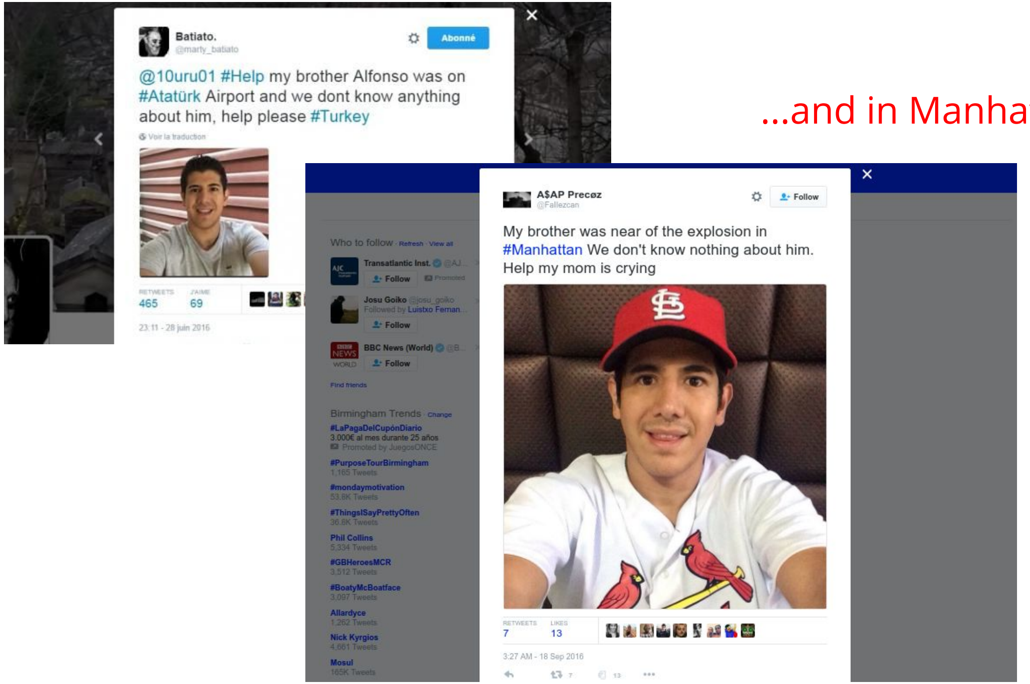
Anthony Cuthbertson | Thursday 18 February 2021 18:32 | comments

Why is generalisation important

Missing in Turkey

Why is generalisation important

...and in Manhattan



Why is generalisation important

...then in Egypt

Why is generalisation important

...also in Orlando

How is generalisation operationalised

Two typical ways of evaluating generalisation:

1. Train on dataset A, test on dataset B.
2. Strategically split dataset A into A1 and A2.

   Train on A1, test on A2.

In this talk

I’ll be covering two main aspects:

1. What we’ve recently done researching *towards* generalisation.

2. Challenges in hate speech detection to improve in terms of generalisation.

Challenge #1: data collection

Detection of social media hoaxes

- Task: veracity classification.

- Problem: datasets are often biased, e.g. collected from fact-checking websites.

- Objective: come up with a scalable methodology for broad data collection.

Celebrity death hoaxes

Collection of social media hoaxes

- Collection of death reports (RIP + person name), e.g.:

”**RIP Elizabeth II**, she was so inspiring.”

”**RIP Elizabeth II** oh dear :(“

”Sad to hear about the passing of **RIP Elizabeth II**”

”Those posting **RIP Elizabeth II**, stop it!”

Collection of social media hoaxes

- Collection of death reports (RIP + person name), e.g.:

”**RIP Elizabeth II,**

”**RIP Elizabeth II**

”Sad to hear abo

”Those posting **R**

FAKE!

Wikidata entry

{"id":"8023",

"name":"Nelson Mandela",

"birth":{"date":"1918-07-18","precision":11},

"death":{"date":"2013-12-05","precision":11},

"description":"former President of South Africa, anti-apartheid activist",

"aliases":["Nelson Rolihlahla Mandela","Mandela","Madiba"]}

Collection of social media hoaxes

1) Collection of tweets with keyword ‘RIP’ in it for 3 years (Jan 2012 – Dec 2014).

2) Sample tweets matching the ‘RIP person-name’ pattern.

3) Sampling, i.e. names with 50+ occurrences on a given day.

4) Semi-automated labelling.

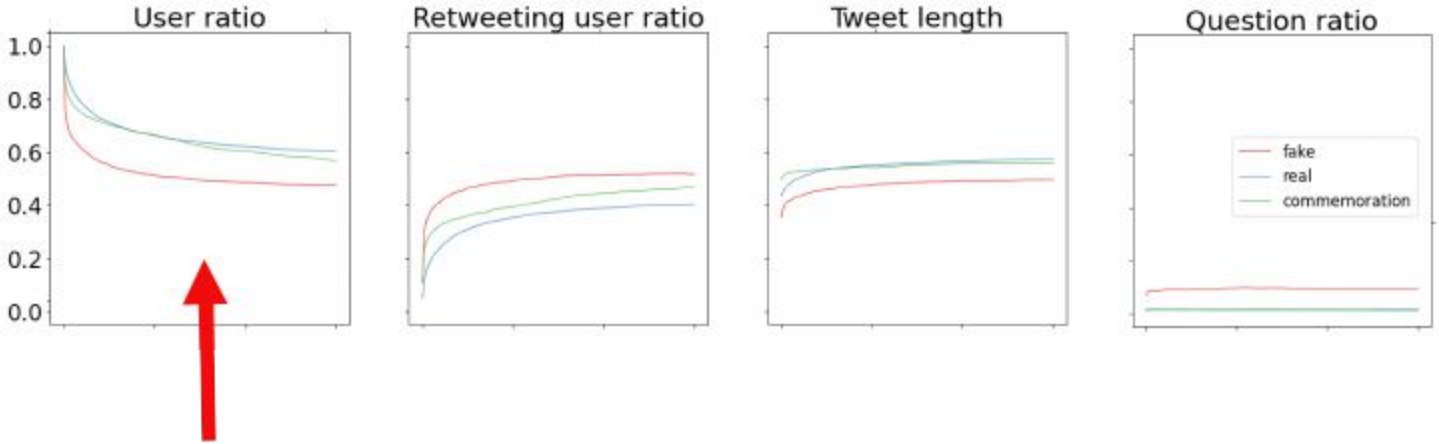
5) 4,007 death reports (13+ million tweets):
    a) 2,301 real deaths.
    b) 1,092 commemorations.
    c) 614 death hoaxes.

Results

	0	1'	2'	5'	10'	15'	30'	60'	120'	300'
social	.427	.495	.509	.510	.510	.528	.535	.577	.594	.591
w2v	.641	*.655*	.658	.663	.667	.670	.680	.696	.699	.698
social+w2v	.612	.634	.661	.671	.671	.677	.675	.709	.709	.724
gw2v	.556	.565	.574	.608	.612	.618	.623	.645	.648	.664
social+gw2v	.569	.590	.599	.616	.633	.647	.663	.679	.688	.686
infersent	.637	.640	.653	.664	.683	.681	.697	.722	.734	.759
social+infersent	*.643*	*.655*	*.670*	*.678*	*.691*	*.688*	*.698*	*.731*	*.748*	**.767**
multiw2v*	**.669**	.676	.691	.703	.714	.722	.723	.721	.738	.741
social+multiw2v*	.647	**.677**‡	**.696**‡	**.707**‡	**.716**‡	**.725**‡	**.724**†	**.744**†	**.752**	*.748*

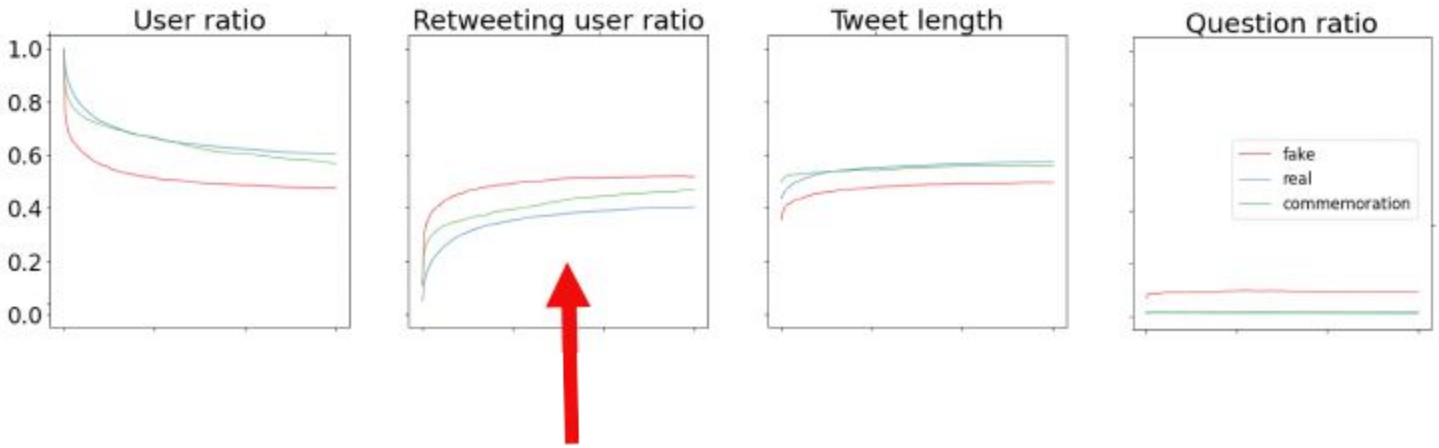
Proposed methods indicated with a star (*). Best method highlighted in bold and second-best method for different types of features highlighted in italic. ‡: statistically significant at $p < .01$, †: statistically significant at $p < .05$.

Analysis of features



Hoaxes tend to have fewer distinct users posting them.

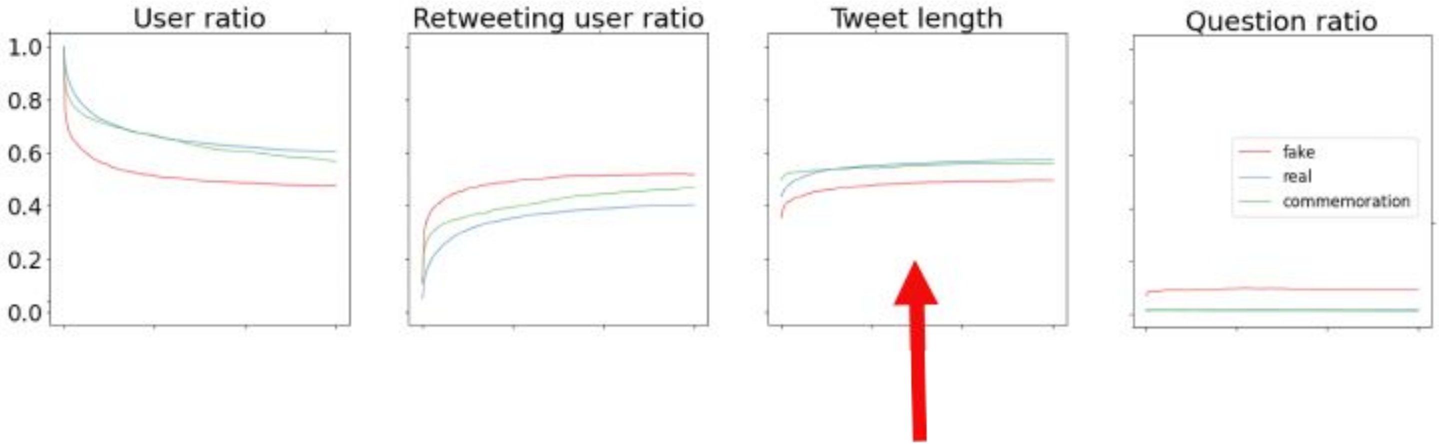
Analysis of features



Hoaxes tend to have fewer distinct users posting them.

BUT they are retweeted by more distinct users!

Analysis of features

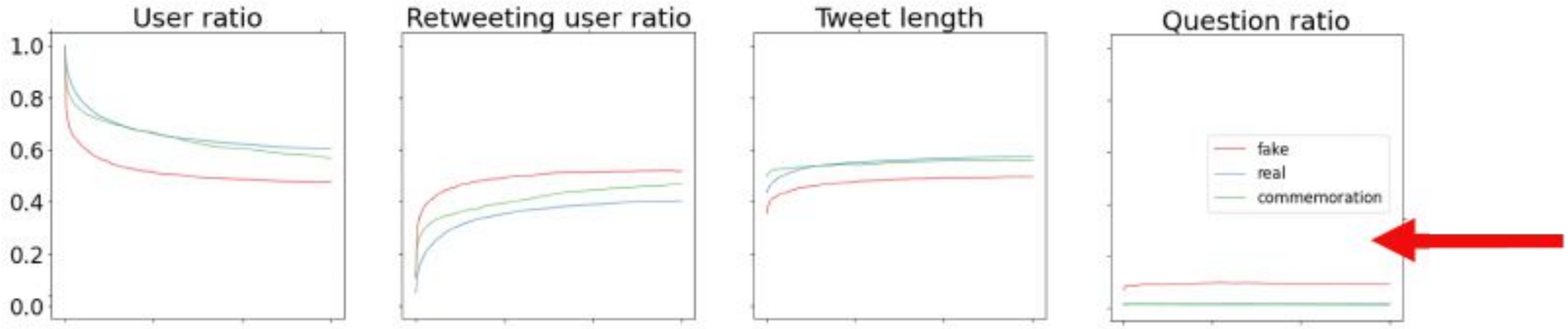


Hoaxes tend to be shorter in length, not as carefully crafted as true stories?

They tend to lack links and pictures.

Presumably less evidence linked to them?

Analysis of features



And hoaxes tend to spark more questions!

Limitations

Hoaxers like having fun:

- “RIP Justin Bieber”

- “RIP Messi”

Limitations

Hoaxers like having fun:

- “RIP Justin Bieber”
  - I mean… he’s a Really Inspiring Person (RIP)

- “RIP Messi”
  - He’s dead after missing that crucial penalty...

Data available

Twitter Death Hoaxes dataset

Version 3 Dataset posted on 25.03.2019, 18:55 by Arkaitz Zubiaga

This is a dataset of death reports collected from Twitter between 1st January, 2012 and 31st December, 2014. It was collected by tracking the keyword 'RIP', and matching those tweets in which a name is mentioned next to RIP. Matching names were identified by using Wikidata as a database of names. For more details, please refer to the paper: https://arxiv.org/abs/1801.07311

RESEARCH-ARTICLE

Early Detection of Social Media Hoaxes at Scale

Authors: Authors Info & Affiliations

Publication: ACM Transactions on the Web • August 2020

Abstract

The unmoderated nature of social media enables the diffusion of hoaxes, which in turn jeopardises the credibility of information gathered from social media platforms. Existing research on automated detection of hoaxes has the limitation of using relatively small datasets, owing to the difficulty of getting labelled data. This, in turn, has limited research

https://figshare.com/articles/Twitter_Death_Hoaxes_dataset/5688811

Zubiaga, A., & Jiang, A. (2020). Early detection of social media hoaxes at scale. ACM Transactions on the Web (TWEB), 14(4), 1-23.

Challenge #2: generalisation across languages

Hate speech detection

- Hate speech refers to the use of language to attack, insult or disparage a person or group based on identity -- such as gender, race, religion, or sexual orientation.

Cross-lingual hate speech detection

- Hate speech detection predominantly done in English.
- Scarcity of hate speech datasets in less-resourced languages.
- If we really want to get rid of hate speech, we need to do it for any language.

Datasets

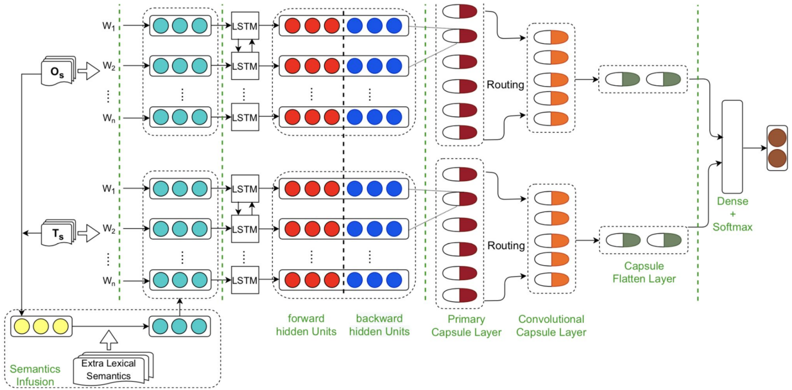
- Gender-based hate speech datasets in English, Spanish, and Italian

- Binary labels -- misogynistic and non-misogynistic

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

CCNL model

- Cross-lingual capsule network, leveraging:
  - Parallel corpora generated through machine translation.
  - Hate speech lexicons for the 3 languages.



Jiang, A., & Zubiaga, A. (2021, August). Cross-lingual Capsule Network for Hate Speech Detection in Social Media. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (pp. 217-223).

Results

+ Use of machine translation (CCNL) can substantially help generalisation.
    - But we depend on machine translation tools.

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	<u>0.666</u>
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	<u>0.635</u>	0.687	0.605	0.497	0.660	0.637
CCNL	0.624	<u>0.719</u>	<u>0.628</u>	**0.584**	**0.735**	**0.668**
CCNL-Ex	**0.651**	**0.729**	**0.629**	<u>0.519</u>	<u>0.736</u>	**0.670**

Results

- + Use of machine translation (CCNL) can substantially help generalisation.
  - But we depend on machine translation tools.
- + Use of lexicons (CCNL-Ex) can boost performance.
  - With exceptions, possibly IT lexicon not as good?

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**

Error analysis

(a) Implicit hate

Analicemos esto: ¿Si te pones unos shorts así, en la calle, ¿qué esperas que te digan? ¿Acoso? ¿O Provocación...
Translation: Let's analyse this: If you wear shorts like this, in the street, what do you expect them to say? Bullying? Or Provocation ...

False negative

Error analysis

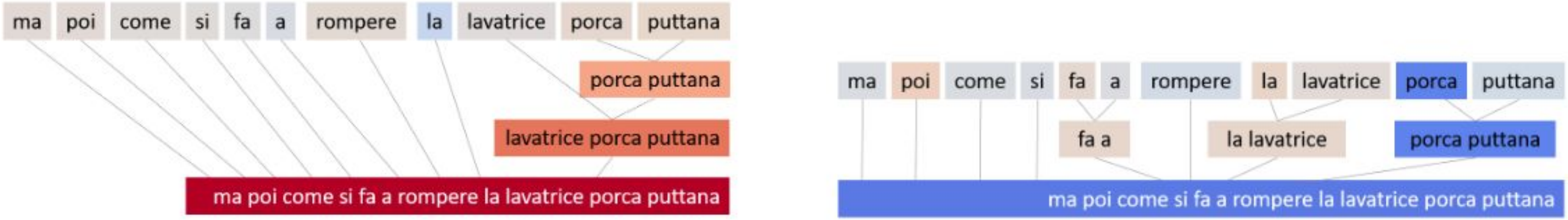
(b) Wrong translation

@user ma se la #culona #tedesca che predica #austerit mi sono perso qualcosa

Translation: @user but if the #culona #german preaching #austerit I missed something

False negative

Generalisation across languages



(a) Misclassified prediction by zero-shot, cross-lingual model trained on English and Spanish and tested on Italian data.

(b) Correct prediction by monolingual model trained on Italian and tested on Italian data.

Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In Proceedings of ACL.

Challenge #3: generalisation across platforms

Generalisation across platforms

Social media platforms have different characteristics:

- Different length restrictions.

Generalisation across platforms

Social media platforms have different characteristics:

- Different length restrictions.

- Different conventions for hashtags, mentions, etc.

Generalisation across platforms

Social media platforms have different characteristics:

- Different length restrictions.
- Different conventions for hashtags, mentions, etc.
- Different types of users who use different language (e.g. more / less formal).

Generalisation across platforms

- Task: can we predict the age of social media users
    - across platforms, i.e. testing on a different platform.

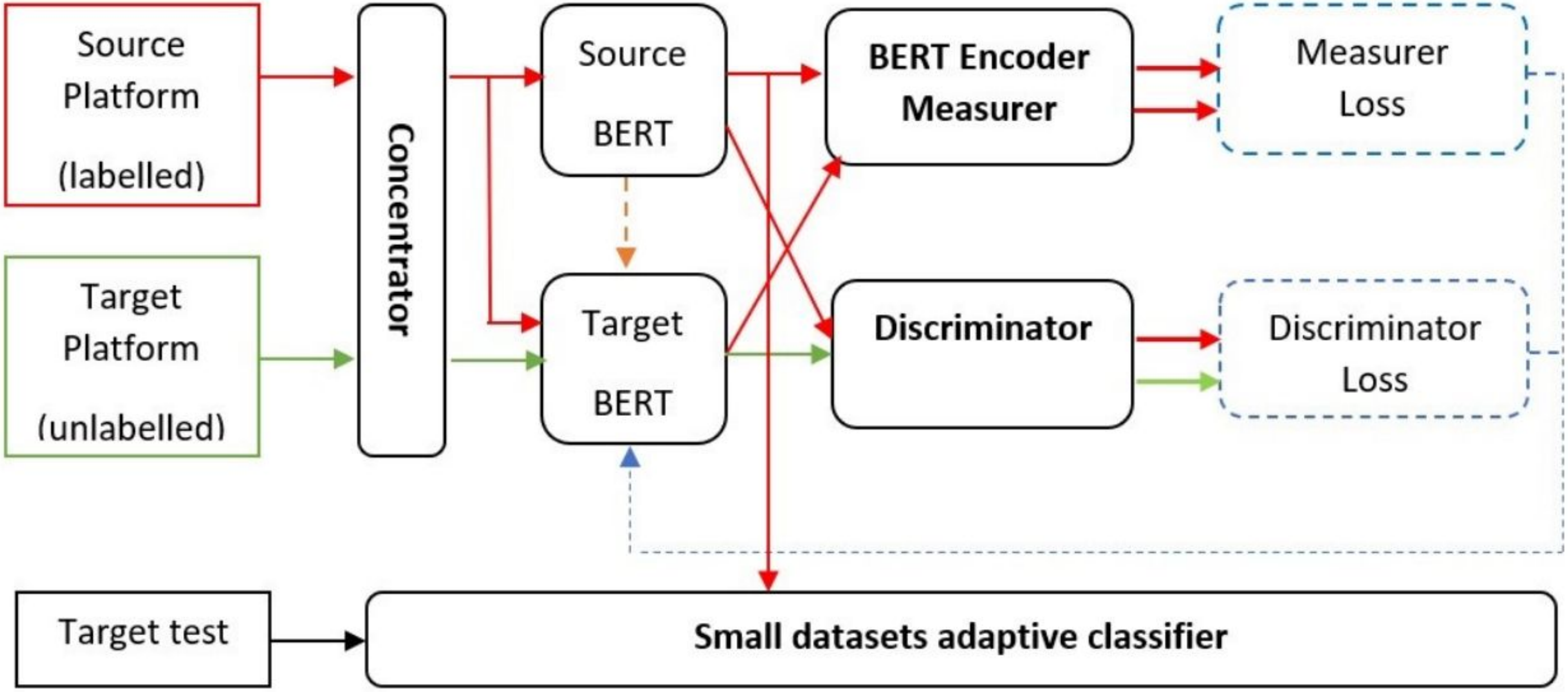
- More specifically: can we determine if a user is an adult or not?
    - Intended for protecting teenagers, e.g. against cyberbullying.

Generalisation across platforms

Platforms	YouTube	Myspace	Blogger	PAN13 (Netlog, Blogspot, Internetwordstats)
Size	3,468	14,813	19,320	236,600
Avg. length	115	17	3766	505
TR	0.2	0.096	0.42	0.08
Year	2020	2011	2009	2013
Source	Elsafoury [6]	Bayzick and Kontostathis [2]	Schler et al. [22]	Rangel et al. [19]

Table 1: Dataset statistics. TR: teenager ratio, as the portion of users in the dataset that are labelled as teenagers.

Generalisation across platforms



Generalisation across platforms

	Baseline	Full model
Source->Target	BERT	AB_CSA
B->Y	0.45	**0.54**
B->M	0.55	**0.58**
B->P	0.48	**0.52**
Y->B	0.37	**0.61**
Y->M	0.49	**0.53**
Y->P	0.47	**0.51**
M->B	0.50	0.45
M->Y	**0.54**	0.53
M->P	0.50	0.50
Average	0.48	**0.53**

Table 2: Cross-platform results.

Yi, P., & Zubiaga, A. (2021, August). Weakly Supervised Cross-platform Teenager Detection with Adversarial BERT. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (pp. 265-270).

Generalisation across platforms

Source->Target	BASE_LINE	AB_CS
B->B	0.86	**0.87**
Y->Y	0.59	0.54
M->M	**0.48**	0.43
P->P	0.49	**0.59**

Table 3: In-platform results

	Baseline	Full model
Source->Target	BERT	AB_CSA
B->Y	0.45	**0.54**
B->M	0.55	**0.58**
B->P	0.48	**0.52**
Y->B	0.37	**0.61**
Y->M	0.49	**0.53**
Y->P	0.47	**0.51**
M->B	0.50	0.45
M->Y	**0.54**	0.53
M->P	0.50	0.50
Average	0.48	**0.53**

Table 2: Cross-platform results.

Yi, P., & Zubiaga, A. (2021, August). Weakly Supervised Cross-platform Teenager Detection with Adversarial BERT. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (pp. 265-270).

Generalisation across platforms

https://sites.google.com/view/icwsm2021datachallenge

Challenge #4: generalisation over time

Generalisation over time

Social media data changes over time:

- New words emerge (e.g. COVID19) and words change meaning.

- Social media conventions change (e.g. 140 → 280 char).

- People's views change over time.

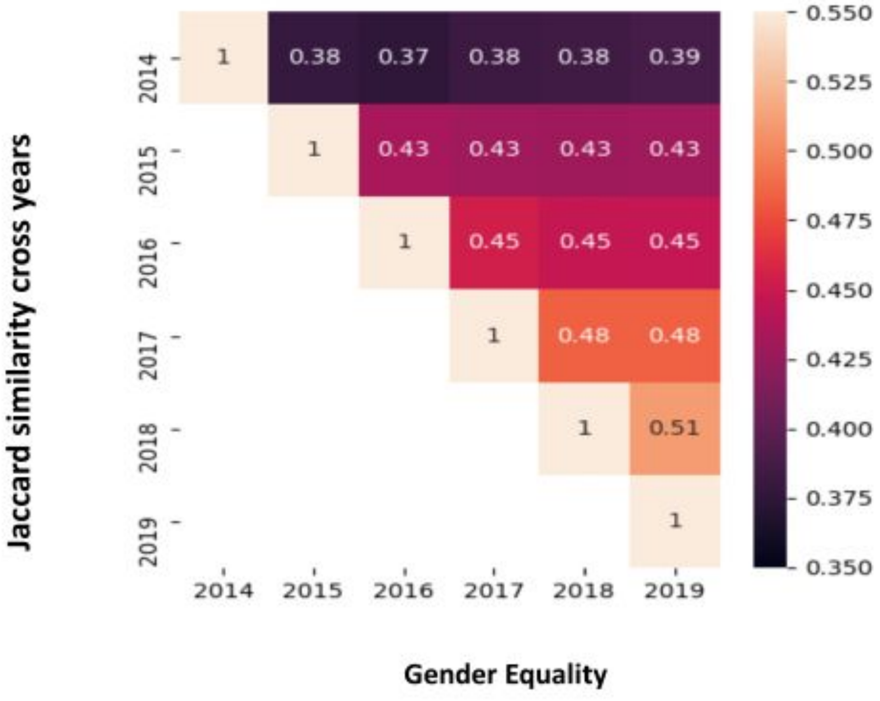
How does this affect our models trained on old data?

Generalisation over time

Collection of longitudinal dataset for stance detection:

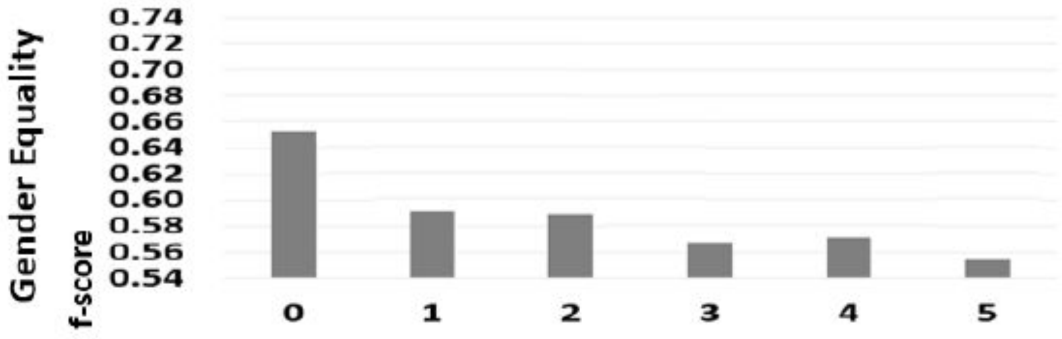
- Distantly supervised collection of tweets:
    - Supporting or opposing gender equality.
- Tweets covering years between 2014 and 2019.
    - ~50K tweets per year.
    - Same distribution for all years: 76.9% support, 23.1% oppose.

Generalisation over time



Generalisation over time

Stance classification with a standard model.



0	1	2	3	4	5
0.653	0.591 (-9.5%)	0.589 (-9.8%)	0.567 (-13.2%)	0.571 (-12.6%)	0.554 (-15.2%)

Generalisation over time

We propose aligning vocabularies to match varying vocabulary, word meanings, etc. making use of Compass.

We tested two alignment settings:

- All years: e.g. align 2014 with 2015, then with 2016, then with 2017...
- Source and target only: align 2014 with 2019, ignore years in between.

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019, July). Training temporal word embeddings with a compass. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 6326-6334).

Generalisation over time

Time gap	0	1	2	3	4	5
No align.	0.653	0.591 (-9.5%)	0.589 (-9.8%)	0.567 (-13.2%)	0.571 (-12.6%)	0.554 (-15.2%)
Align all years	**0.704**	**0.649** (-7.8%)	0.631 (-10.4%)	0.613 (-12.9%)	**0.620** (-11.9%)	0.617 (-12.4%)
Align src+tgt	0.653	0.639 (-2.1%)	**0.633** (-3.1%)	0.624 (-4.4%)	0.615 (-5.8%)	0.618 (-5.4%)

Generalisation over time

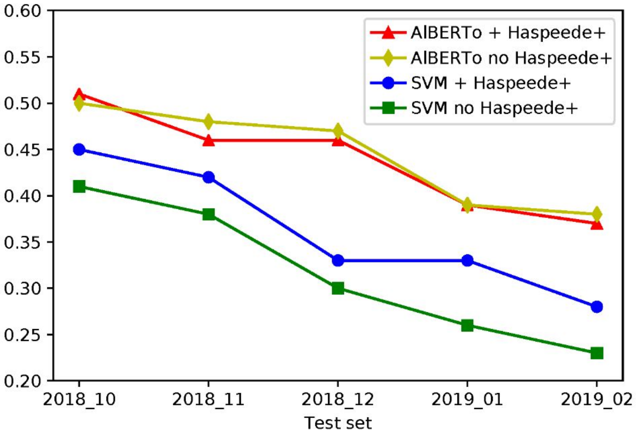
Time gap	0	1	2	3	4	5
No align.	0.653	0.591 (-9.5%)	0.589 (-9.8%)	0.567 (-13.2%)	0.571 (-12.6%)	0.554 (-15.2%)
Align all years	**0.704**	**0.649** (-7.8%)	0.631 (-10.4%)	0.613 (-12.9%)	**0.620** (-11.9%)	0.617 (-12.4%)
Align src+tgt	0.653	0.639 (-2.1%)	**0.633** (-3.1%)	0.624 (-4.4%)	0.615 (-5.8%)	0.618 (-5.4%)

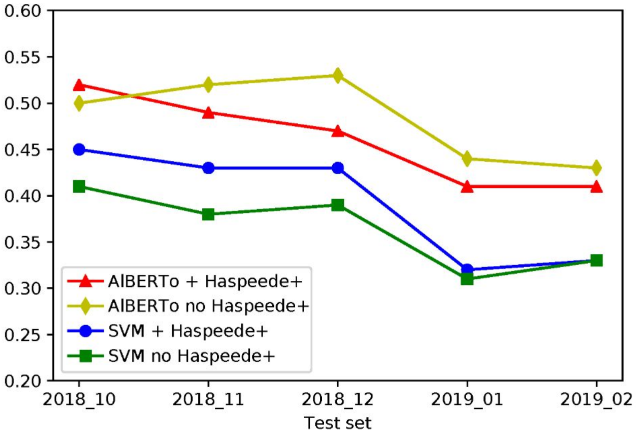
- Aligning definitely helps!
- Aligning source + target only best for reduced performance drop.

Alkhalifa, R., Kochkina, E., & Zubiaga, A. (2021). Opinions are made to be changed: Temporally adaptive stance classification. Proceedings of ACM Hypertext (OASIS).

Generalisation over time

Similar problem with hate speech detection.





Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. Applied Sciences, 10(12), 4180.

Challenges in generalisable hate speech detection

Challenges in generalisable hate speech detection

PeerJ Computer Science

Towards generalisable hate speech detection: a review on obstacles and solutions

Wenjie Yin and Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

ABSTRACT

Hate speech is one type of harmful online content which directly attacks or promotes hate towards a group or an individual member based on their actual or perceived aspects of identity, such as ethnicity, religion, and sexual orientation. With online hate speech on the rise, its automatic detection as a natural language processing task is gaining increasing interest. However, it is only recently that it has been shown that existing models generalise poorly to unseen data. This survey paper attempts to summarise how generalisable existing hate speech detection models are and the reasons why hate speech models struggle to generalise, sums up existing attempts at addressing the main obstacles, and then proposes directions of future research to improve generalisation in hate speech detection.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science, 7, e598.

Model performance often overestimated

Model, training dataset	Dataset			
	W	T1*	T2	T3
LR char, W	(0.86)	0.37	0.50	0.24
MLP char, W	(0.86)	0.38	0.50	0.25
CNN+GRU, T1*	0.11	(0.70)	0.48	0.51
CNN+GRU, T2	0.14	0.28	(0.83)	0.44
CNN+GRU, T3	0.13	0.48	0.50	(0.81)
LSTM, T2	0.23	0.33	(0.78)	0.47

Testing on same dataset

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech detection. In Proceedings of the 11th ACM workshop on artificial intelligence and security (pp. 2-12).

Model performance often overestimated

Model, training dataset	Dataset			
	W	T1*	T2	T3
LR char, W	(0.86)	0.37	0.50	0.24
MLP char, W	(0.86)	0.38	0.50	0.25
CNN+GRU, T1*	0.11	(0.70)	0.48	0.51
CNN+GRU, T2	0.14	0.28	(0.83)	0.44
CNN+GRU, T3	0.13	0.48	0.50	(0.81)
LSTM, T2	0.23	0.33	(0.78)	0.47

Testing on same dataset

Cross-platform

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech detection. In Proceedings of the 11th ACM workshop on artificial intelligence and security (pp. 2-12).

Challenges in generalisable hate speech detection

We can argue there are substantial **differences across datasets**:

Challenges in generalisable hate speech detection

We can argue there are substantial **differences across datasets**:

- Definitions & labelling criteria.

Challenges in generalisable hate speech detection

We can argue there are substantial **differences across datasets**:

- Definitions & labelling criteria.
- Proportion of abuse vs non-abuse.

Challenges in generalisable hate speech detection

We can argue there are substantial **differences across datasets**:

- Definitions & labelling criteria.
- Proportion of abuse vs non-abuse.
- Sampling criteria.

Challenges in generalisable hate speech detection

We can argue there are substantial **differences across datasets**:

- Definitions & labelling criteria.
- Proportion of abuse vs non-abuse.
- Sampling criteria.

But the problem is also in the **models overfitting a dataset**.

- Where possible, do try and test models on and across different datasets.

Challenges in generalisable hate speech detection

We discuss 3 key obstacles to generalisation:

- Non-standard grammar & vocabulary.

- Limited & biased data.

- Implicit expressions of hate.

Non-standard grammar & vocabulary

- Can lead to many false negatives, e.g.:
  - Words with hidden meanings: Skype, Google, banana.
  - Intended misspellings: feck.

Non-standard grammar & vocabulary

- Can lead to many false negatives, e.g.:
  - Words with hidden meanings: Skype, Google, banana.
  - Intended misspellings: feck.

- Beyond word embeddings:
  - Subword & char embeddings useful, e.g. Indurthi et al. (2019) won Hateval 2019.
  - Sentence embeddings.

Non-standard grammar & vocabulary

- Can lead to many false negatives, e.g.:
  - Words with hidden meanings: Skype, Google, banana.
  - Intended misspellings: feck.

- Beyond word embeddings:
  - Subword & char embeddings useful, e.g. Indurthi et al. (2019) won Hateval 2019.
  - Sentence embeddings.

- Possibly spelling correction: Gong et al. (2019)

Category	Metric	Original	Our system
racist	Precision	0.630	**0.640**
	Recall	0.617	0.681
	F1 score	0.623	**0.660**
sexist	Precision	0.641	0.630
	Recall	0.775	0.767
	F1 score	0.701	0.692

Limited & biased data

- Small data size
    - Abuse-specific embeddings: Caselli et al. (2020).

    - Transfer learning from e.g. sentiment analysis: Uban & Dinu (2019).

Limited & biased data

- Small data size
    - Abuse-specific embeddings: Caselli et al. (2020).
    - Transfer learning from e.g. sentiment analysis: Uban & Dinu (2019).
- Sampling & representation bias
    - Multi-task learning: Waseem et al. (2018).
    - Use more data(sets), e.g. Park et al. (2018).
    - Data augmentation, e.g. Dixon et al. (2018).

Challenges in generalisable hate speech detection

Implicit hate speech, e.g.:

“Hey Brianne - get in the kitchen and make me a samich. Chop Chop” (Gao and Huang, 2017)

Challenges in generalisable hate speech detection

Implicit hate speech, e.g.:

"Hey Brianne - get in the kitchen and make me a samich. Chop Chop" (Gao and Huang, 2017)

Still in its infancy, some attempts include:

- Consider context beyond just a post, e.g. De Gibert et al (2018).
- Paraphrase implicit statements, e.g. Sap et al (2020).
- Labelling of implicit hate, e.g. Caselli et al (2020), ElSherief et al. (2021).

Questions?

Thanks to:

- Wenjie Yin (@)
- Rabab Alkhalifa (@)
- Aiqi Jiang (@)
- Peiling Yi (@)
- Xia Zeng (@)
- Parisa Jamadi (@)
- Raneem Alharthi (@)
- Amani Abumansour (@)
- Aida Halitaj (@)
- Dina Pisarevskaya (@)
- Noman Ashraf
- Elena Kochkina (@)