FONDAZIONE
BRUNO KESSLER

BUILDING TOOLS AND DATASETS TO DETECT ONLINE HATE SPEECH

Current studies on cyberbullying and islamophobia

Sara Tonelli
"Digital Humanities" research group
Fondazione Bruno Kessler, Trento
satonelli@fbk.eu

Disclaimer

These slides contain examples of language that may be offensive to some readers. Of course, they do not reflect the views of the presenter

HATE SPEECH

Speech that attacks **a person or a group** on the basis of attributes such as **race**, **religion**, **ethnicity**, **sexual orientation**, **disability**, etc.

In some countries it is defined and punished by law, therefore other expressions are used such as **abusive language** or **offensive language**

FONDAZIONE BRUNO KESSLER

WHY WE NEED (multilingual) DATA

Study the phenomenon

Train automatic detection systems

FONDAZIONE BRUNO KESSLER

How datasets are built

Typically focus on **Twitter** for large-scale analyses

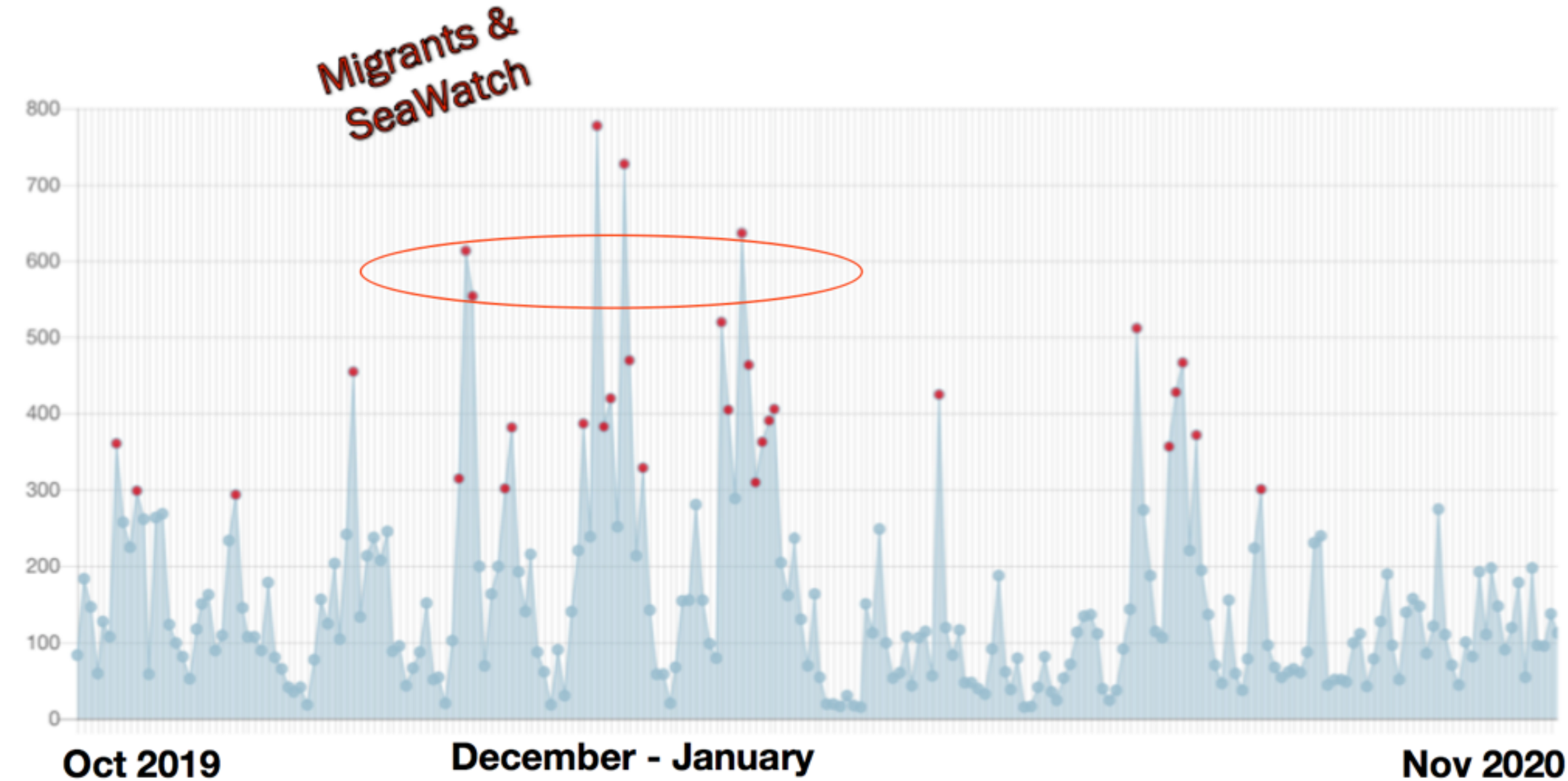
Choose few **hashtags** that are offensive/hateful
e.g. #banIslam, #StopAlleMoschee

Retrieve messages containing the selected hashtags

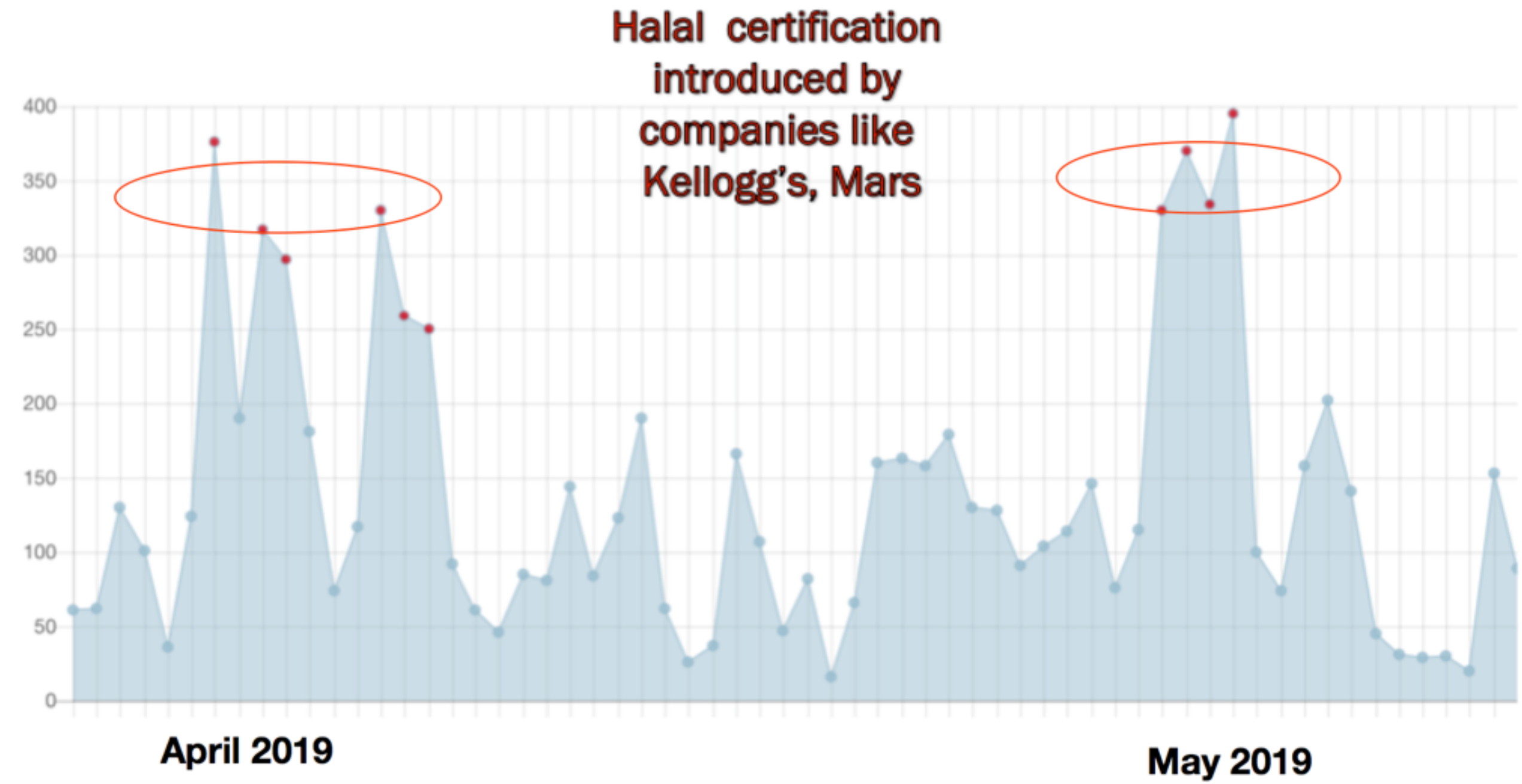
Use data for **quantitative studies**, trend analysis, key issue is how hashtags are chosen

FONDAZIONE BRUNO KESSLER

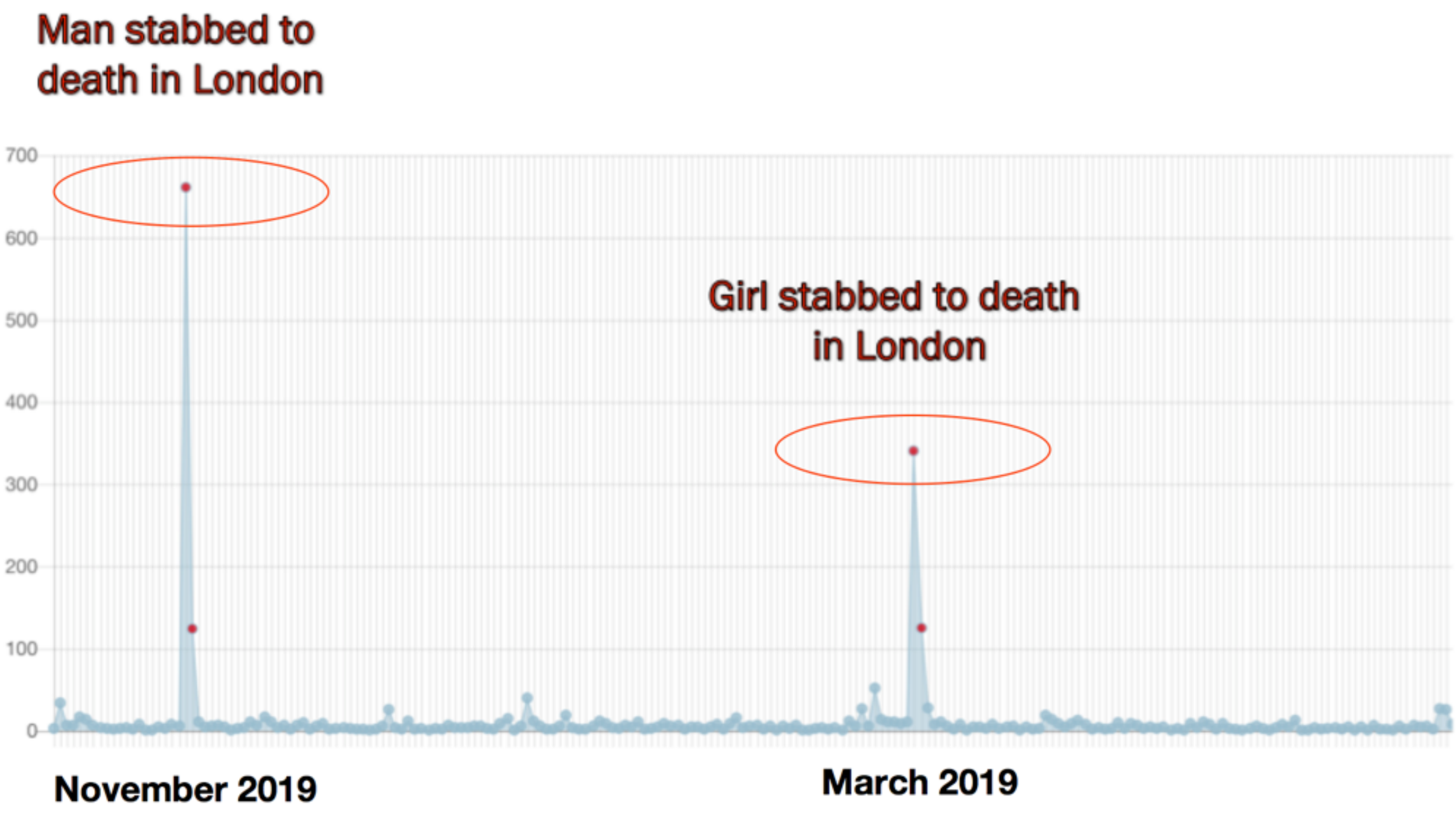
Italian: #stopInvasione



French: #GrandRemplacement



English: #londonistan



Issues with unfiltered datasets

Data Content: Hashtags must be unambiguous and stable over time

In any case, hashtags or search terms are not enough to identify hateful content

Gianrico Carofiglio
@GianricoCarof

La fantasiosa locuzione “fascista liberale” elaborata da un politico in (lunga) vacanza per descrivere suo padre, offre lo spunto per altri esperimenti linguistici: nazista caritatevole; brigatista nonviolento; jihadista tollerante; grillino autoironico. È un mondo difficile.

Translate Tweet

9:12 PM · Nov 17, 2018 · Twitter for iPhone

2.3K Retweets **7.6K** Likes

FONDAZIONE BRUNO KESSLER

Solution: Data annotation

Existing benchmarks and shared tasks: see link at http://hatespeechdata.com

Italian: AMI dataset (Anzovino et al., 2019), HaSpeeDe (Bosco et al. 2019), shared tasks at Evalita 2020 evaluation campaign

Annotation process:

- retrieve data starting from (hateful) hashtags or keywords and then extend with additional (non-hateful) data.
- define annotation guidelines
- annotate data (experts or crowd-sourcing), multiple judgements per item

FONDAZIONE BRUNO KESSLER

Issues

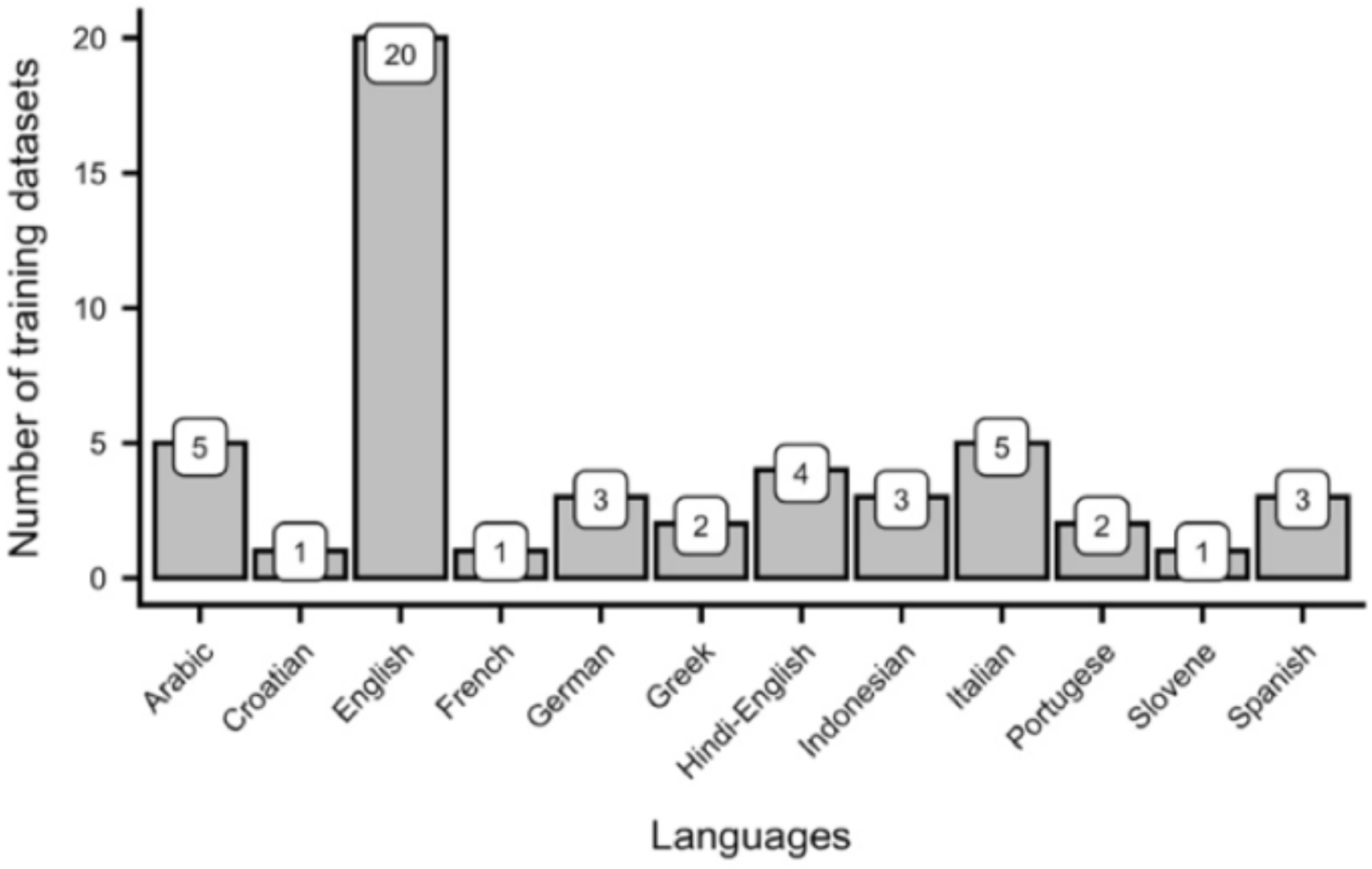
Data Content: Dataset is not representative of Twitter data, biased towards the domain associated with the selected hashtags (*Wiegand et al., 2019*)

Annotator role: Annotators may be more or less engaged in the annotation, their gender, ethnicity, background make them sensitive to some topics or language choices (*Sap et al., 2019*)

Privacy: difficult to share these data for research purposes without infringing GDPR and social media privacy policies

FONDAZIONE BRUNO KESSLER

Language Diversity



Source: Vidgen and Derczynski (2020)

An alternative approach

With financial support of the/Con il supporto finanziario di EIT Digital Call 2018 - Digital Wellbeing Action Line

:C R E E:P

CREEP - CybeRbullying EffEcts Prevention

WhatsApp corpus creation

Cyberbullying **living lab**, 4 classes involved.

Role play: around 10 students in each class, plot provided to start the discussion

Roles assigned with the help of teachers: **cyberbully**, cyberbully **assistant**, victim assistant, **victim**

Pros: WhatsApp dataset, annotation of messages in context (chats)

FONDAZIONE BRUNO KESSLER

Example scenarios

Scenario	Type of addressed problem
Your shy male classmate has a great passion for classical dance. Usually he does not talk much, but today he has decided to invite the class to watch him for his ballet show.	Gendered division of sport practices
Your classmate is very good at school, but does not have many friends, due to his/her haughty and 'teacher's pet' attitude. Few days ago, s/he realised that his/her classmates brought cigarettes to school and snitched on them with the teacher. Now they will be met with a three days suspension, and they risk to fail the year.	Interference in others' businesses
Your classmate is very good at school, and everyone think s/he is an overachiever. S/He studies a lot and s/he never goes out. S/He does not speak much with his/her classmates, that from time to time tease him/her for his/her unsocial life. Things have slightly changed recently: your classmates mum convinced teachers to increase the homework for all the students. A heedless teacher revealed the request to the class, and now some students are very angry at him/her.	Lack of independence, parental intromission.
Your shy classmate is good in all subjects but in gymnastics. For this reason, his/her classmates often tease on him/her when s/he exercises. Recently, the class has found out a video on the social network Musical.ly, where s/he dances gracelessly, on a 90s song that no one has never heard before.	Web virality

Corpus description

10 chats, ~14,600 tokens, manually annotated by two linguists (IAA 0.80 Dice)

Offensive messages are 41% of the chats

Label not only if the message is offensive, but also the type of offence, categories inspired by *Van Hee et al. (2015)*

FONDAZIONE BRUNO KESSLER

Corpus statistics

OFFENCE TYPE	
Defense	381
General insult	313
Curse or Exclusion	200
Threat or blackmail	81
Encouragement to harassment	63
Body shame	45
Discrimination - Sexism	45
Attacking relatives	28
Defamation	23
Other	24
TOTAL	**1,203**

FONDAZIONE BRUNO KESSLER

Discussion & Lessons learned

Religious and **racial** offence not present in the data

Plausibility of the exchanges: assessment by teachers, students and parents

Replicability of the approach, although consent from parents is still an issue

More than >9,000 messages to be released, currently not annotated with offence type

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini and Enrico Piras. *Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying*. Proceedings of the 2nd Workshop on Abusive Language Online, Bruxelles, 2018

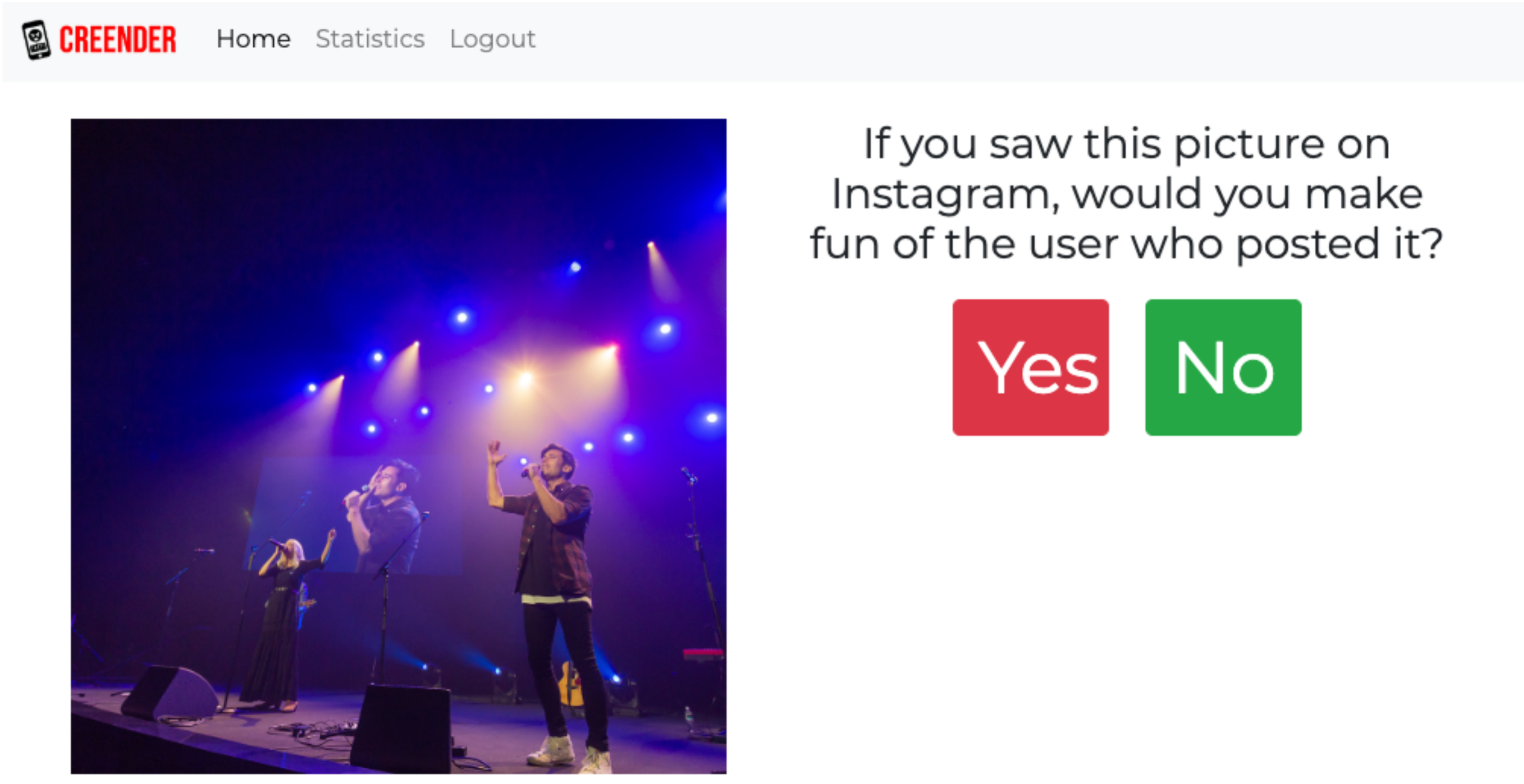
More data for Italian?

Teenagers use mainly video or image-based social networks such as **Instagram**, TikTok, musical.ly

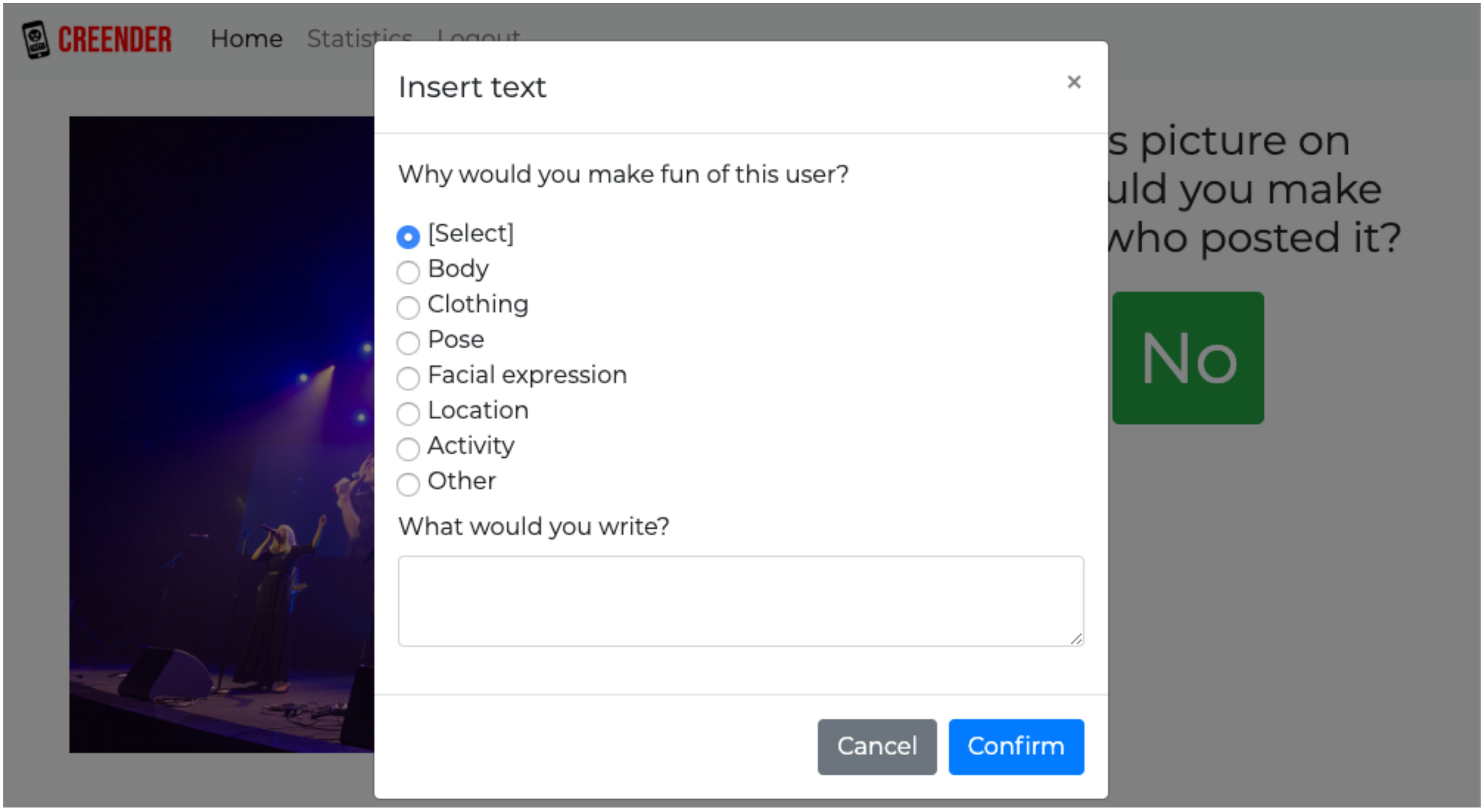
These data are difficult to collect and to study, Instagram stories disappear after 24 hours

Extend simulation exercises to images + text similar to Instagram, 95 high-school students involved

Image + text annotation

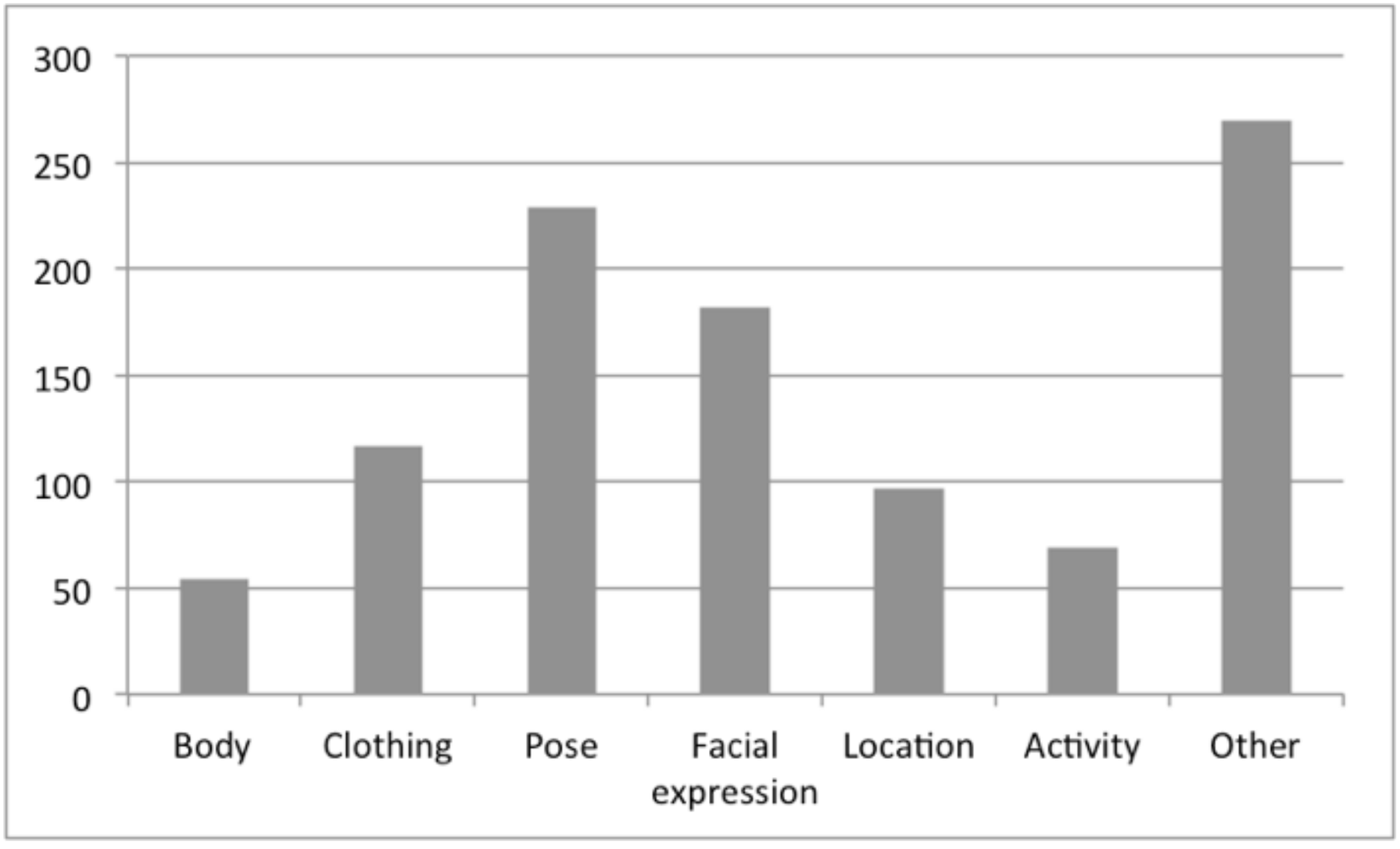


CREENDER categories



Data analysis

17,912 images have been judged at least once. 1,018 of them (5.7%) have received at least one offensive comment



Data analysis

Manual annotation of the picture subject for the images having at least one offensive comment

	Females	Males	Mixed	Nobody
Body	27	20	3	4
Clothing	66	30	9	12
Pose	114	99	11	5
Facial Expression	68	90	17	7
Location	16	17	7	57
Activity	12	14	7	36
Other	72	63	22	113
Total	377	318	76	252

Data analysis

Manual annotation of 3,200 pictures randomly taken from those with no comments, and compare the distribution with pictures having at least one comment

	% Yes	% No
Females	36.85	32.14
Males	31.09	19.00
Mixed	7.43	9.33
Nobody	24.63	39.53

Pictures with no human subject are less likely to get an offensive comment. Female subjects are the most commented, statistically significant differences between the two distributions.

FONDAZIONE BRUNO KESSLER

From annotation to automatic detection

For English best approaches are supervised and use deep learning = data-hungry

Challenges:

- Multilinguality
- Create robust systems dealing with different platforms, datasets are platform-specific
- Multimodality
- Detection in context

Multilinguality

OffensEval 2020 shared task on Multilingual Offensive Language Identification in Social Media include English, Danish, Turkish, Arabic and Greek

Best system on English 0.92 F1, best system on Danish 0.81

Multilingual transformer-based models (e.g. multilingual BERT) are a promising research direction

https://sites.google.com/site/offensevalsharedtask/home

Robustness

- Datasets created with data from a social media platform do not yield good results when used to train a classifier for another platform
- Our findings (*Corazza et al., 2019*): merging data from different platforms improves classification results, except for Twitter. Single-source data are better only when the training set is big (few thousand annotated examples)

FONDAZIONE BRUNO KESSLER

Multimodality

- Most offensive messages consist of a combination of different modalities (images/videos + text)

- NLP community focuses on text, but a really effective classification should be able to integrate all components of the communication

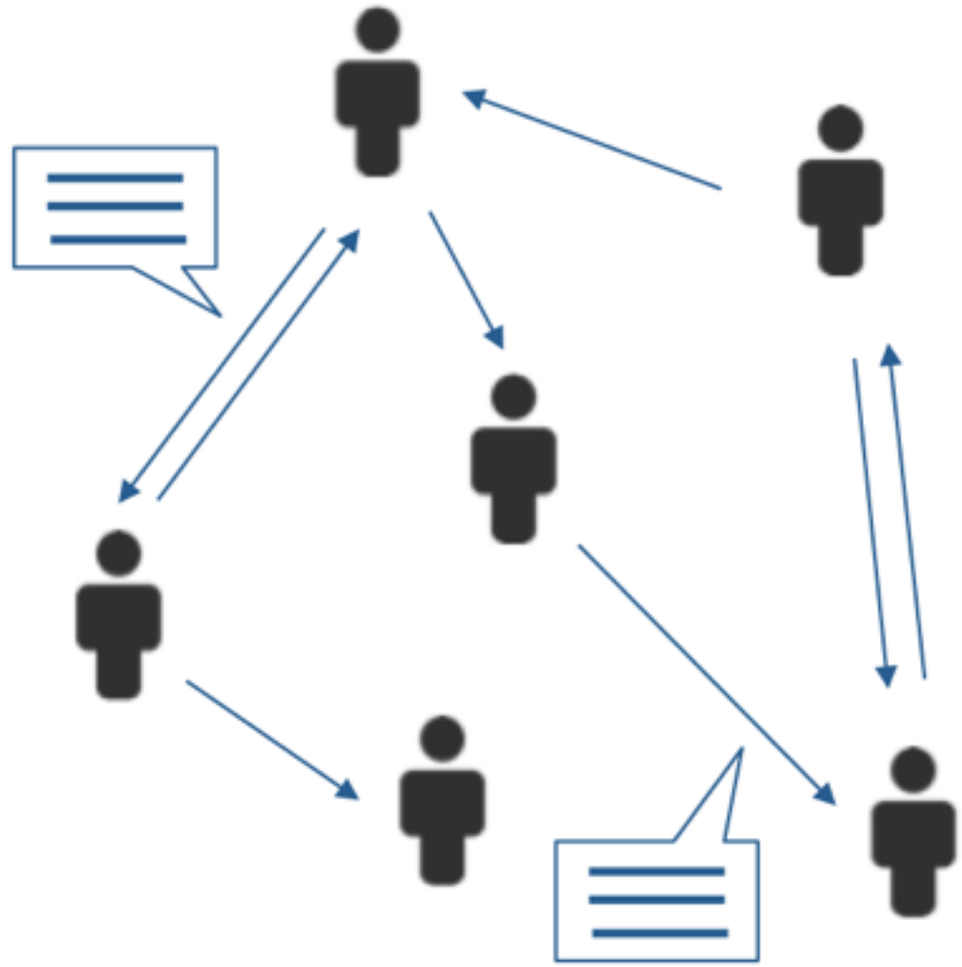
(*Gomez et al., 2020*)

RT @CarSalesBossMan: I'm not sexist but http://t.co/tkYiFGEs16

Context

- Most datasets for hate speech detection contain **single messages** with no context, and systems learn to classify single messages

- Messages need (**textual**) **context** to be interpreted as offensive or not, as well as information on the **user network** where conversations take place (*Menini et al., 2019*)



Conclusions

- Hate speech analysis and detection is a **very difficult task**, starting from the definition and (manual) identification of what is offensive

- For automatic detection, several **sources of error**:
  - use of dialect and slang "*en se ponno senti*"
  - bad orthography "*Io no nesdune delle due*"
  - sarcasm
  - world knowledge "*un certo Adolf sarebbe utile ancora oggi*"
  - metaphorical expressions / creative language "*ruspali*"

- In the future, focus on different languages (not only English), multimodality and context to build better datasets and better detection systems

FONDAZIONE BRUNO KESSLER

References

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, Serena Villata. *A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis*. In Proceedings of the 3rd Workshop on Abusive Language Online, 2019.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini and Enrico Piras. *Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying*. Proceedings of the 2nd Workshop on Abusive Language Online, Bruxelles, 2018

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, Serena Villata. *Cross-Platform Evaluation for Italian Hate Speech Detection*. In Proceedings of CLIC-it 2019, Bari, Italy, 2019.

Raul Gomez, Jaume Gibert, Lluis Gomez and Dimosthenis Karatzas. *Exploring Hate Speech Detection in Multimodal Publications* In Proceedings of WACV, Aspen, CO, 2020.

Bertie Vidgen and Leon Derczynski. *Directions in Abusive Language Training Data*. arXiv, 2020.

Martin Wiegand, Josef Ruppenhofer, Thomas Kleinbauer. *Detection of Abusive Language: The Problem of Biased Datasets* In Proceedings of NAACL, 2019.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, Noah Smith. *The Risk of Racial Bias in Hate Speech Detection* In Proceedings of ACL, Florence, Italy, 2019.

FONDAZIONE BRUNO KESSLER

Thank you !

satonelli@fbk.eu
https://dh.fbk.eu

FONDAZIONE BRUNO KESSLER