me

Valerio Basile
Postdoc at University of Turin

Content-centered Computing

Hate speech monitoring group
http://hatespeech.di.unito.it/

TWO PROJECTS against HS:

2016-2018: HS & Social Media
(*Hate Speech and Social Media*)

2017-2019: IHatePrejudice
(*Immigrants, Hate and Prejudice in Social Media*)

A COMMON GOAL:
building a framework for collecting, analyzing and displaying big data about HS

Compagnia
di San Paolo

Fondazione
CRT

HS & Social Media

The main goal of the project is the development of a framework for collecting, analyzing and displaying big data, which can be exploited by teachers and students (within high schools) for the improvement of their knowledge about the HS in Piedmont, and for promoting the intercultural and multi-ethnic integration

IHatePrejudice

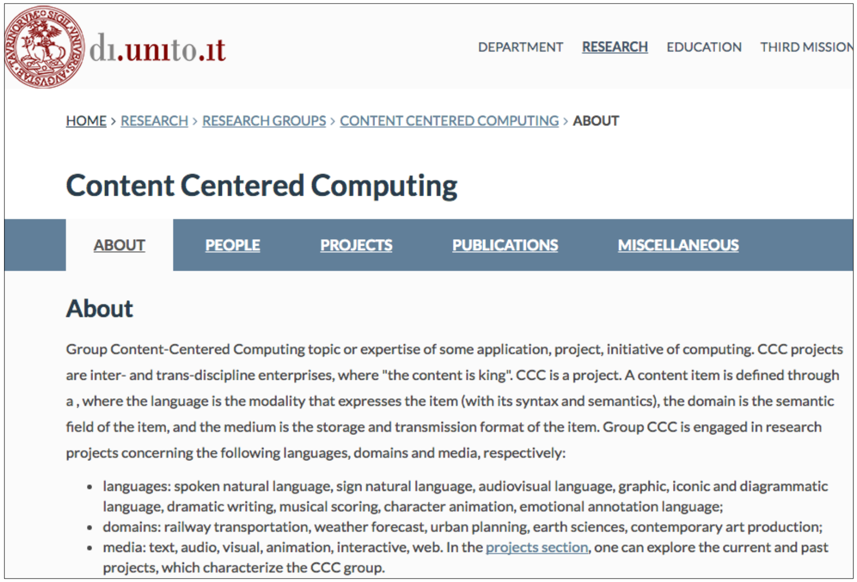
The main goal of the project is the development of a framework for collecting, analyzing and displaying big data about the HS in Piedmont in particular when oriented against immigrants, for promoting the intercultural and multi-ethnic integration sharing such knowledge with operators working in the local area but also with citizens for increasing awareness and contribute to the work of policy-makers.

Competences involved in the projects

- Natural Language Processing Sentiment Analysis and Opinion Mining, development of linguistic resources

- Data Analysis applied to social media for detecting the dynamics for the diffusion of the HS

- Data Visualization interactive visualization of complex information for allowing the acces to data previously collected and analyzed

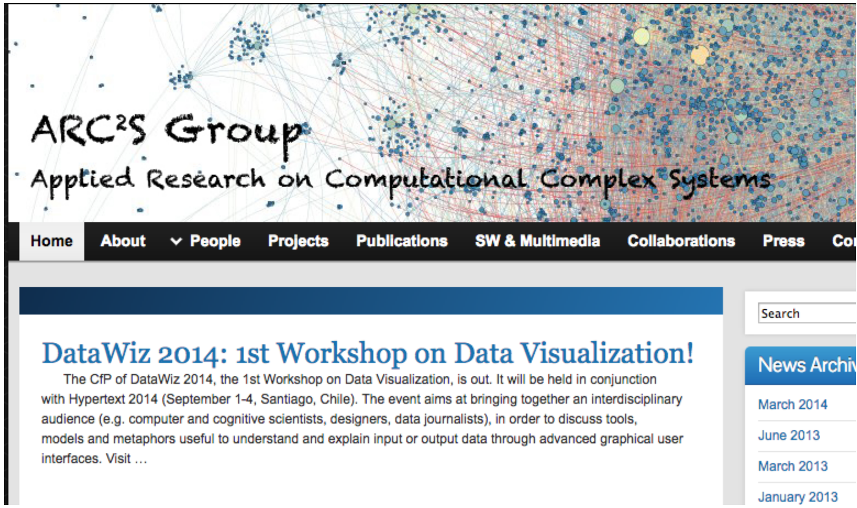


Content Centered Computing

About

Group Content-Centered Computing topic or expertise of some application, project, initiative of computing. CCC projects are inter- and trans-discipline enterprises, where "the content is king". CCC is a project. A content item is defined through a , where the language is the modality that expresses the item (with its syntax and semantics), the domain is the semantic field of the item, and the medium is the storage and transmission format of the item. Group CCC is engaged in research projects concerning the following languages, domains and media, respectively:

- languages: spoken natural language, sign natural language, audiovisual language, graphic, iconic and diagrammatic language, dramatic writing, musical scoring, character animation, emotional annotation language;
- domains: railway transportation, weather forecast, urban planning, earth sciences, contemporary art production;
- media: text, audio, visual, animation, interactive, web. In the projects section, one can explore the current and past projects, which characterize the CCC group.



DataWiz 2014: 1st Workshop on Data Visualization!

The CfP of DataWiz 2014, the 1st Workshop on Data Visualization, is out. It will be held in conjunction with Hypertext 2014 (September 1-4, Santiago, Chile). The event aims at bringing together an interdisciplinary audience (e.g. computer and cognitive scientists, designers, data journalists), in order to discuss tools, models and metaphors useful to understand and explain input or output data through advanced graphical user interfaces. Visit …

Motivations

- In the last years several Italian and European laws have been promulgated for contrasting the public incitement to hatred towards e.g. ethnic and religious minorities

- Nevertheless the *Hate Speech* (HS) is continuously increasing, together with the change of the society determined by the immigration from Africa and East countries

Motivations

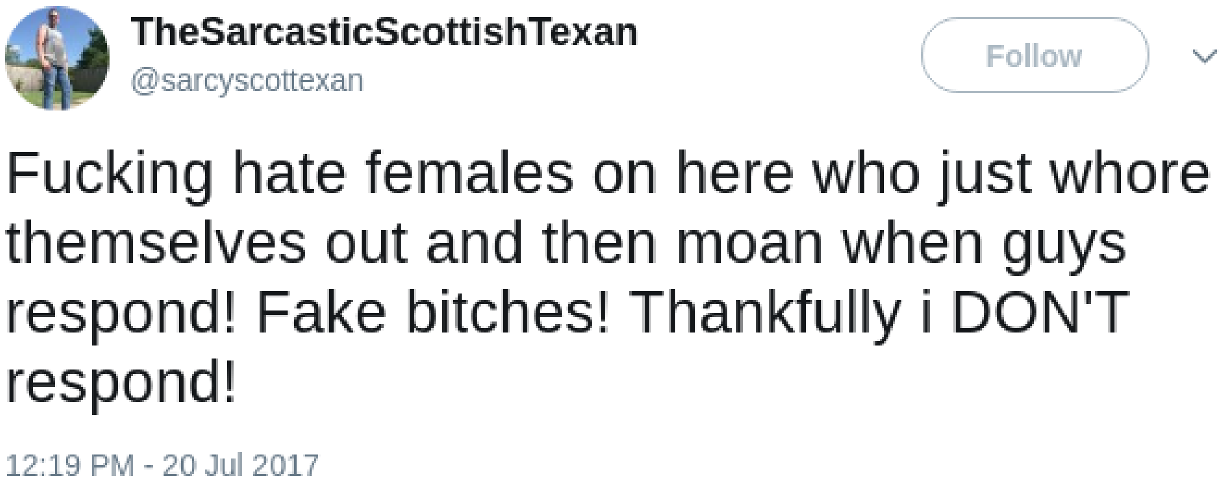
The need for preventive actions against HS is crucial within the education area and schools, where the percentage of students born in Italy by families of migrants is growing steadily (more than 12% in 2014)

Since blogs, fora and social networks can be often vectors for HS, informed preventive actions can be based on the analysis of texts from social media

motivation

- European Union Commission directives.
- Automatic techniques not available.
- Lack of data about hate speech.
- Hate speech removal.
- Quality of service.

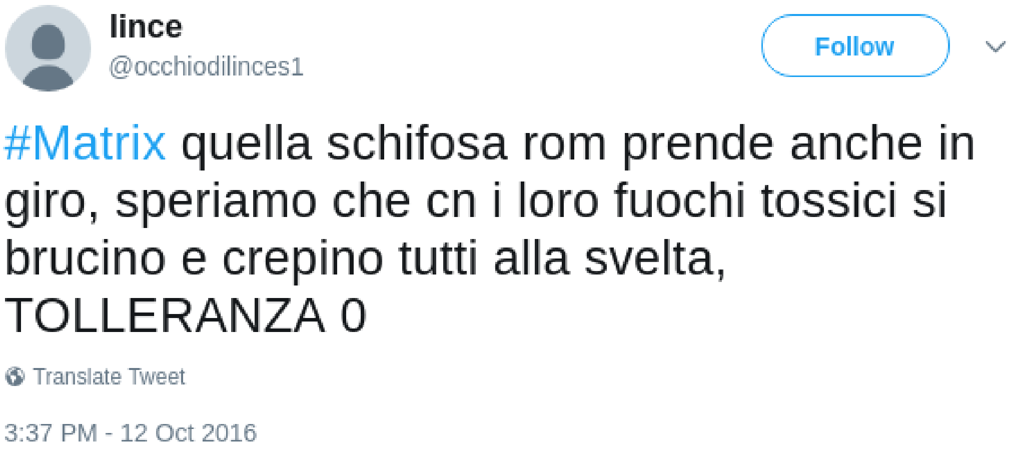
hate speech in social media



TheSarcasticScottishTexan
@sarcyscottexan

Fucking hate females on here who just whore themselves out and then moan when guys respond! Fake bitches! Thankfully i DON'T respond!

12:19 PM - 20 Jul 2017



lince
@occhiodilinces1

#Matrix quella schifosa rom prende anche in giro, speriamo che cn i loro fuochi tossici si brucino e crepino tutti alla svelta, TOLLERANZA 0

Translate Tweet

3:37 PM - 12 Oct 2016

Demos (UK)
2014: 10,000-15,000 racist messages
2016: More than 200,000 sexist messages
DAILY

hate speech in social media

“(language that is) abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination.

It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth”

(Erjavec and Kovacic, 2012)

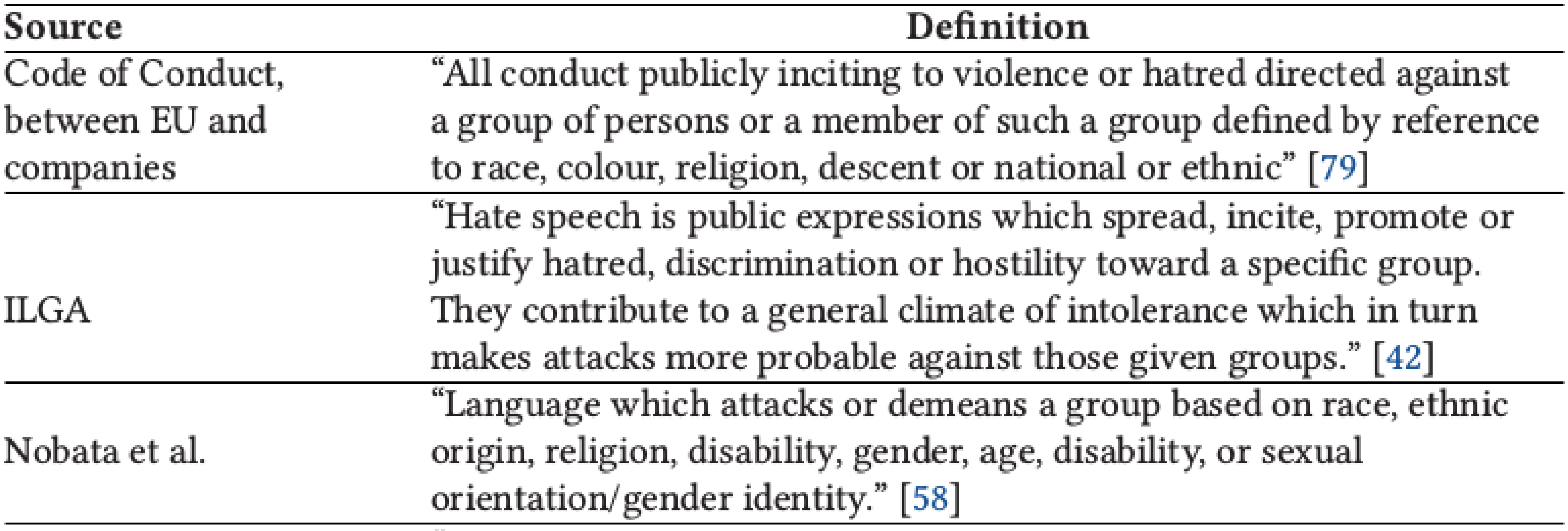
hate speech in social media

A Survey on Automatic Detection of Hate Speech in Text

Paula Fortuna and Sergio Nunes

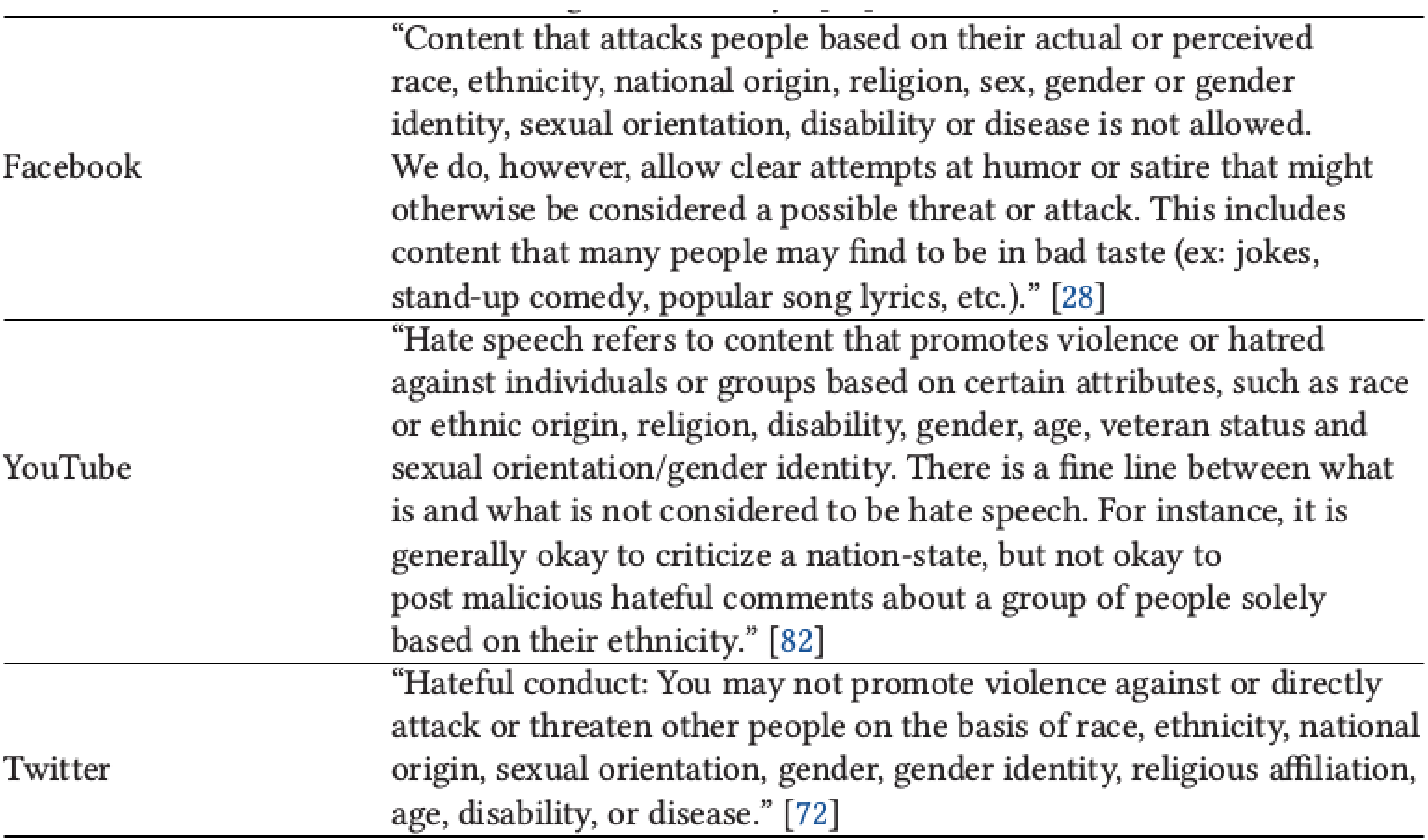
ACM Computing Survey 51, 4, Article 85 (July 2018)

hate speech in social media



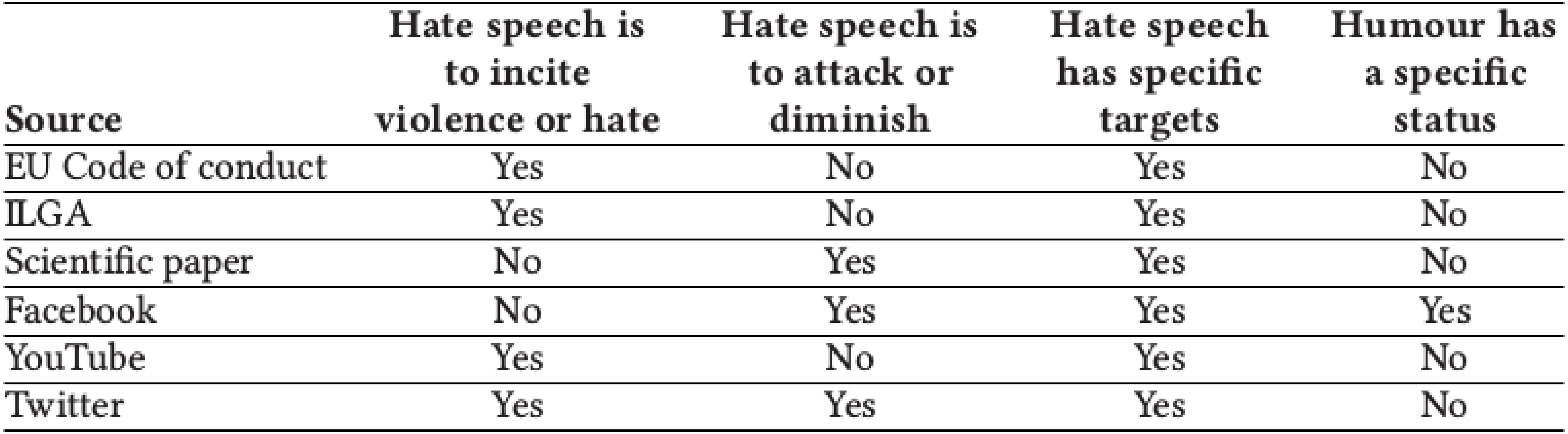
Source	Definition
Code of Conduct, between EU and companies	“All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic” [79]
ILGA	“Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups.” [42]
Nobata et al.	“Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity.” [58]

hate speech in social media



Facebook	"Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.)." [28]
YouTube	"Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity." [82]
Twitter	"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease." [72]

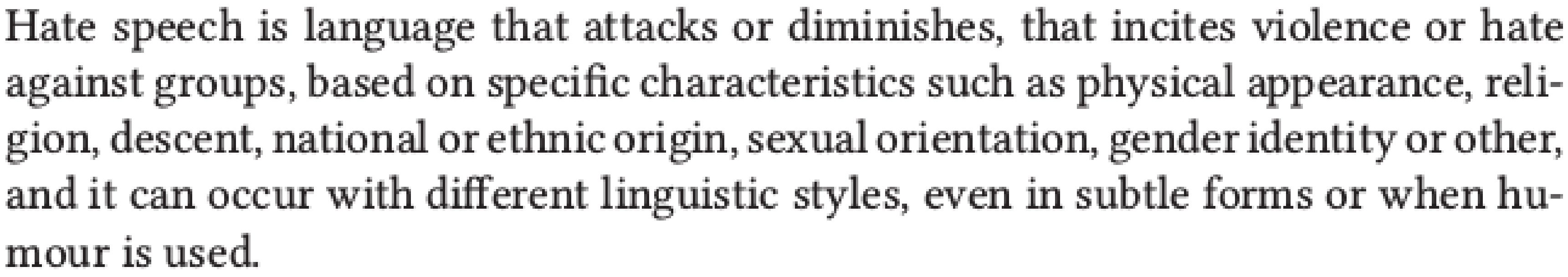
hate speech in social media



Source	Hate speech is to incite violence or hate	Hate speech is to attack or diminish	Hate speech has specific targets	Humour has a specific status
EU Code of conduct	Yes	No	Yes	No
ILGA	Yes	No	Yes	No
Scientific paper	No	Yes	Yes	No
Facebook	No	Yes	Yes	Yes
YouTube	Yes	No	Yes	No
Twitter	Yes	Yes	Yes	No

hate speech in social media

Definition by Fortuna and Nunes



Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.

hate speech in social media

Definition by Poletto and Sanguinetti

Whenever both factors happen to co-occur in the same tweet, we consider it as a HS case:

- the tweet should be addressed, or just refer to, one of the minority groups identified as HS targets, or to an individual considered for its membership in that Category;
- the action, or more precisely the illocutionary force of the utterance, in that it is capable of spreading, inciting, promoting or justifying violence against a target.

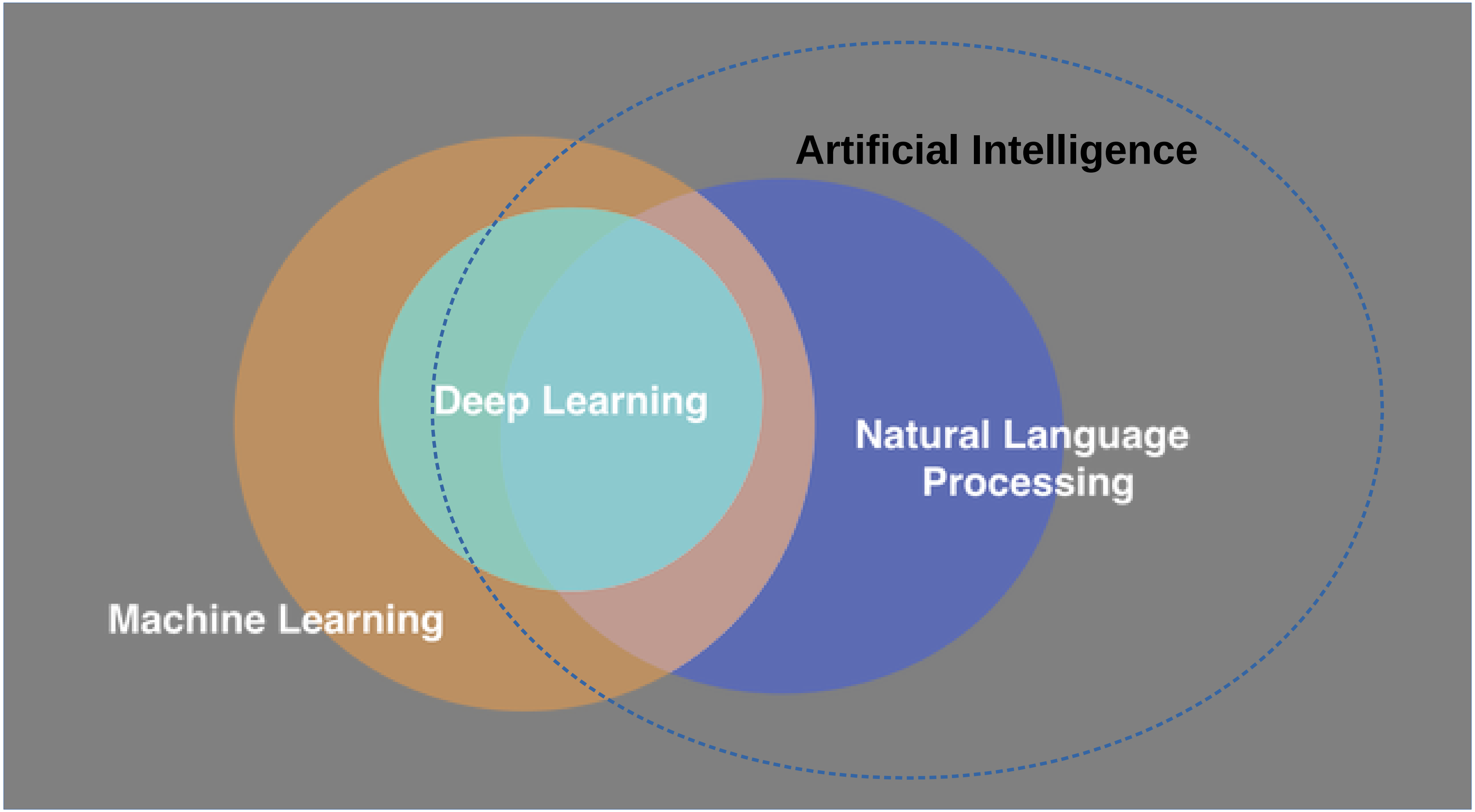
hate speech detection

Typically addressed as a text classification task

Binary or multi-label

Supervised

natural language processing



from quora.com

machine learning and NLP

Example: Support Vector Machine with Bags of Words

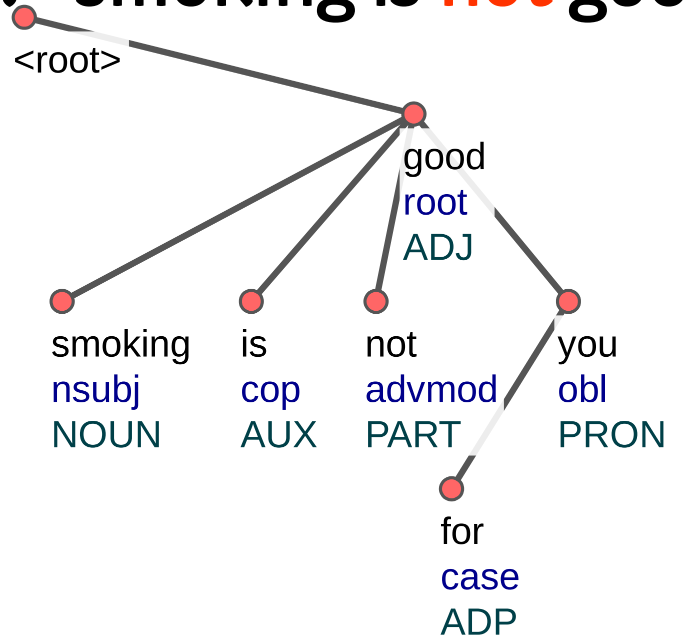
14-ExLab@UniTo:
Automatic Misogyny Detection at IberEval 2018
1st place on English (91.3% accuracy)
and Spanish (81.5% accuracy)

SVM with Bags of Words
+ Twitter-specific features
+ target-specific features

http://ceur-ws.org/Vol-2150/AMI_paper2.pdf

neural and deep

Words in natural language are not isolated.
e.g. "smoking is not good for you"



<root>

good
root
ADJ

smoking
nsubj
NOUN

is
cop
AUX

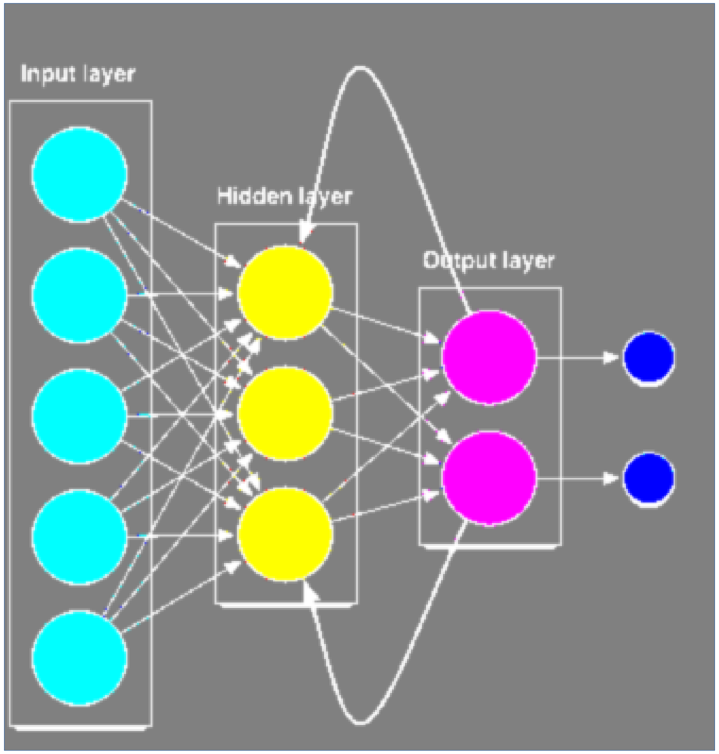
not
advmod
PART

you
obl
PRON

for
case
ADP

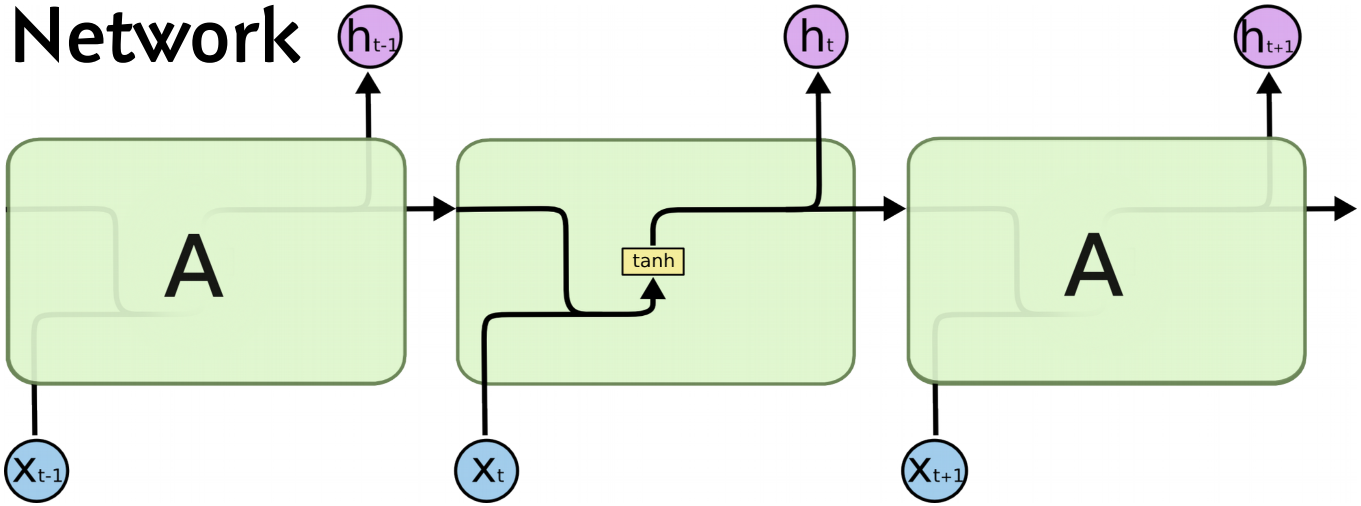
If words make features,
we need to model feature interaction

Natural language comes in sequences
→ recurrent architectures



neural and deep

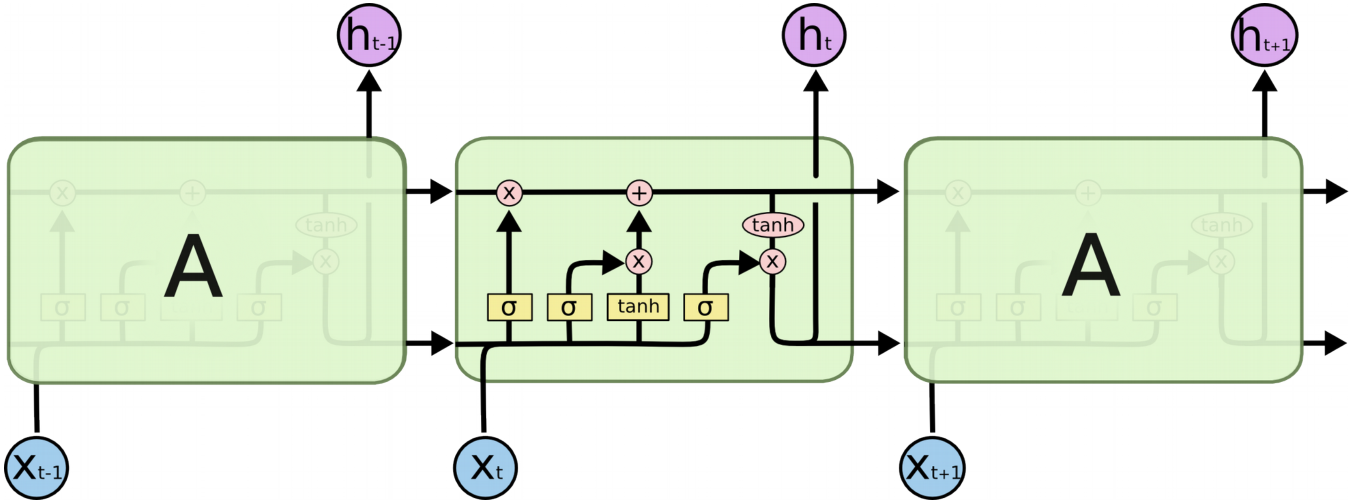
Recurrent Neural Network



Long Short-term Memory network

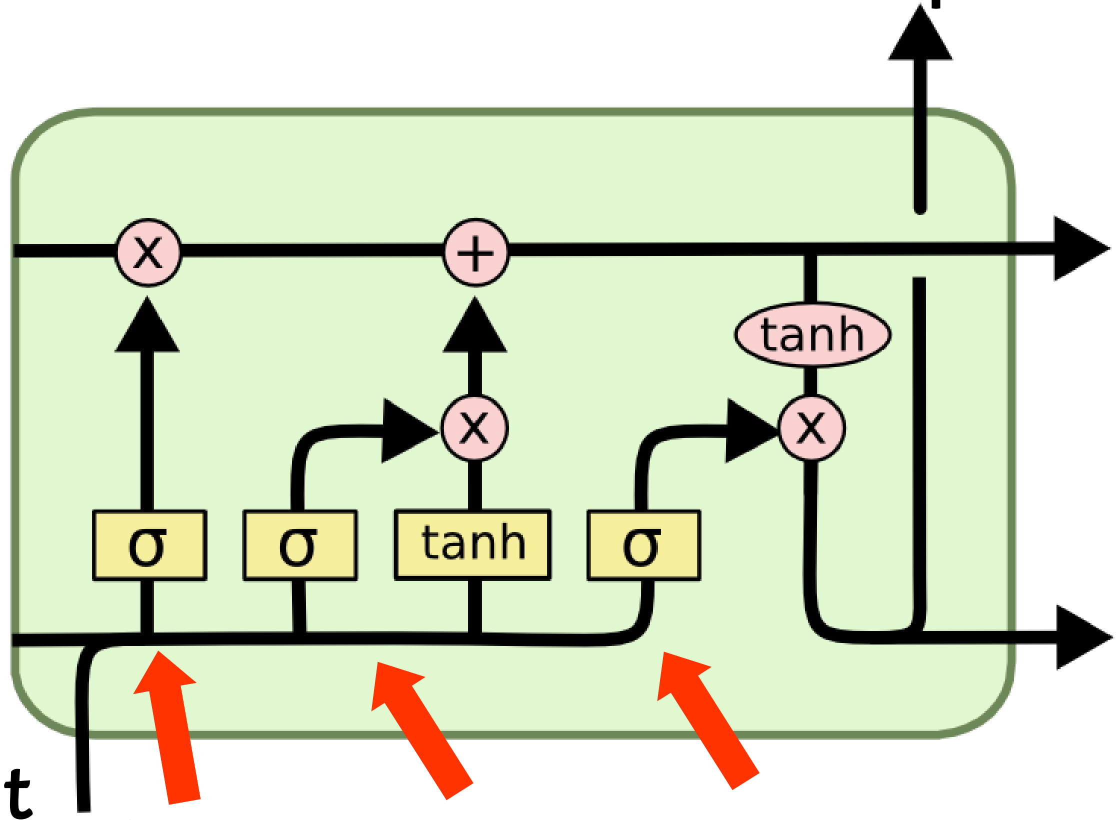
by Hochreiter
& Schmidhuber
(1997)

One word
at a time!



neural and deep

LSTM unit



output

state

input

forget gate

input gate

output gate

from Chris Olah's blog http://colah.github.io/

Neural language models

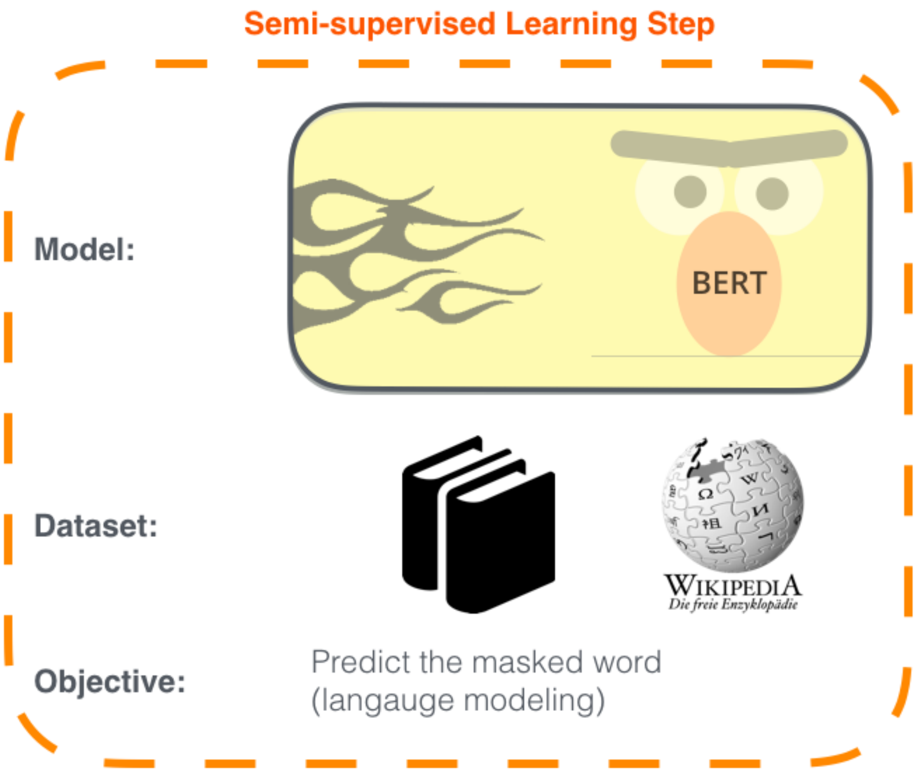
A new generation of language models based on deep learning (e.g. Transformer)

- GPT(-2)
- ELMo
- BERT
- XLnet

Neural language models

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



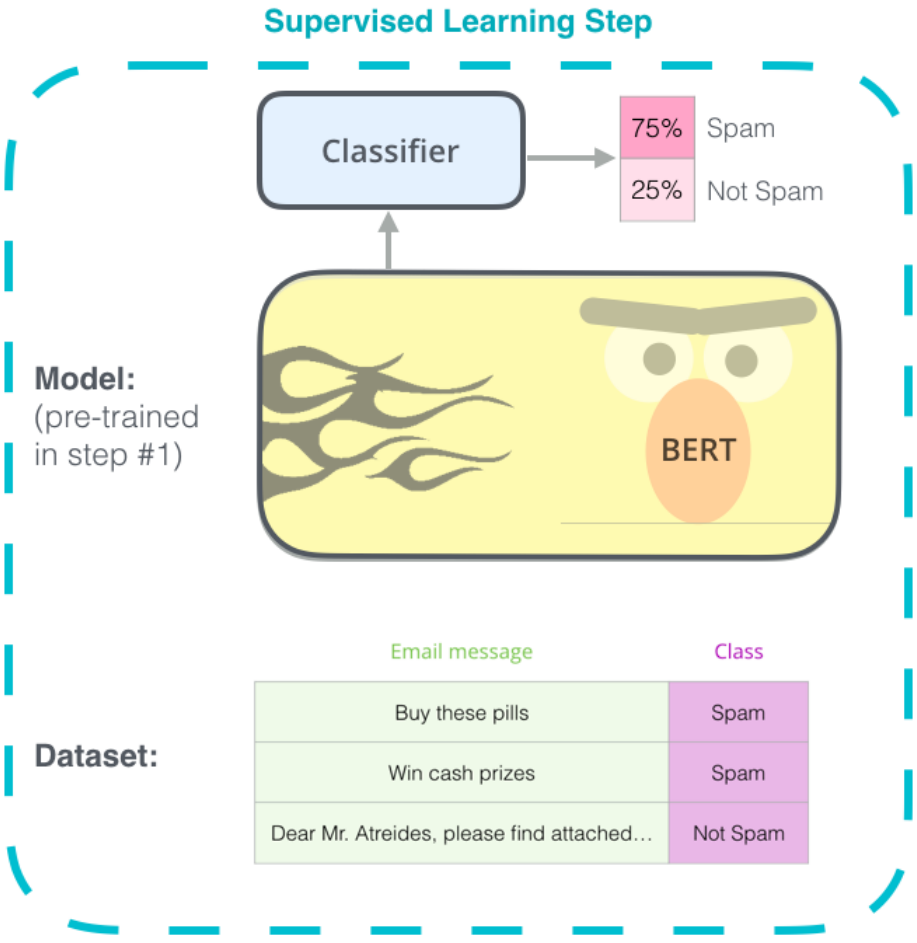
Semi-supervised Learning Step

Model:

Dataset:

Objective: Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.



Supervised Learning Step

75%	Spam
25%	Not Spam

Model: (pre-trained in step #1)

Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

SemEval-2019 Task 5

Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter

Valerio Basile, Cristina Bosco,
Elisabetta Fersini, Debora Nozza,
Viviana Patti, Francisco Rangel,
Paolo Rosso, Manuela Sanguinetti

SemEval-2019 Task 5

- Hate Speech (HS):
  Any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. (Nockleby, 2000)
- Key aspects feature online HS, such as virality, or presumed anonymity, which distinguish it from offline communication and make it potentially also more dangerous and hurtful.
- Targets: Women (Manne, 2017) and Immigrants (Bosco et al., 2017)

Task Description

- Subtask A
  - Hate Speech (HS, binary classification)
- Subtask B
  - Target (TR, individual/group)
  - Aggressiveness (AG, binary classification)
- Source: Twitter
- Languages: English and Spanish

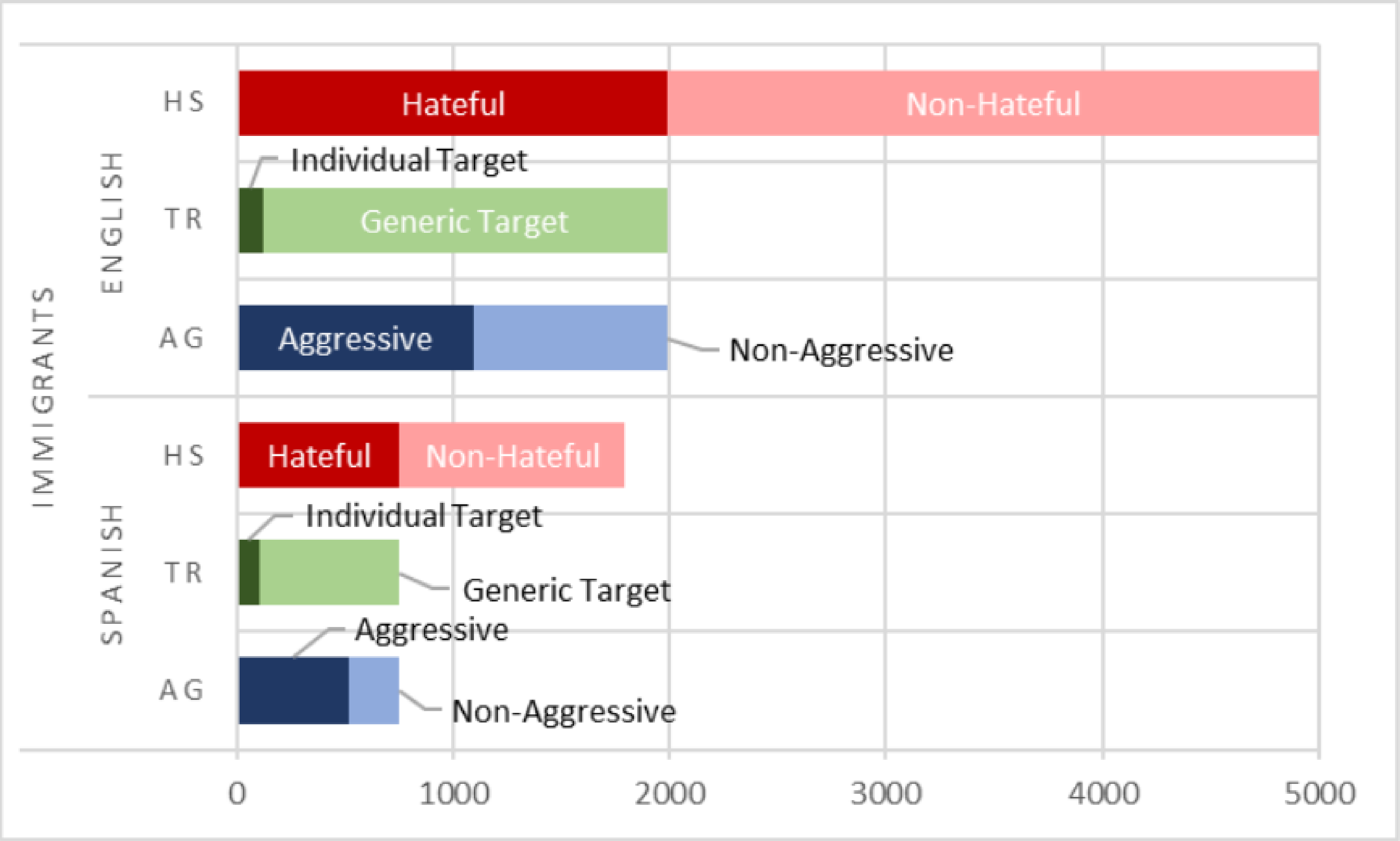
Data

- Keyword-driven approach
  - neutral keywords (Sanguinetti et al., 2018)
  - derogatory words against the targets
  - highly polarized hashtags
- Women target only:
  - monitoring potential victims of hate accounts
  - history of identified haters
- Collected from July to September 2018
  - Except for Women-targeted training (data from two AMI tasks)

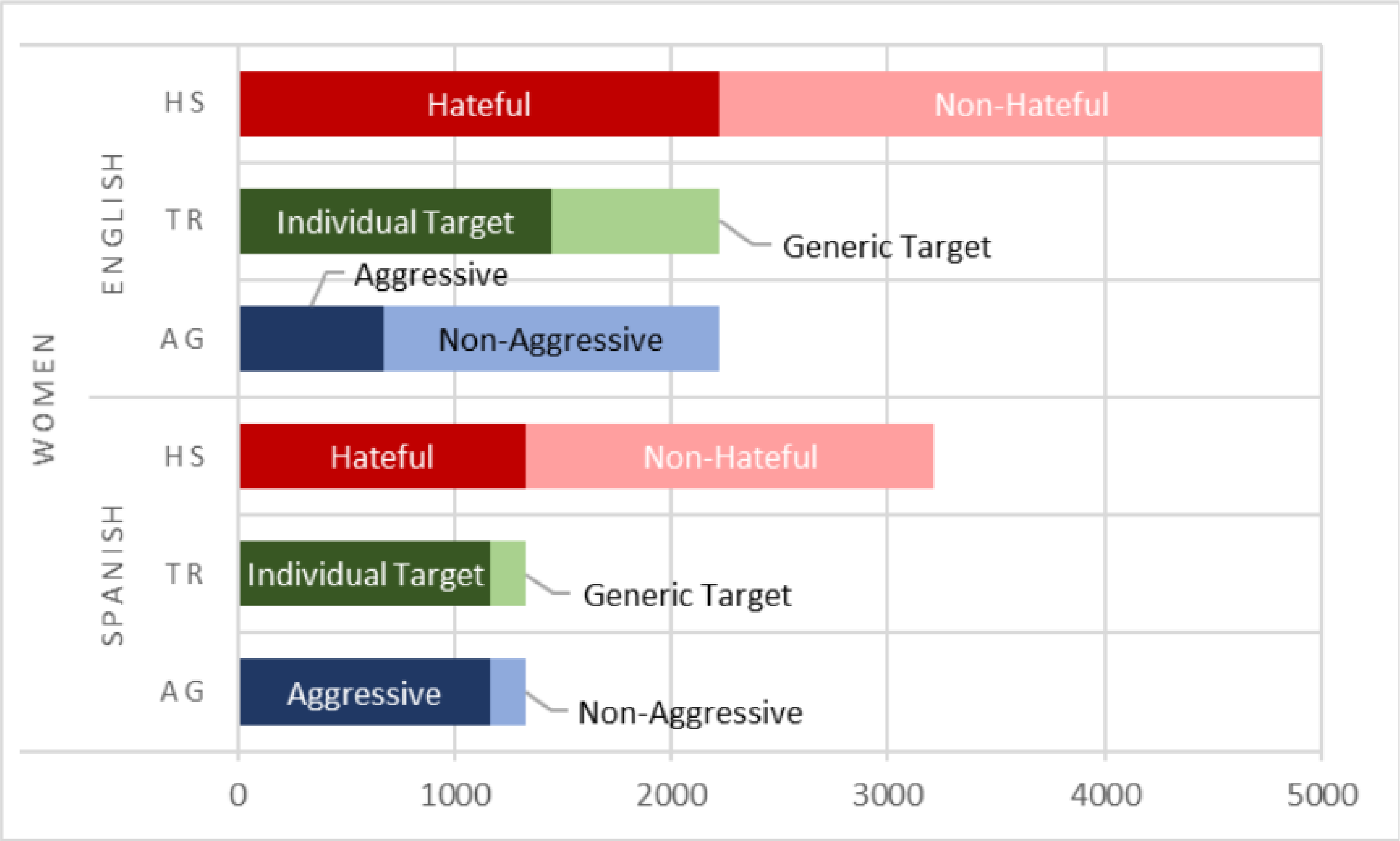
Annotation

- Crowdsourcing (Figure Eight)
- Guidelines in English and Spanish
  - Definition for hate speech against the two targets
  - Definition of aggressiveness
  - List of examples
- Two additional expert annotators (Basile et al., 2018)
- HS distribution is over-represented
- AG and TR distributions are natural

Data Distribution: Immigrants



Data Distribution: Women



Evaluation

- Subtask A
  - Accuracy, Precision, Recall, (macro-)F1
- Subtask B
  - Macro-F1
  - Exact Match Ratio
- Baselines
  - Most Frequent Class (MFC)
  - Support Vector Machine (SVM) based on a TF-IDF representation

Participants

- 74 teams
- 108 runs for Subtask A
- 70 runs for Subtask B.
- 22 teams participated to all the subtasks for the two languages
- 534 subscribers to CodaLab
- 236 subscribers to the Google Group

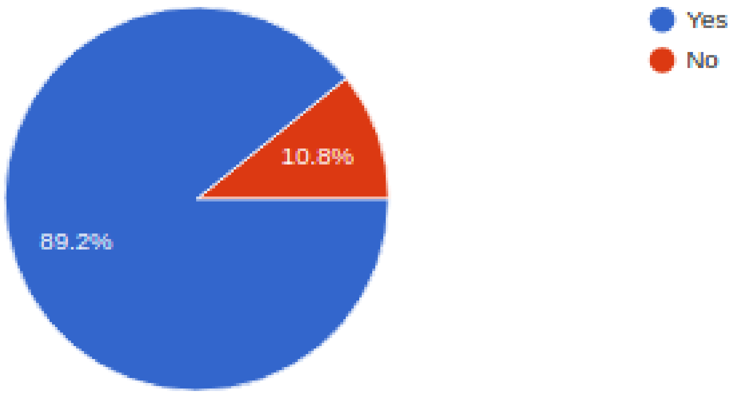
Results

- Approaches
  - Deep Learning (RNN in particular) → more than 1/2
  - Word Embeddings (GloVe mostly)
- Preprocessing
  - Mostly standard
  - Twitter-driven: hashtag segmentation, slang conversion, emoji translation
- Custom hate lexicons

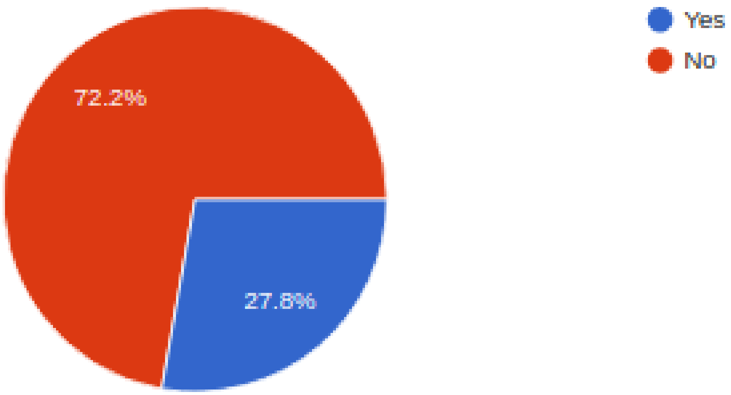
Results

Is your system trained on the officially distributed training set only?

37 responses



Did you participate in other SemEval 2019 tasks with this system?



Discussion

- MFC outperformed all systems in EN B subtask
- Target distinction has been ignored
- Beyond text classification?
- Definition of HS → Eurocentric?
- Many participants, little analysis

Hurtlex

Multilingual lexicon of "words to hurt"

53 languages

17 categories + stereotype

asf	no	puss
cds	no	fiend
re	no	miscreants
ddp	yes	stupefy
an	no	diddlysquat
asf	no	boob
cds	no	fib
cds	no	puke
pr	no	streetwalker
re	no	terraist
cds	no	police-man
cds	no	gangs
cds	no	hypersexuals
asf	no	imbecility
ps	yes	stupidhead
asm	no	sap

http://hatespeech.di.unito.it/resources.html

Related tasks

Sentiment Analysis (SemEval)

Stance Detection (SemEval)

Irony and Sarcasm (SemEval)

Fake news (Fake news challenge)

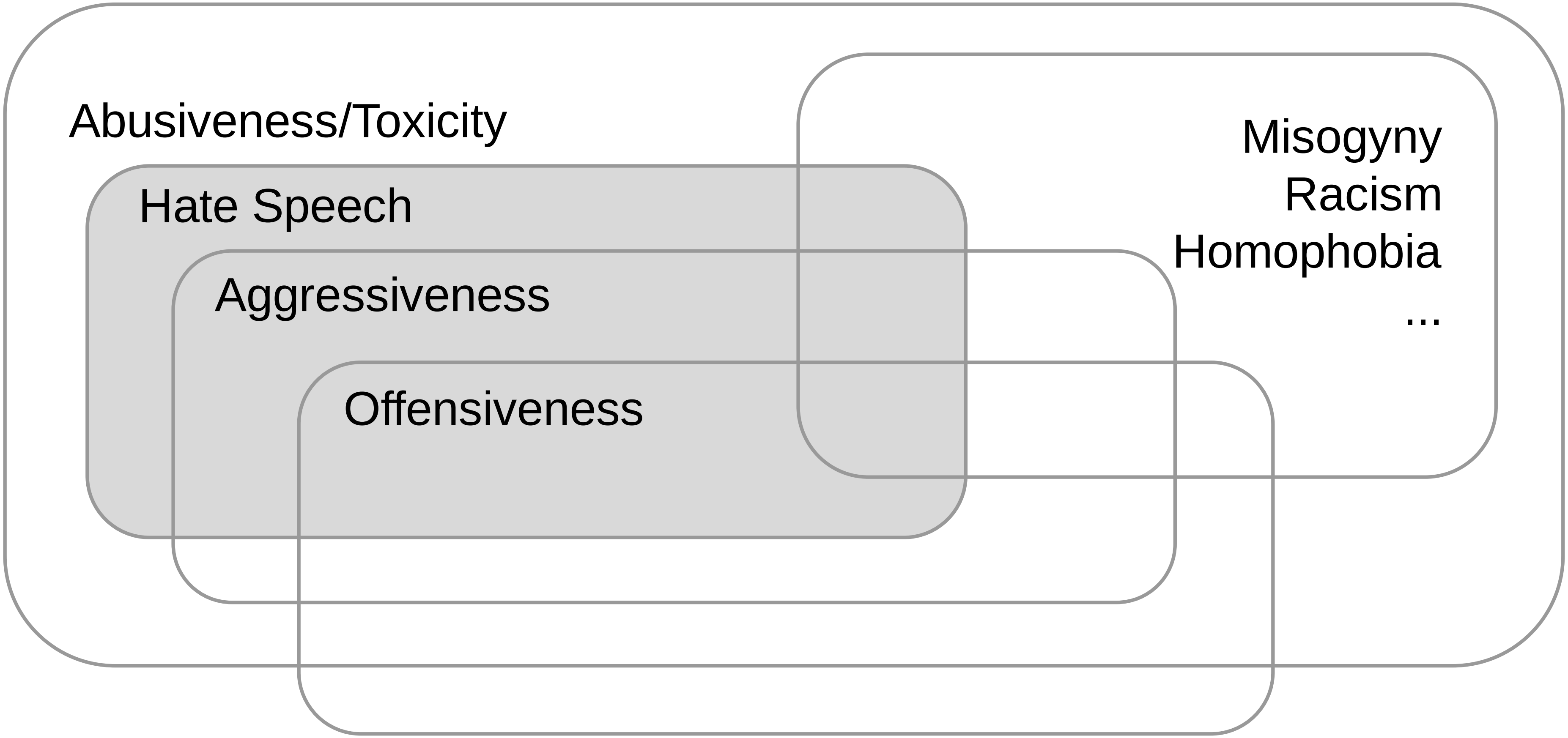
Troll identification

Rumor detection (e.g. RumourEval)

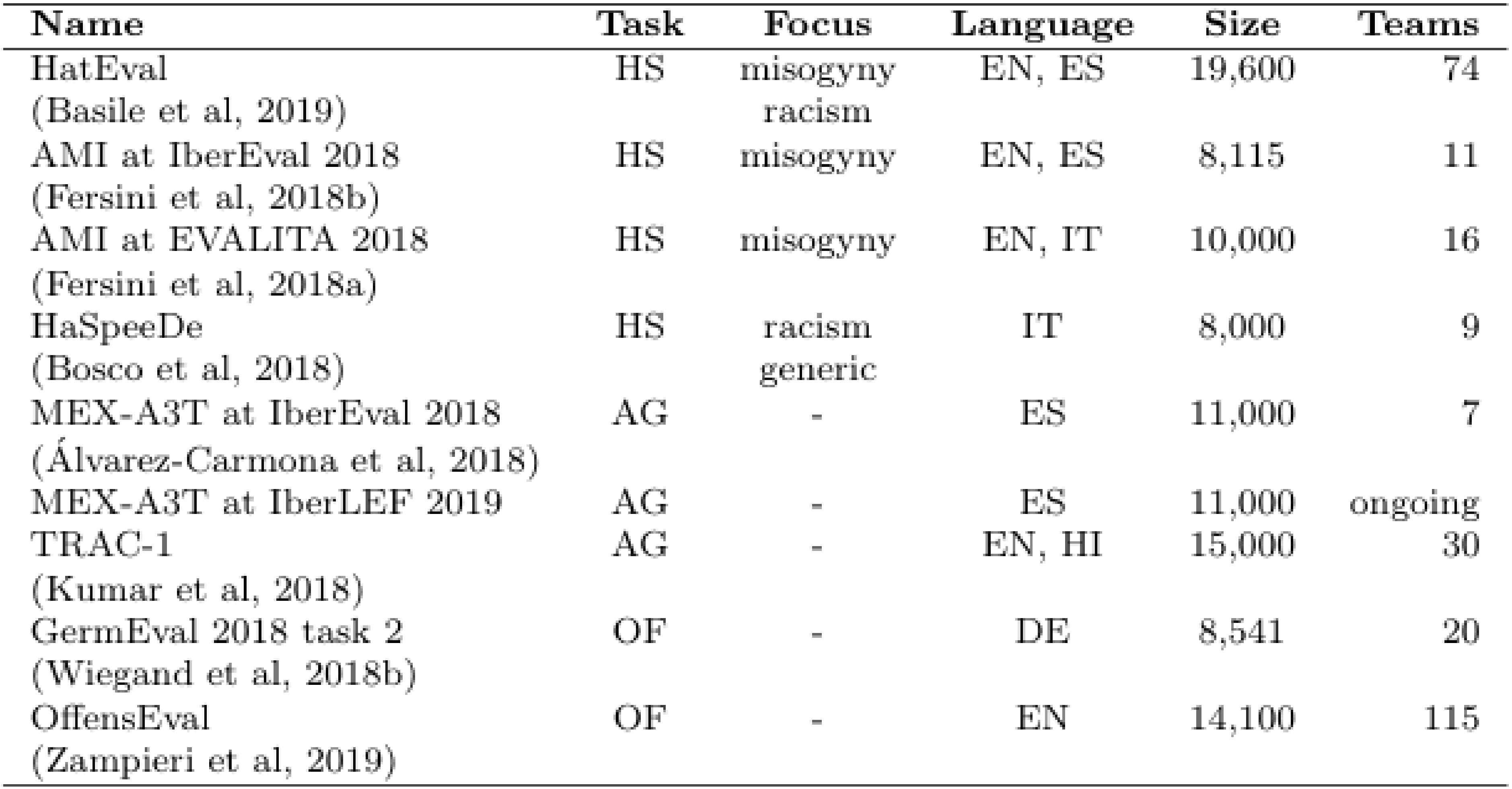
Terrorism and threat identification

…

Related tasks



Related tasks



Name	Task	Focus	Language	Size	Teams
HatEval (Basile et al, 2019)	HS	misogyny racism	EN, ES	19,600	74
AMI at IberEval 2018 (Fersini et al, 2018b)	HS	misogyny	EN, ES	8,115	11
AMI at EVALITA 2018 (Fersini et al, 2018a)	HS	misogyny	EN, IT	10,000	16
HaSpeeDe (Bosco et al, 2018)	HS	racism generic	IT	8,000	9
MEX-A3T at IberEval 2018 (Álvarez-Carmona et al, 2018)	AG	-	ES	11,000	7
MEX-A3T at IberLEF 2019	AG	-	ES	11,000	ongoing
TRAC-1 (Kumar et al, 2018)	AG	-	EN, HI	15,000	30
GermEval 2018 task 2 (Wiegand et al, 2018b)	OF	-	DE	8,541	20
OffensEval (Zampieri et al, 2019)	OF	-	EN	14,100	115

Issues: definition

> The definition of hate speech is responsibility of the judge, to the linguist
>
> – T. Caselli

What has Legal Informatics to say about HS?

Issues: agreement

Low agreement on the definition of HS leads to low inter-annotator agreement

→ low quality data

Crowdsourcing is hardly an option

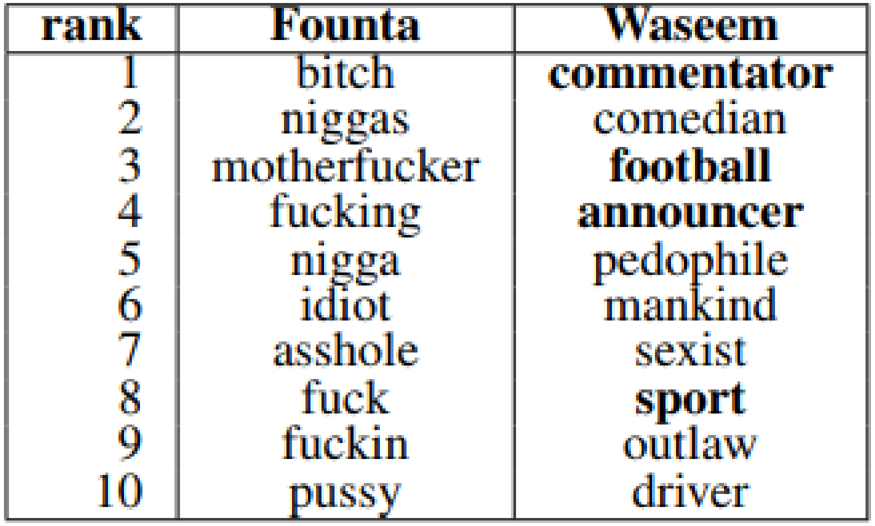
Issues: data bias

Detection of Abusive Language: the Problem of Biased Datasets

Michael Wiegand

et al.

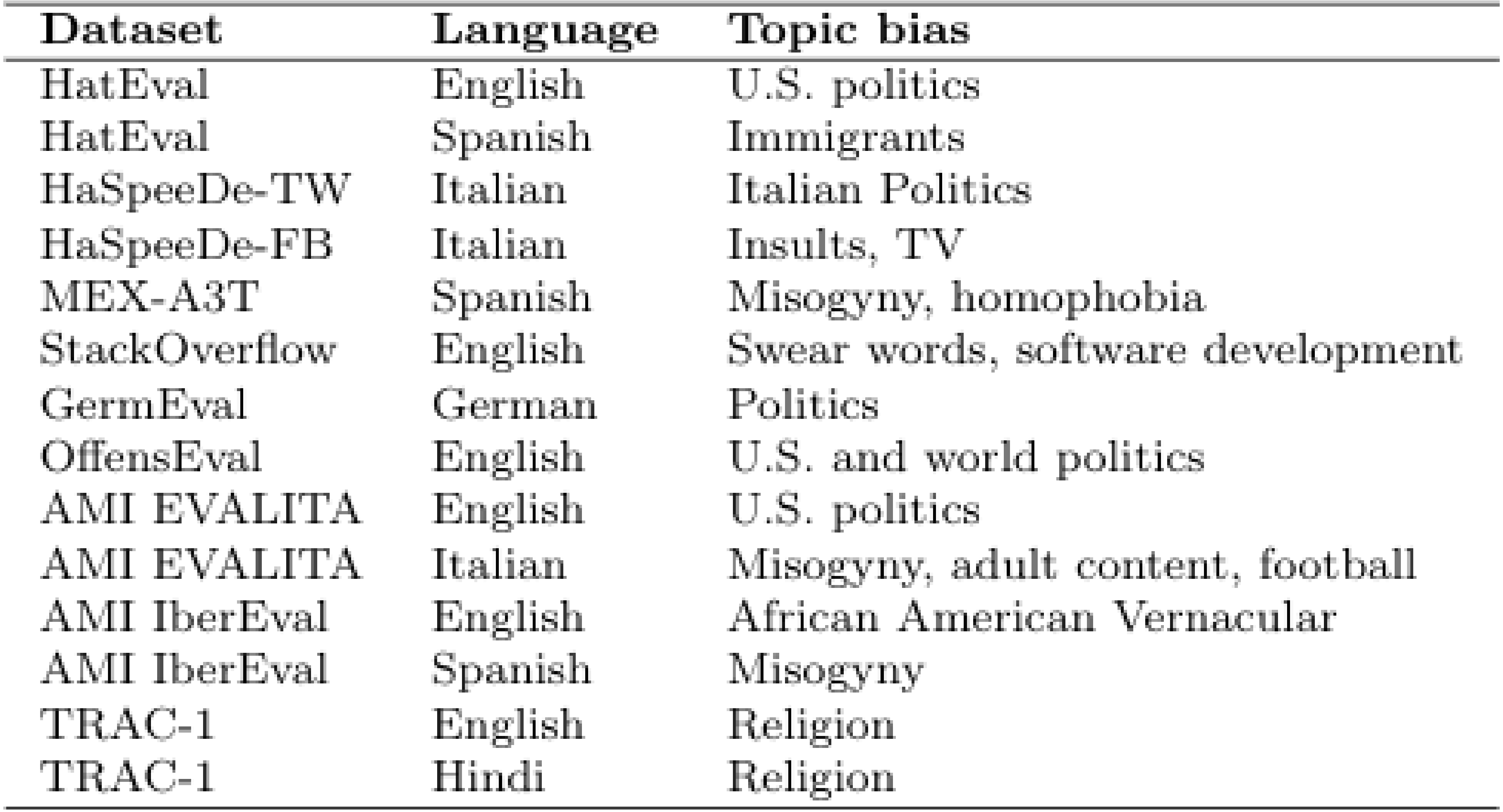
NAACL-HLT 2019



rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

Table 2: Top 10 words having strongest correlation with abusive microposts according to PMI on *Founta* (dataset representing almost random sample) and *Waseem* (dataset produced by biased sampling).

Issues: data bias



Dataset	Language	Topic bias
HatEval	English	U.S. politics
HatEval	Spanish	Immigrants
HaSpeeDe-TW	Italian	Italian Politics
HaSpeeDe-FB	Italian	Insults, TV
MEX-A3T	Spanish	Misogyny, homophobia
StackOverflow	English	Swear words, software development
GermEval	German	Politics
OffensEval	English	U.S. and world politics
AMI EVALITA	English	U.S. politics
AMI EVALITA	Italian	Misogyny, adult content, football
AMI IberEval	English	African American Vernacular
AMI IberEval	Spanish	Misogyny
TRAC-1	English	Religion
TRAC-1	Hindi	Religion

Issues: implicit vs. explicit

Not all HS is expressed in a lexically explicit way.

Implication, world knowledge, rhetorical expressions...

> *@USER @USER @USER Have you ever seen ANTIFA burning college campuses and trashing them any time a conservative comes to speak ? Educate yourself please !*

Issues: implicit vs. explicit

One major distinction that has been proposed in the literature is the division into explicitly and implicitly abusive language (Waseem et al., 2017).
The former are microposts that employ some abusive words, while the latter represents the more difficult case in which the abusive nature is conveyed by other means, such as sarcasm, jokes, and particularly the usage of negative stereotypes, e.g.:

- i havent had an intelligent conversation with a woman.
- Jews don’t marry children. Muslims do. All the time.

(Also from Wiegand et al. 2019)

Where to go now?

Are we hitting the plateau of NLP performance on HS detection?

Where to go now?

We are able from extract almost all the information present in the text.

Hence, what is missing is the output IS NOT in the language.

→ link to Ontologies, Knowledge Graphs, …

Where to go now?

Are we benchmarking correctly?

Where to go now?

In creating gold standard data, we assume that there is ONE ground truth.

Perhaps it is time for the annotators’ background to be part of the equation.