Hate Speech:
Misogyny, stereotypes, irony/sarcasm, hurtful humour

Paolo Rosso

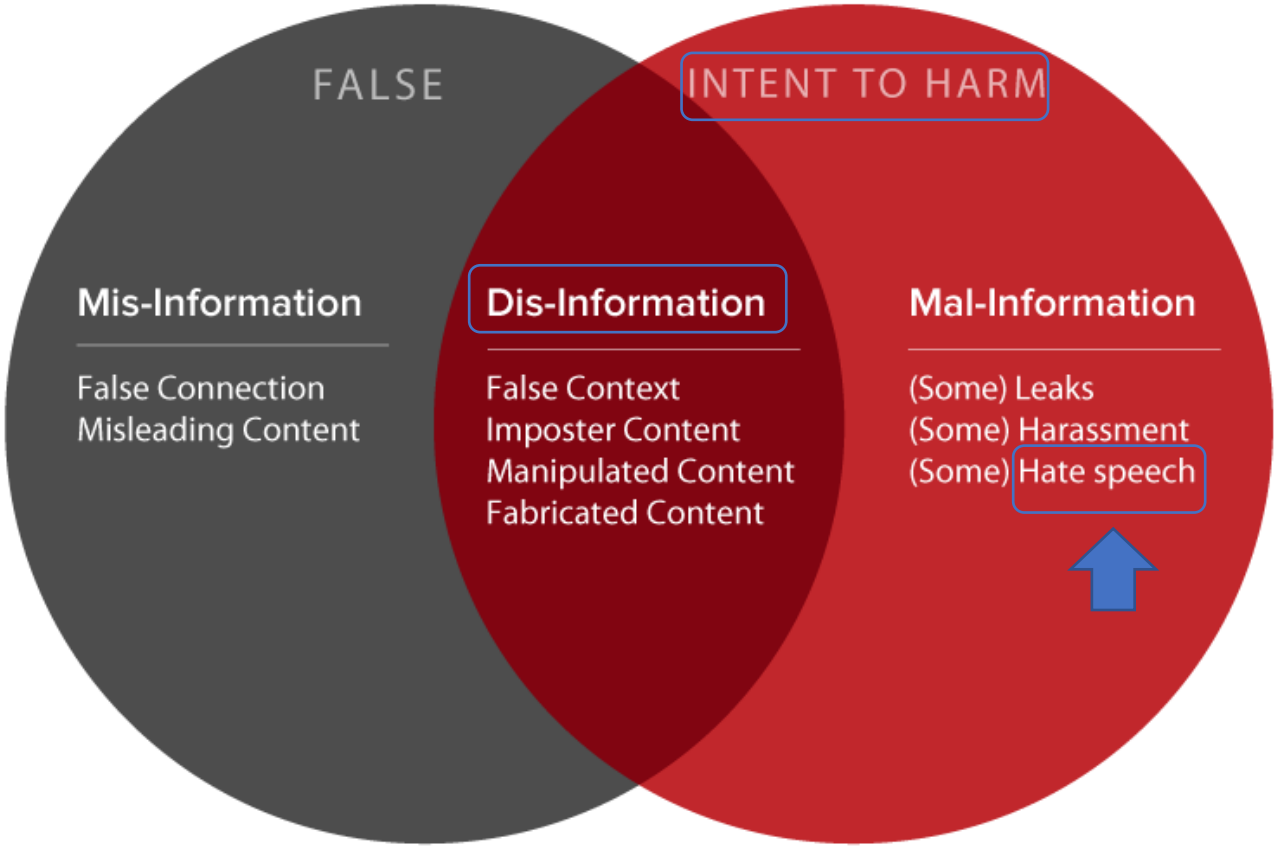
2023-2024

DSIC

MIARFID
Official Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Information disorder online: Harmful information



C. Wardle, H. Derakhshan. One year on, we are still not recognizing the complexity of **information disorder online**.

https://firstdraftnews.org/latest/coe_infodisorder/

Hate Speech

- **Related concepts and shared tasks**
- Misogyny identification
- Spanish observatory on racism and xenophobia
- Implicit HS: stereotypes, irony/sarcasm, hurtful humour
- HS in memes
- Strategies against HS

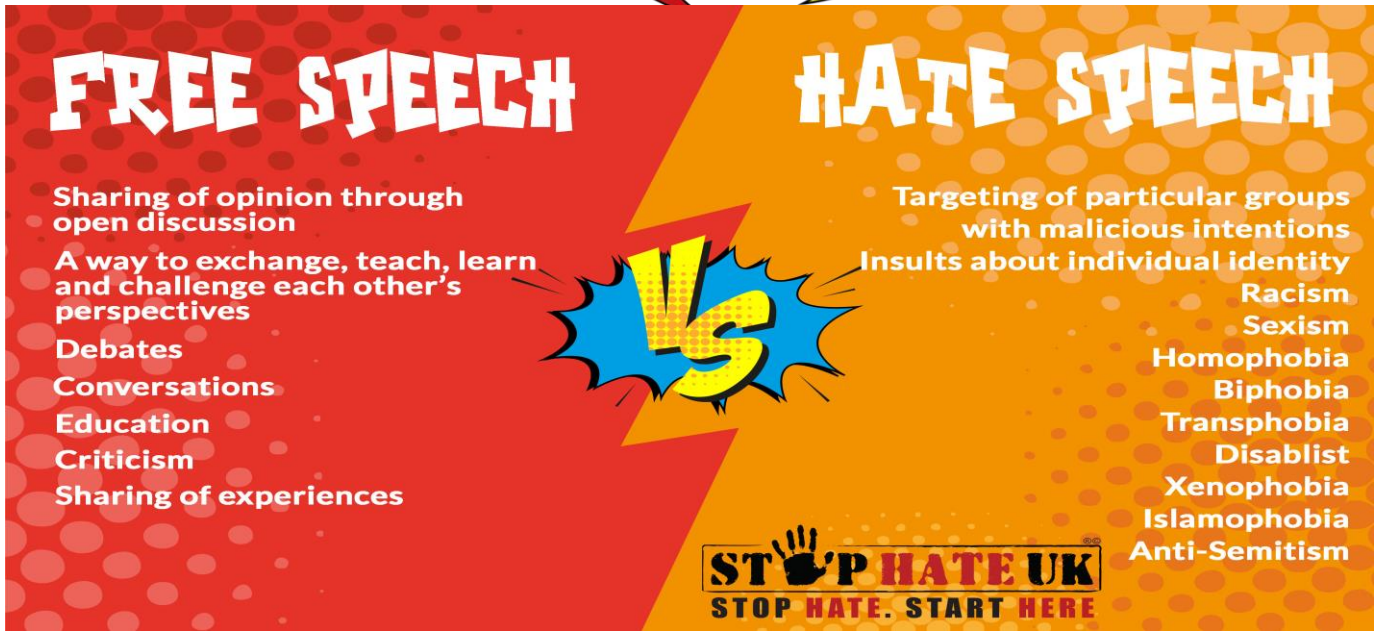
Hate speech

- Hate speech (HS) is commonly defined as any communication that disparages **a person or a group** on the basis of some characteristics such as **race, color, ethnicity, gender, sexual orientation, nationality, religion**, or other.

- Expressions that: (i) incite **discrimination or violence** due to racial hatred, xenophobia, sexual orientation and other types of intolerance; (ii) **foster hostility** through **prejudice and intolerance**.

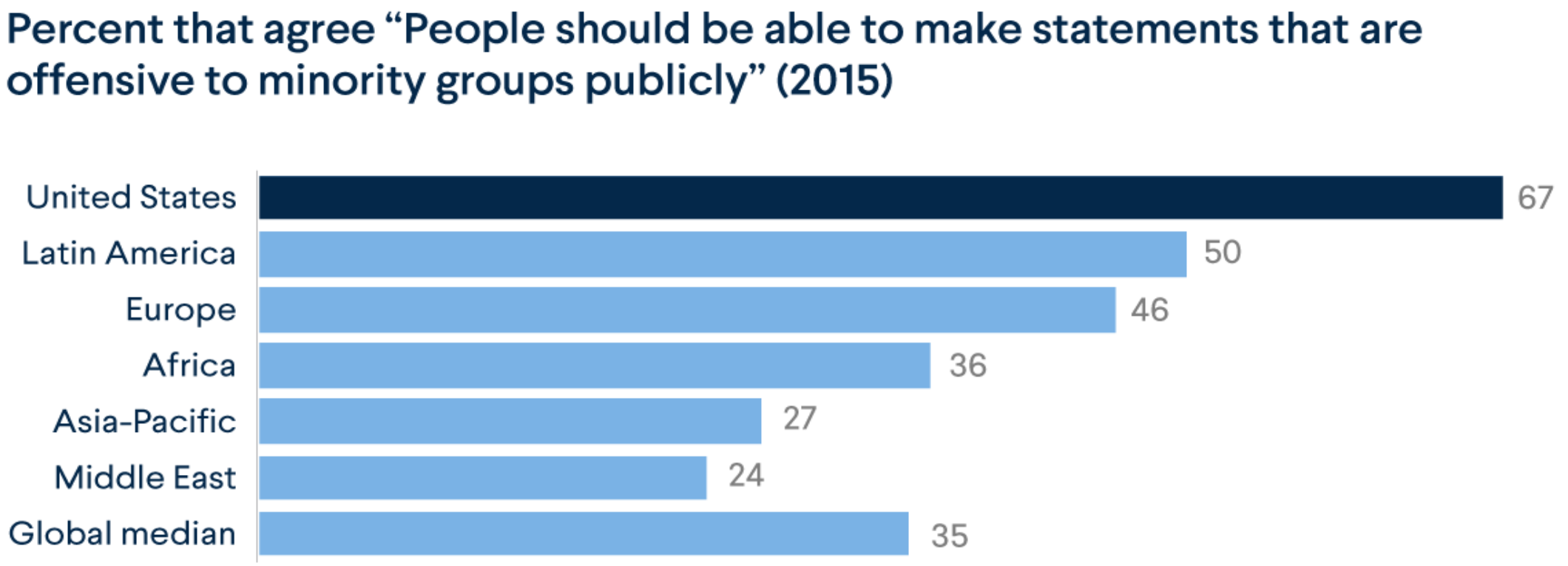
J. T. Nockleby (2000). **Hate speech**. Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan), pp. 1277–1279

Freedom of expression vs suppression of HS



Tolerance vs intransigence

Free vs Hate Speech



Note: Displays the median among countries included in the survey.

Source: Pew Research Center.

COUNCIL on FOREIGN RELATIONS

Online hate speech

- Decentralised communication
- Massive scale
- **Multiplication potential**
- Widespread use of pseudonyms and **anonymity**
- **Virtuality helps people lose their inhibitions**
- **Unlimited time content until removal**

Inflammatory content

Related concepts

Abusive language detection:

- Covers all the **hurtful language**
- Includes **hate speech**
- Many researches refer it to as **offensive language**

Cyberbullying detection:

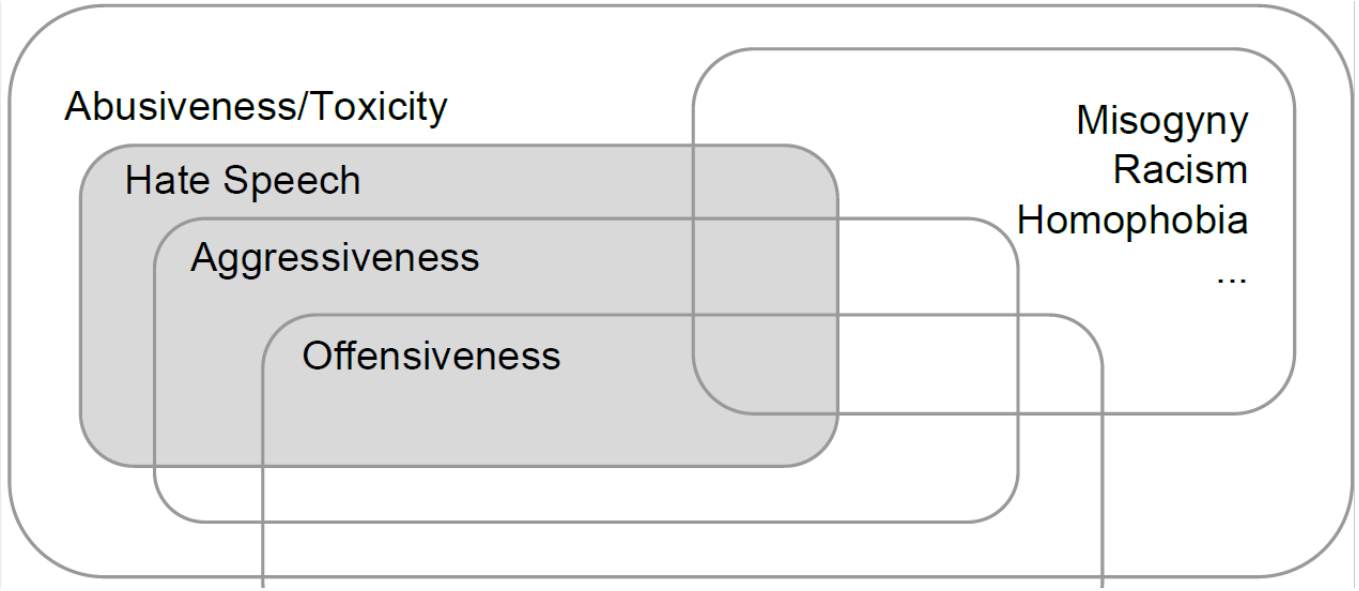
- The online form of traditional bullying
- **Harassment**: intent to harm an **individual** (target) who is unable to defend herself

Radicalization detection:

- Motivates **violent extremism**

Related concepts

One of the major issues consists in the intrinsic complexity in defining HS and in a widespread vagueness in the use of related terms (such as **abusive, toxic, dangerous, offensive or aggressive language**), that often **overlap** and are prone to strongly **subjective interpretations**



F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti (2021). **Resources and benchmark corpora for hate speech detection: a systematic review**. Language Resources & Evaluation, 55: 477–523

Shared tasks

- Identification of offensive language @ GermEval 2018, … 
- HS detection @ EVALITA 2018, 2020 
- HS and offensive content identification @ FIRE 2019, … 
- OffensEval @ SemEval 2019, … 
- HatEval @ SemEval 2019 
- DETOXIS @ IberLEF 2021, DETESTS @ IberLEF 2022, 2024 
- Profiling HATERS @ PAN 2021 
- AMI @ IberEval 2018, EVALITA 2018, 2020 
- MAMI @ SemEval 2022 

Repository of regionalized, multilingual HS

HATEBASE

3,725	569,646	97	178
TERMS	SIGHTINGS	LANGUAGES	COUNTRIES

All terms

All ethnicities

All religions

All languages

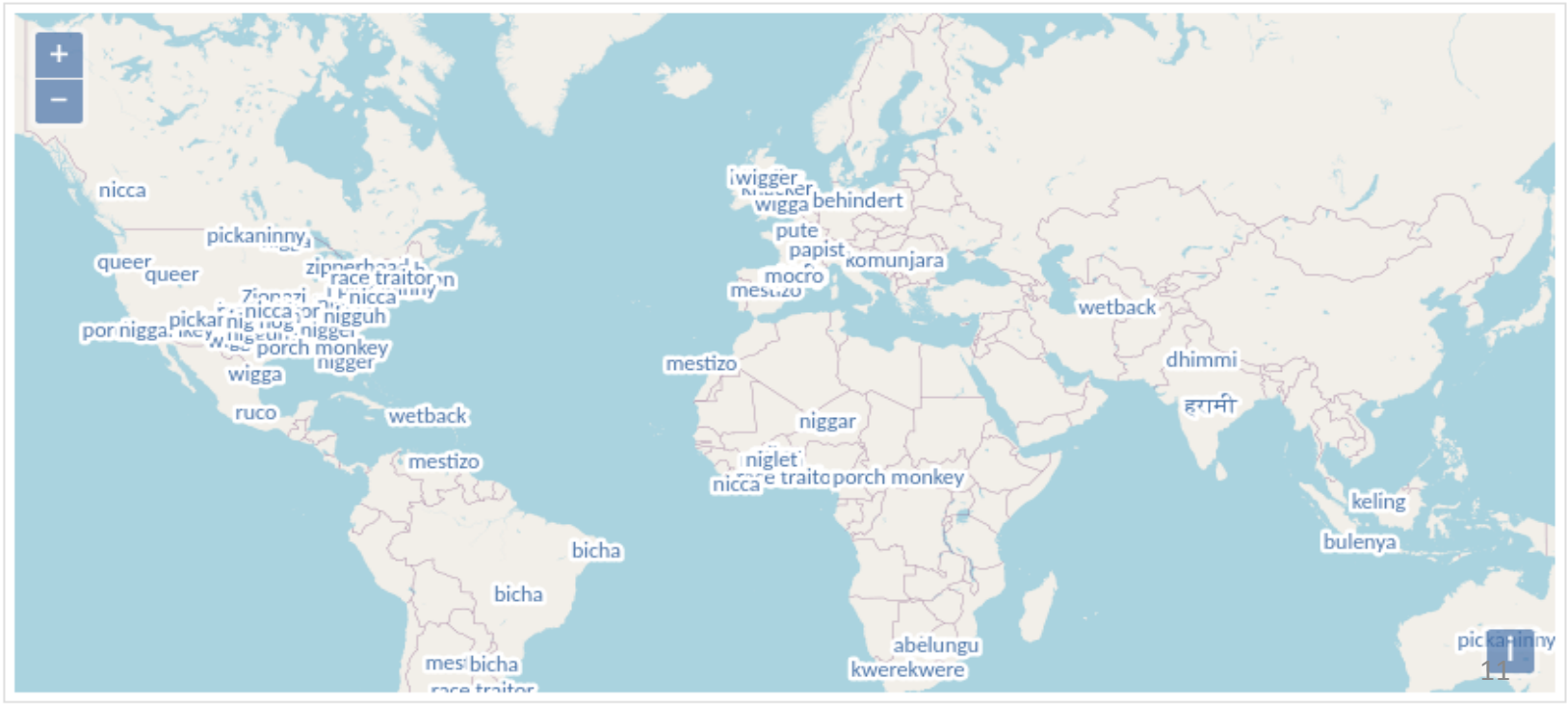
- [ ] Pertains to gender
- [ ] Pertains to sexual orientation
- [ ] Pertains to disability
- [ ] Pertains to class

Search Clear

https://hatebase.org/

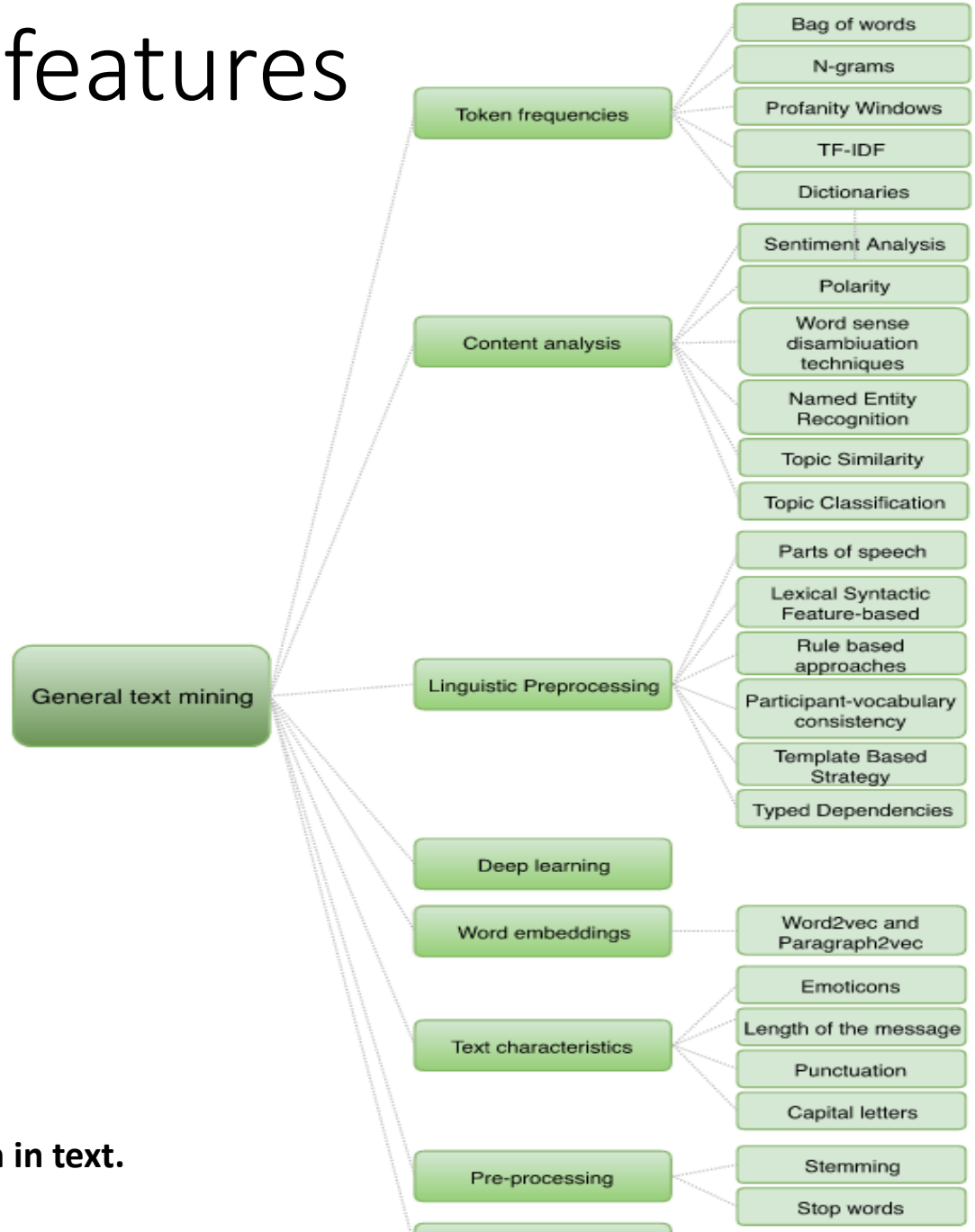
Recent Sightings

MAP TABLE



Generic approaches and features

- Supervised methods
- Semi-supervised methods
- Unsupervised methods

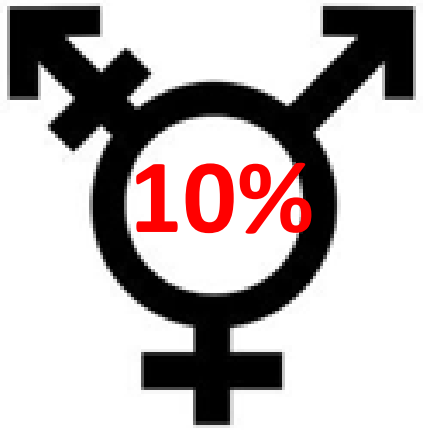


P. Fortuna, S. Nunes (2018). **A survey on automatic detection of hate speech in text.** ACM Computing Surveys (CSUR) 51.4

Hate Speech in Twitter

Women

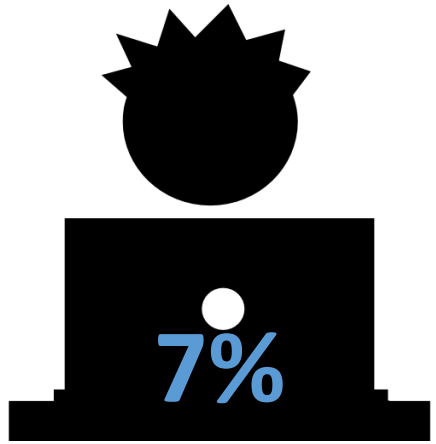
Migrants



LGBT

ITALY 2015-2016

2,6 million of tweet



Disabled

https://www.repubblica.it/tecnologia/social-network/2016/07/13/news/l_odio_corre_su_twitter_nel_mirino_soprattutto_le_donne-143983944/

hatEval @ SemEval 2019

Multilingual detection of HS against **immigrants** and **women** in Twitter

- Task A: Multilingual detection of hate
- Task B: **Aggressive** behaviour and **target** classification

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel., P. Rosso, M. Sanguinetti (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. Proc. SemEval 2019

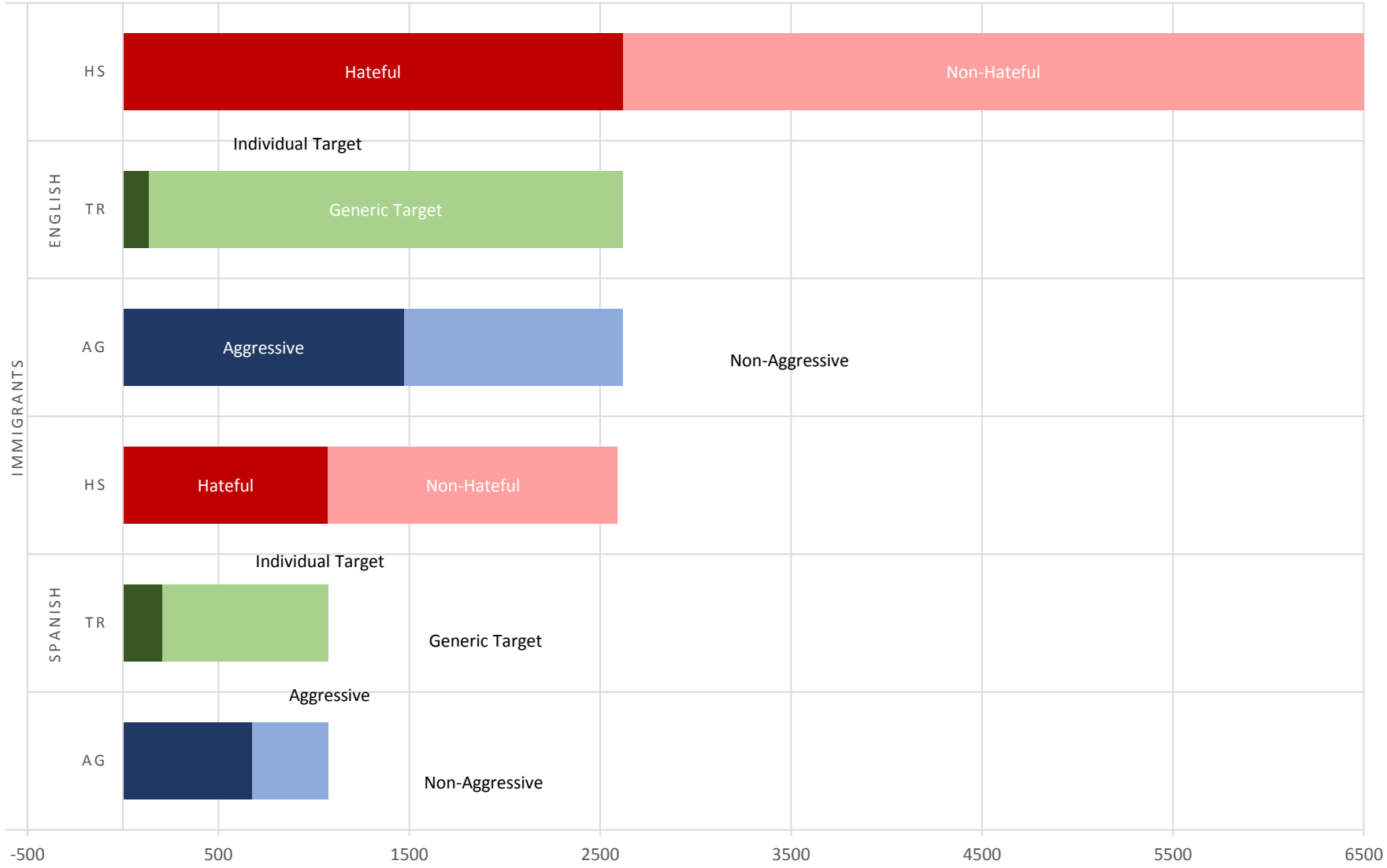
Dataset

- **Keyword-driven approach**
  - neutral keywords
  - **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 + data from AMI tasks
- Statistics: 19,600 tweets (**13,000 EN; 6,600 ES**)
- Target: **immigrants**: 9k; **women**: 10k approx.

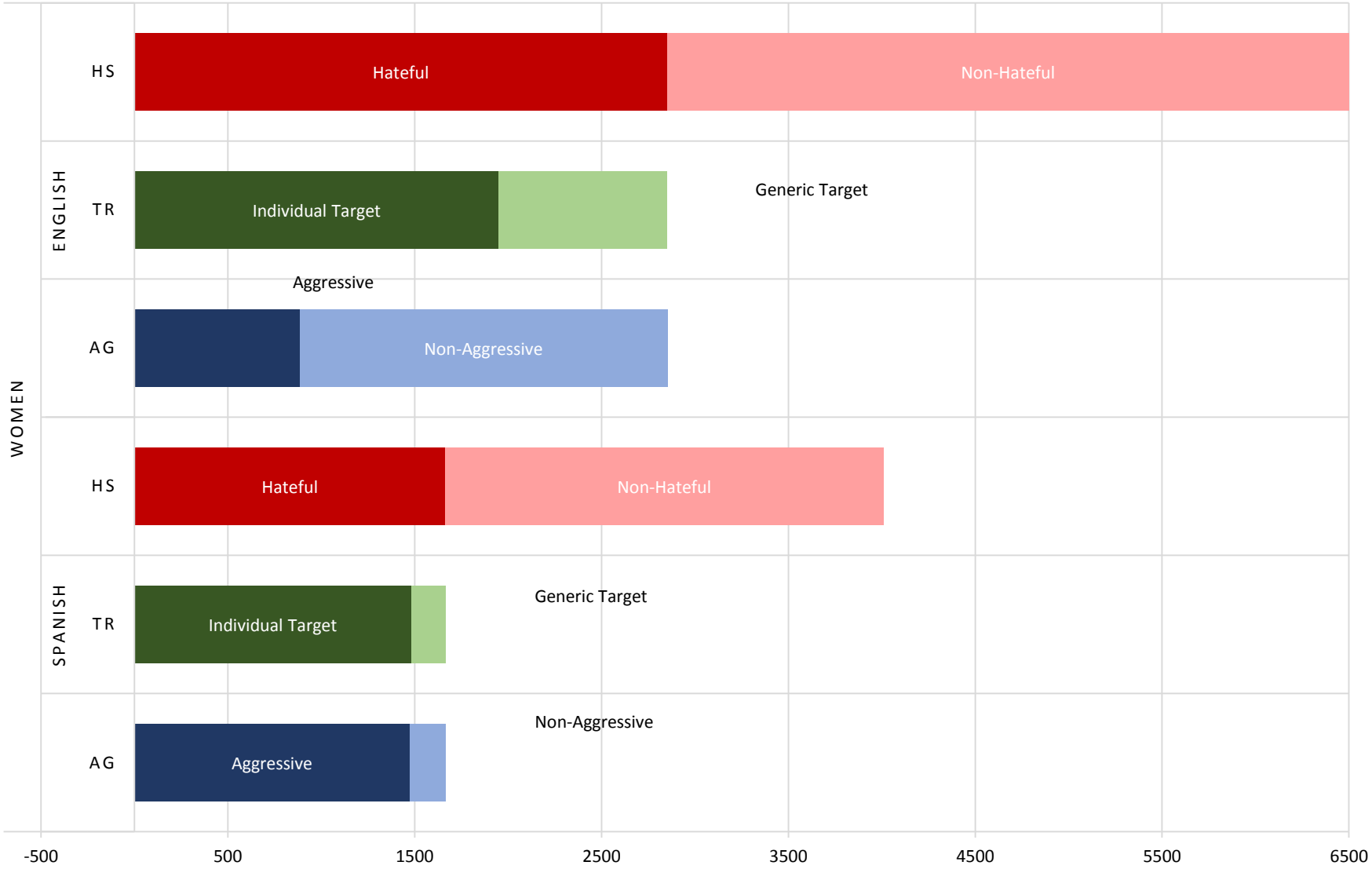
Annotation

- **Crowdsourcing**
- Guidelines in English and Spanish
  - Definition for hate speech against the two targets
  - Definition of aggressiveness
  - List of examples
- Two additional expert annotators
- **HS distribution is over-represented**
- **AG and TR distributions are natural**

Dataset: Immigrants



Dataset: Women



Evaluation

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

Techniques

- Approaches
  - Deep learning (more than half): RNN in particular
- Features
  - Word embeddings: mostly GloVe
  - Custom hate lexicons
- Preprocessing
  - Mostly standard
  - Twitter-driven: hashtag segmentation, slang conversion, emoji translation

Results

74 teams

- Task A (Multilingual detection of hate): 108 runs
- Task B (Aggressive behaviour and target classification): 70 runs

Approaches:

- **Task A** (accuracy). **EN**: SVM + sentence embeddings from Google's Universal Sentence Encoder (0.65); CNN, LSTM; **ES**: SVM (0.73); BERT
- **Task B**. **EN**: MFC baseline; SVM (best); LR, LSTM; **ES**: LR (best); SVM

@ SemEval 2019 & 2020

SUBTASK A

1. Offensive language identification
Offensive / Not offensive

SUBTASK B

2. Automatic categorization of offense types
Targeted insult / Untargeted

SUBTASK C

3. Offense target identification
Individual/Group/Other

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar (2019). **Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval).** Proc. SemEval 2019
M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, C. Çöltekin (2020). **SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020).** Proc. SemEval 2020

Evaluation

Macro-averaged F1-score

2019. Best system: BERT

Sub-task A		Sub-task B		Sub-task C	
Team Ranks	F1 Range	Team Ranks	F1 Range	Team Ranks	F1 Range
1	0.829	1	0.755	1	0.660
2	0.815	2	0.739	2	0.628
3	0.814	3	0.719	3	0.626
4	0.808	4	0.716	4	0.621
5	0.807	5	0.708	5	0.613
6	0.806	6	0.706	6	0.613
7	0.804	7	0.700	7	0.591
8	0.803	8	0.695	8	0.588
9	0.802	9	0.692	9	0.587
CNN	**0.800**	**CNN**	**0.690**	10	0.586
10	0.798	10	0.687	11-14	.571-.580
11-12	.793-.794	11-14	.680-.682	15-18	.560-.569
13-23	.782-.789	15-24	.660-.671	19-23	.547-.557
24-27	.772-.779	**BiLSTM**	**0.660**	24-29	.523-.535
28-31	.765-.768	25-29	.640-.655	30-33	.511-.515
32-40	.750-.759	**SVM**	**0.640**	34-40	.500-.509
BiLSTM	**0.750**	30-38	.600-.638	41-47	.480-.490
41-45	.740-.749	39-49	.553-.595	**CNN**	**0.470**
46-57	.730-.739	50-62	.500-.546	**BiLSTM**	**0.470**
58-63	.721-.729	**ALL TIN**	**0.470**	**SVM**	**0.450**
64-71	.713-.719	63-74	.418-.486	46-60	.401-.476
72-74	.704-.709	75	0.270	61-65	.249-.340
SVM	**0.690**	76	0.121	**All IND**	**0.210**
75-89	.619-.699	**All UNT**	**0.100**	**All GRP**	**0.180**
90-96	.500-.590			**ALL OTH**	**0.090**
97-103	.422-.492				
All NOT	**0.420**				
All OFF	**0.220**				
104	0.171				

BERT

2020. Subtask A (English)

Best system: Ensemble of Tranformer models

#	Team	Score	#	Team	Score	#	Team	Score
1	UHH-LT	0.9204	29	UTFPR	0.9094	57	OffensSzeged	0.9032
2	Galileo	0.9198	30	IU-UM@LING	0.9094	58	FBK-DH	0.9032
3	Rouges	0.9187	31	TAC	0.9093	59	RGCL	0.9006
4	GUIR	0.9166	32	SSN_NLP_MLRG	0.9092	60	byteam	0.8994
5	KS@LTH	0.9162	33	Hitachi	0.9091	61	ANDES	0.8990
6	kungfupanda	0.9151	34	CoLi @ UdS	0.9091	62	PUM	0.8973
7	TysonYU	0.9146	35	XD	0.9090	63	NUIG	0.8927
8	AlexU-BackTranslation-TL	0.9139	36	UoB	0.9090	64	I2C	0.8919
9	SpurthiAH	0.9136	37	PAI-NLP	0.9089	65	sonal.kumari	0.8900
10	amsqr	0.9135	38	PingANPAI	0.9089	66	IJS	0.8887
11	m20170548	0.9134	39	VerifiedXiaoPAI	0.9089	67	IR3218-UI	0.8843
12	Coffee_Latte	0.9132	40	nlpUP	0.9089	68	TeamKGP	0.8822
13	wac81	0.9129	41	NLP_Passau	0.9088	69	UNT Linguistics	0.8820
14	NLPDove	0.9129	42	TheNorth	0.9087	70	janecek1	0.8744
15	UJNLP	0.9128	43	problemConquero	0.9085	71	Team Oulu	0.8655
16	ARA	0.9119	44	Lee	0.9084	72	TECHSSN	0.8655
17	Ferryman	0.9115	45	Wu427	0.9081	73	KDELAB	0.8653
18	ALT	0.9114	46	ITNLP	0.9081	74	HateLab	0.8617
19	SINAI	0.9105	47	Better Place	0.9077	75	IASBS	0.8577
20	MindCoders	0.9105	48	IIITG-ADBU	0.9075	76	IUST	0.8288
21	IRLab_DAIICT	0.9104	49	doxaAI	0.9075	77	Duluth	0.7714
22	erfan	0.9103	50	NTU_NLP	0.9067	78	RTNLU	0.7665
23	Light	0.9103	51	FERMI	0.9065	79	KarthikaS	0.6351
24	KAFK	0.9099	52	AdelaideCyC	0.9063	80	Bodensee	0.4954
25	PALI	0.9098	53	INGEOTEC	0.9061		**Majority Baseline**	0.4193
26	PRHLT-UPV	0.9097	54	PGSG	0.9060	81	IRlab@IITV	0.0728
27	YNU_oxz	0.9097	55	SRIB2020	0.9048			
28	IITP-AINLPML	0.9094	56	GruPaTo	0.9036			

Best system

Model: Masked Language Modeling RoBERTa-large ensemble

Fine-tuned 10 times with OffensEval 2019 dataset

Prediction: Majority vote on the 10 predictions

Other evaluations (fine-tuning with OffensEval 2019):

- BERT_base
- BERT_large
- RoBERTa-base
- RoBERTa-large
- XLM-RoBERTa
- ALBERT_large-v1
- ALBERT_large-v2
- ALBERT_xxlarge-v1
- ALBERT_xxlarge-v2

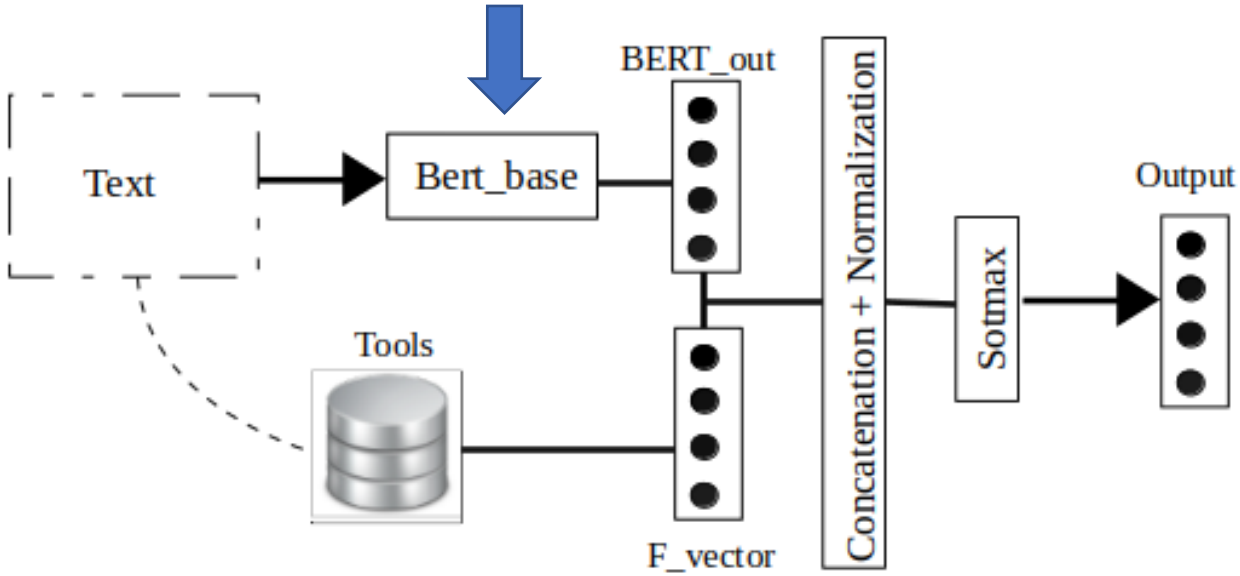
Results (10-fold cross-validation):

	NOT			OFF				
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro F1**	**Acc.**
RoBERTa-large	98.96	91.49	95.07	81.54	97.49	88.76	91.93	93.15
RoBERTa-large MLM-ft	**99.15**	**91.53**	**95.18**	**81.66**	**97.96**	**89.06**	**92.12**	**93.31**

	NOT			OFF				
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro F1**	**Acc.**
Baselines								
All NOT	72.21	100.00	41.93	-	0.00	0.00	41.93	72.21
All OFF	-	0.00	0.00	27.78	100.00	43.49	21.74	27.79
Single pre-trained transformer models								
BERT-base	99.06	90.2	94.42	79.34	97.78	87.60	91.01	92.31
BERT-large	99.65	90.35	94.77	79.81	99.17	88.44	91.60	92.80
RoBERTa-base	99.45	90.70	94.88	80.33	98.70	88.57	91.73	92.93
RoBERTa-large	99.53	90.92	95.03	80.73	98.89	88.89	91.96	93.13
XLM-RoBERTa	99.03	91.31	95.01	81.22	97.69	88.69	91.85	93.08
ALBERT-large-v1	98.87	90.24	94.36	79.32	97.31	87.40	90.88	92.20
ALBERT-large-v2	98.87	90.20	94.34	79.26	97.31	87.36	90.85	92.18
ALBERT-xxlarge-v1	98.35	91.09	94.58	80.57	96.02	87.62	91.10	92.46
ALBERT-xxlarge-v2	98.47	91.73	94.98	81.76	96.30	88.44	91.71	93.00
Ensembles of pre-trained transformer models								
All	**99.65**	90.95	95.10	80.83	**99.17**	89.06	92.08	93.23
BERT	99.42	91.16	95.11	81.11	98.61	89.01	92.06	93.23
RoBERTa	99.57	90.84	95.01	80.62	98.98	88.86	91.93	93.11
ALBERT-all	98.23	**92.66**	**95.36**	**83.37**	95.65	89.00	92.23	**93.49**
ALBERT-xxlarge	98.70	92.16	95.32	82.62	96.85	**89.17**	**92.25**	93.47

G. Wiedemann, S.M. Yimam, C. Biemann (2020). **UHH-LT & LT2 at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection.** arXiv preprint arXiv:2004.11493

Our proposal



Feature vector (F_vector)

- Basic features:
  - length of the tweets
  - number of misspelled words
  - use of punctuation marks
- Semantic features:
  - use of emoticons
  - use of noun phrases

Model	Subtask A - Macro F1
Best system	0.9204
Proposal	0.9097
Average	0.8709
Last system	0.0728

G. De La Peña, P. Rosso (2020). **PRHLT-UPV at SemEval-2020 task 12: BERT for multilingual offensive language detection**. Proc. SemEval 2020

Hate Speech

- Related concepts and shared tasks
- **Misogyny identification**
- Spanish observatory on racism and xenophobia
- Implicit HS: stereotypes, irony/sarcasm, hurtful humour
- HS in memes
- Strategies against HS

AMI: Automatic Misogyny Identification

AMI shared task @ IberEval 2018

AMI shared task @ Evalita 2018

1. **Automatic Misogyny Identification**
   - *Misogynous* vs *Not Misogynous*

SUBTASK A

2. **Misogynistic Misbehavior Classification**
   - *Derailing*
   - *Dominance*
   - *Discredit*
   - *Stereotype & Objectification*
   - *Sexual Harassment & Threats of Violence*
3. **Target Classification**
   - *Individual* vs *Generic*

SUBTASK B

E. Fersini, P. Rosso, M. Anzovino (2018). Overview of the task on automatic misogyny identification at IberEval 2018. Proc. IberEval 2018

E. Fersini, D. Nozza, P. Rosso (2018). Overview of the Evalita 2018 task on automatic misogyny identification (AMI). Proc. EVALITA 2018

E. Fersini, D. Nozza, P. Rosso (2018). AMI @ EVALITA2020: Automatic Misogyny Identification. Proc. EVALITA 20

Misogyny

MISOGYNY is defined as the **hate** or **prejudice against women**, can be linguistically manifested in numerous ways, denoting social exclusion, discrimination, hostility, threats of violence and sexual objectification.

The proposed taxonomy

MISOGYNY

- ***Misogynous:*** a text that represent a form of hate towards women;

  *I've yet to come across a nice girl. They all end up being bit**es in the end*

- ***Not Misogynous:*** a text that does not express hating towards women.

  *Karma is a bi**ch!*

Anzovino M., Fersini E., Rosso P. Automatic identification and classification of misogynistic language on Twitter.
In: Proc. 23rd Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2018, Springer-Verlag, LNCS(10859), pp. 57-64

The proposed taxonomy

MISOGYNY CATEGORY

- ***Discredit:*** slurring over women with no other larger intention.

*@melaniatrump stupid fuc**ing bi**ch*

- ***Stereotype & Objectification:*** a widely held but fixed and oversimplified image or idea of a woman; description of women's physical appeal and/or comparisons to narrow standards;

Women are good only into the kitchen..#makemeasandwich

- ***Dominance***: to assert the superiority of men over women to highlight gender inequality.

Women are inferior to men..so shut up please!

The proposed taxonomy

MISOGYNY CATEGORY

- ***Sexual Harassment & Threats of Violence:*** to describe actions as sexual advances, requests for sexual favors, harassment of a sexual nature; intent to physically assert power over women through threats of violence.

  *Stupid bi**ch I'll put you down on the floor and I'll rape you! You should be scared!*

- ***Derailing****:* to justify woman abuse, rejecting male responsibility; an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men.

  Wearing a tiny skirt is "asking for it". Your teasing a (hard working, taxes paying) dog with a bone. That's cruel. #YesAllMen

The proposed taxonomy

TARGET

- ***Active (Individual):*** the text includes offensive messages purposely sent to a specific target;

@JulieB stupid crazy psychopathic woman..you should die...

- ***Passive (Generic)****:* it refers to messages posted to many potential receivers (e.g groups of women).

Women: just an inferior breed!!!

The proposed taxonomy

AGGRESSIVENESS

- ***Aggressive:*** a message is considered aggressive if it (implicitly or explicitly) presents, incites, threatens, implies, suggests, or alludes to:
  - attitudes, violent actions, hostility or commission of offenses against women;
  - social isolation towards women for physical or psychological characteristics;
  - justification or legitimization aggressive actions against women.

@JulieB I'll torture you until the end of my life! It's a promise! It will hurt..a lot…

- ***Not Aggressive:*** If none of the previous conditions hold.

Datasets

Three approaches were employed to **collect** misogynistic text on Twitter:

- Streaming download using a set of representative **keywords**
- Monitoring of potential **victims' accounts**
- Downloading the history of **misogynist**

DATASETS - AMI SHARED TASKS

AMI@IBEREVAL 2018: English + Spanish

AMI@EVALITA 2018: English + Italian

AMI@EVALITA 2020: Italian

Crowdsourcing platform: **annotation** with 3 annotators (inter-rater annotator agreement)

- **English** dataset: 0.81 "misogynous", 0.45 "misogyny category"
- **Spanish** dataset: 0.84 "misogynous", 0.66 "misogyny category"
- **Italian** dataset: 0.96 "misogynous", 0.68 "misogyny category"

AMI@IberEval 2018

	Training		Testing	
	Spanish	**English**	**Spanish**	**English**
Misogynistic	1649	1568	415	283
Non-misogynistic	1658	1683	416	443
Discredit	978	943	287	123
Sexual Harassment & Threats of Violence	198	410	51	32
Derailing	20	29	6	28
Stereotype & Objectification	151	137	17	72
Dominance	302	49	54	28
Active	1455	942	370	104
Passive	194	626	45	179

Computational results

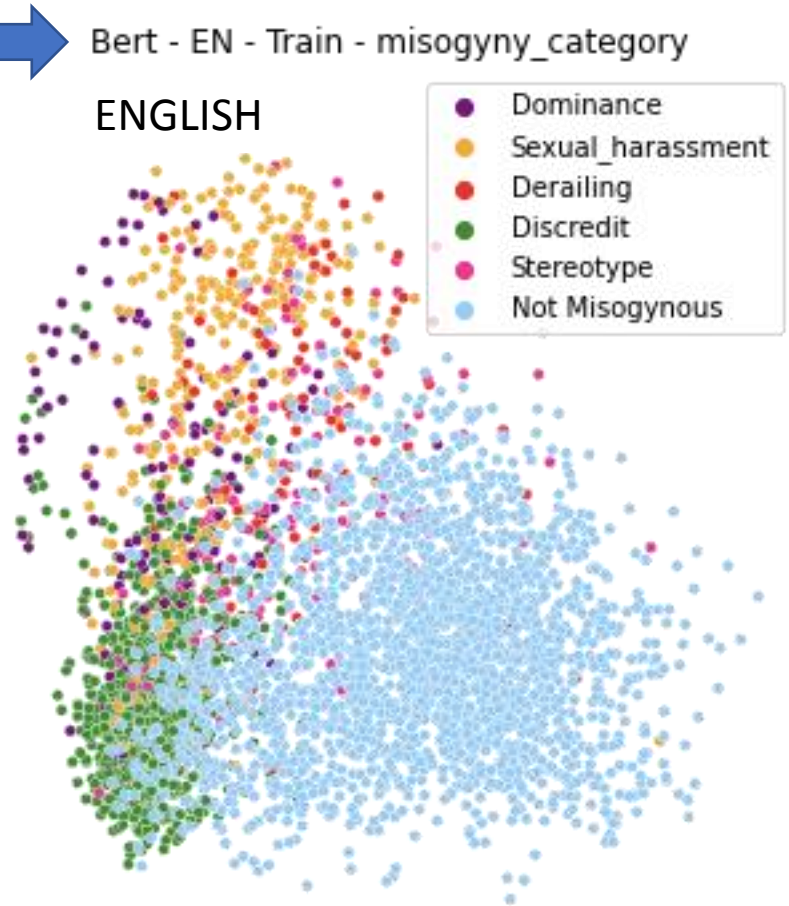
- **MISOGYNY**: 86% English, 85% Spanish, 86% Italian

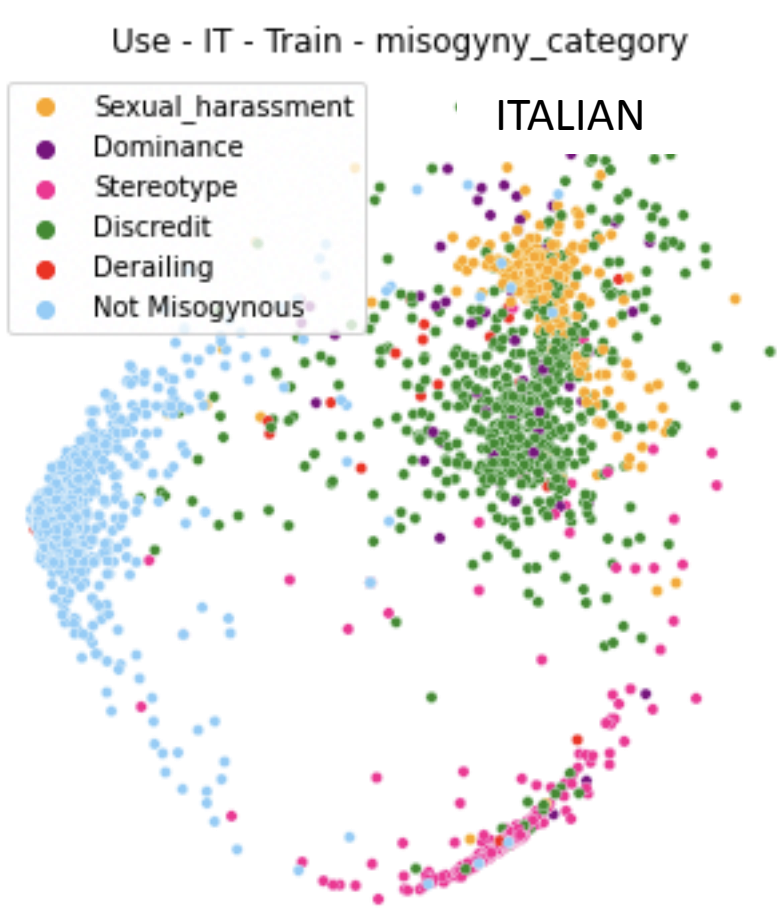
- **MISOGYNISTIC BEHAVIOUR**: 37% English, 34% Spanish, 59% Italian

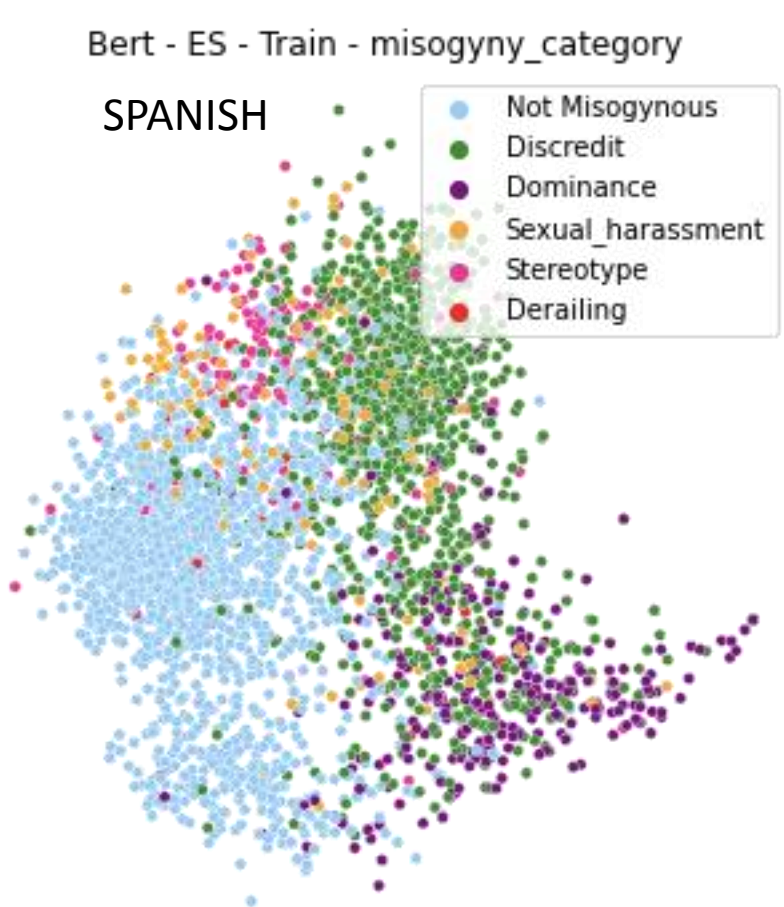
Qualitative interpretation

- How difficult is this task across languages?

MISOGYNY CATEGORY







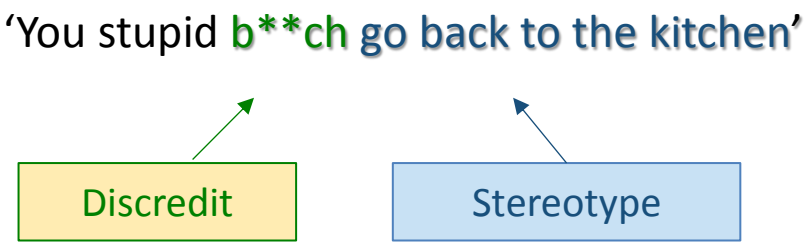
XAI: eXplainable AI

Qualitative interpretation

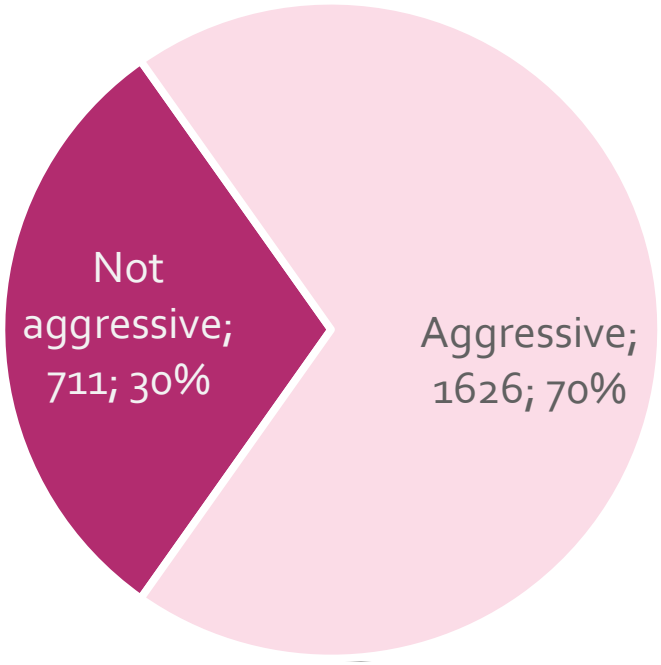
Misogynistic Behaviour

Token n-grams can capture the collocation and *misogynous qualifier* describing the misogynistic behavior.

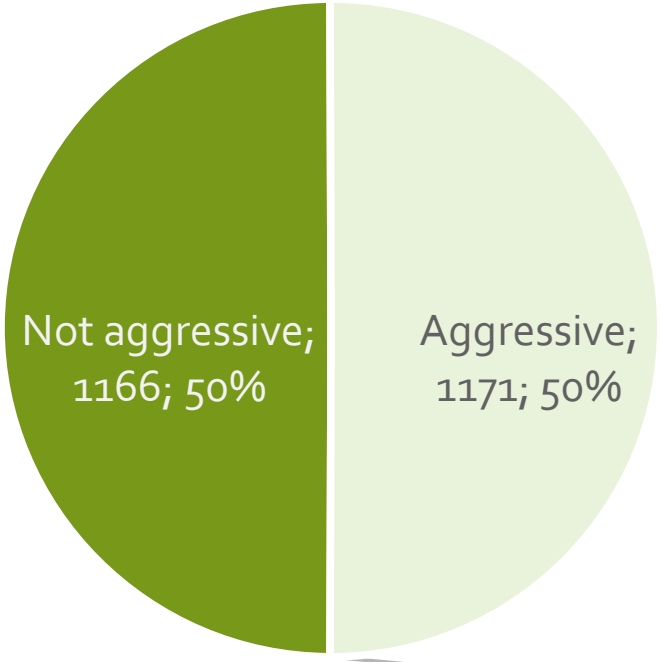
BUT the task is more difficult since the potential **overlapping among categories**:



Misogyny: a gender perspective



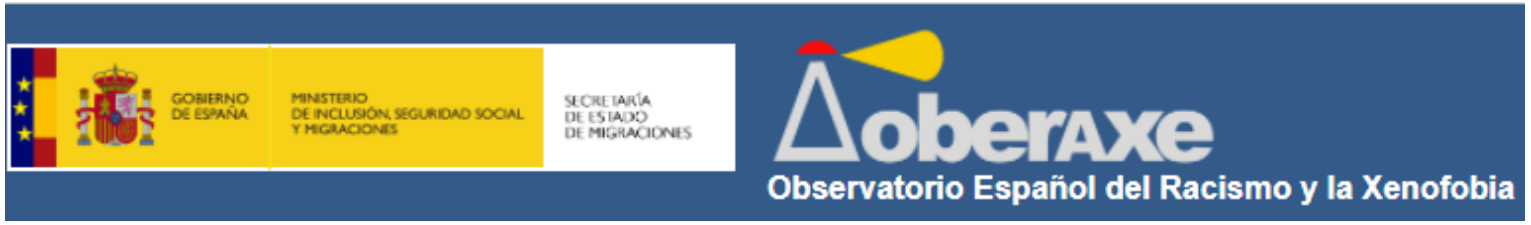
Women



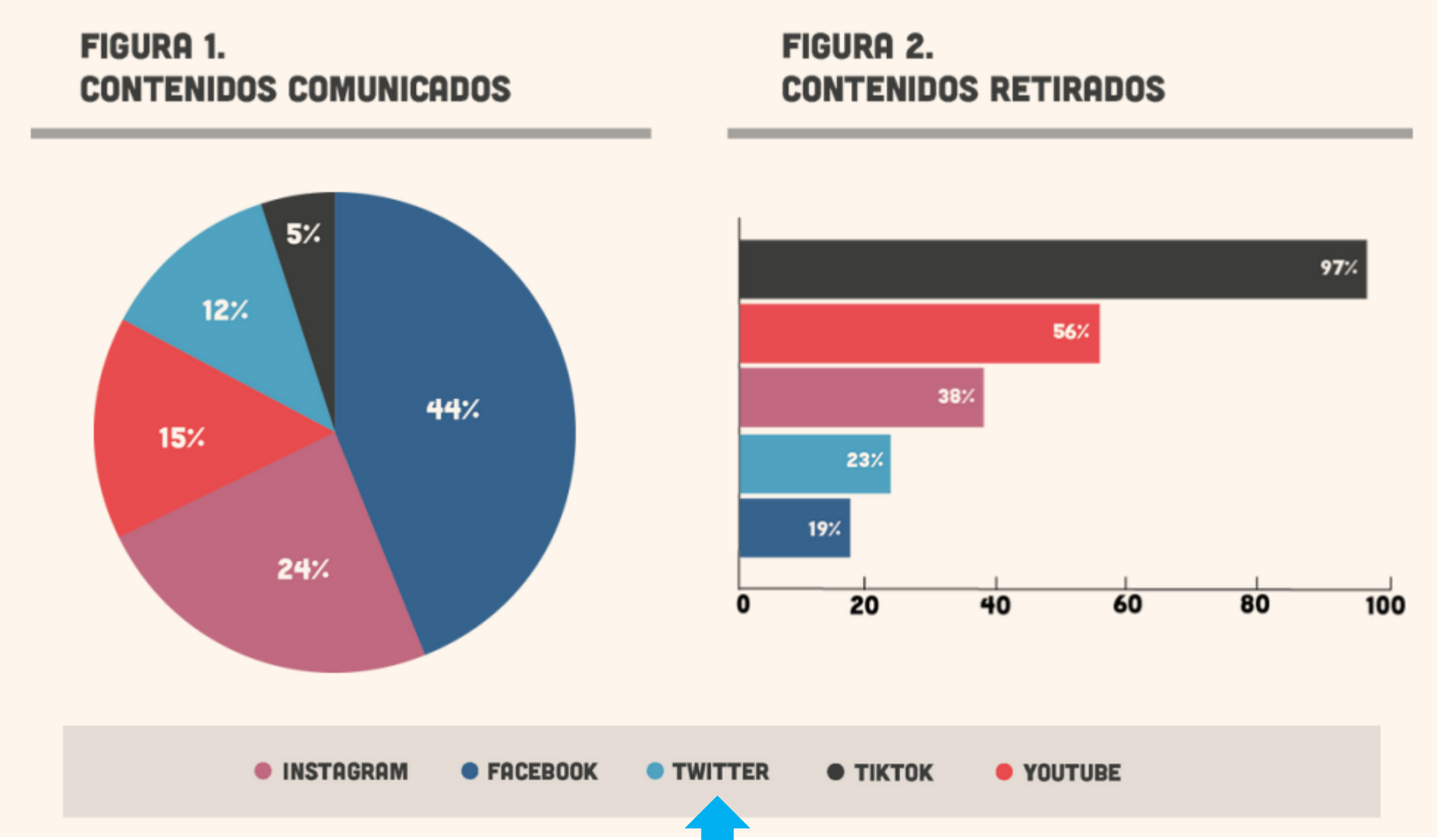
Men

Hate Speech

- Related concepts and shared tasks
- Misogyny identification
- **Spanish observatory on racism and xenophobia**
- Implicit HS: stereotypes, irony/sarcasm, hurtful humour
- HS in memes
- Strategies against HS

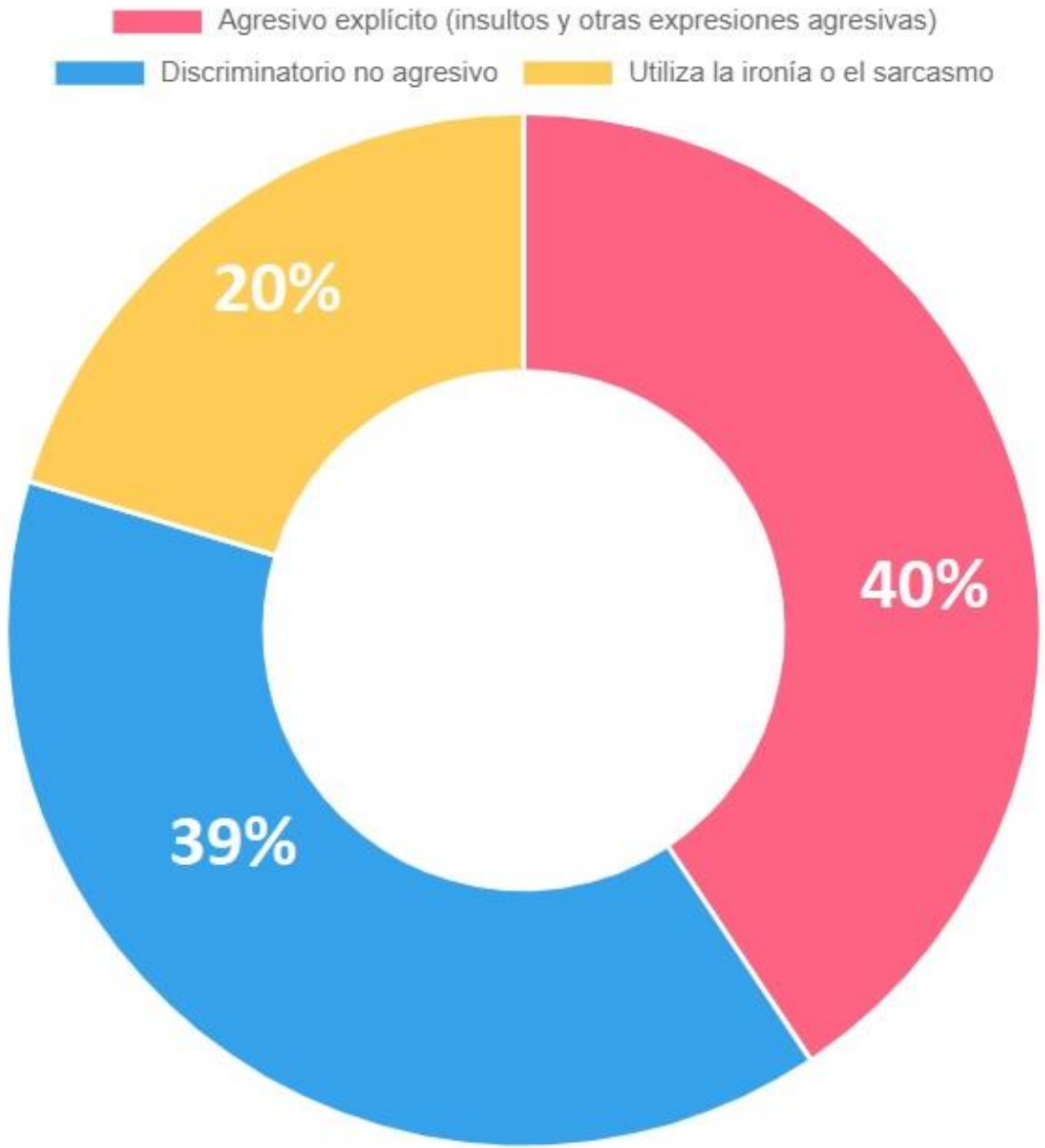


Oberaxe (Sept-Oct 2022): flagged vs removed content

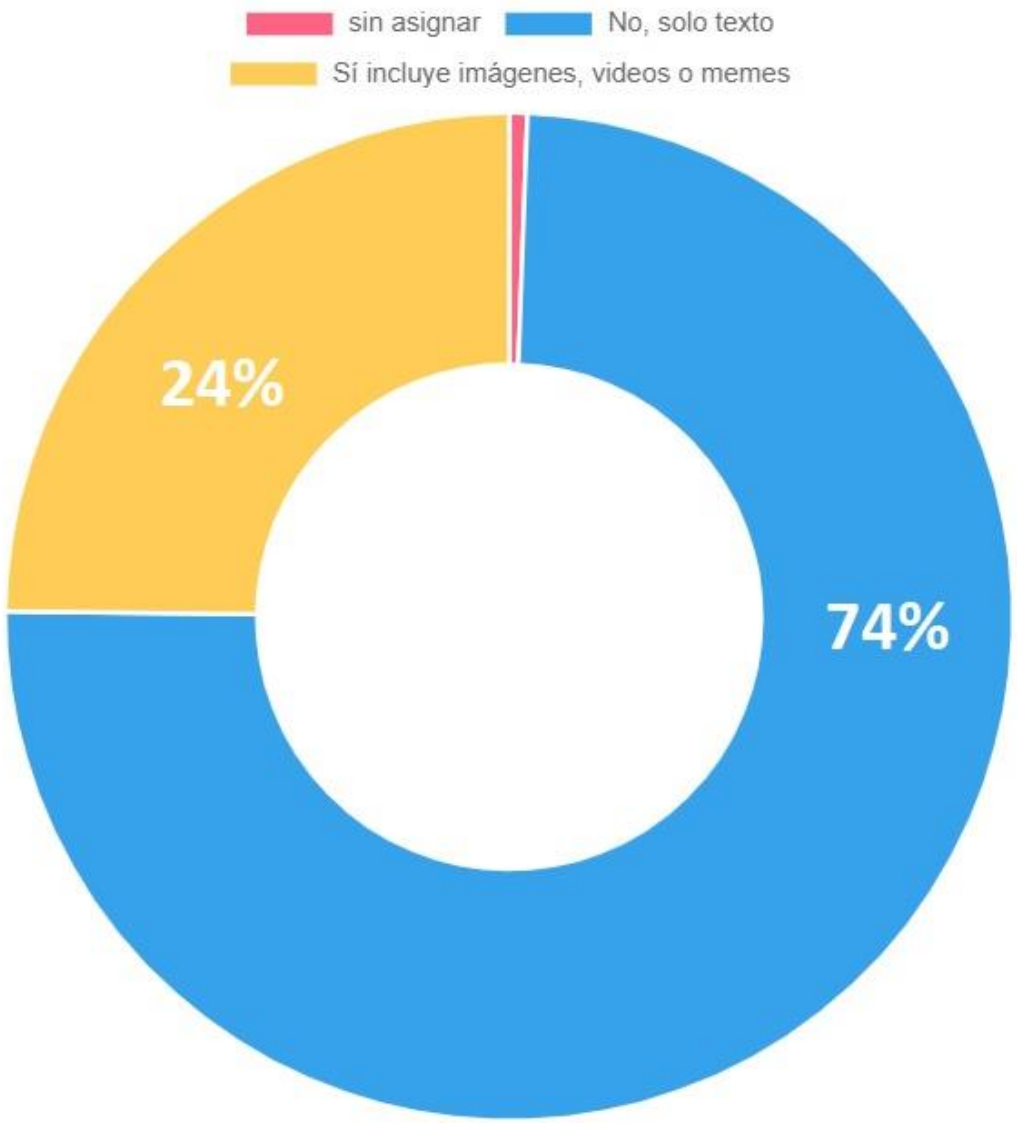


Oberaxe (Sept-Oct 2022): target of hate speech

Oberaxe (Sept-Oct 2022): not only aggressiveness



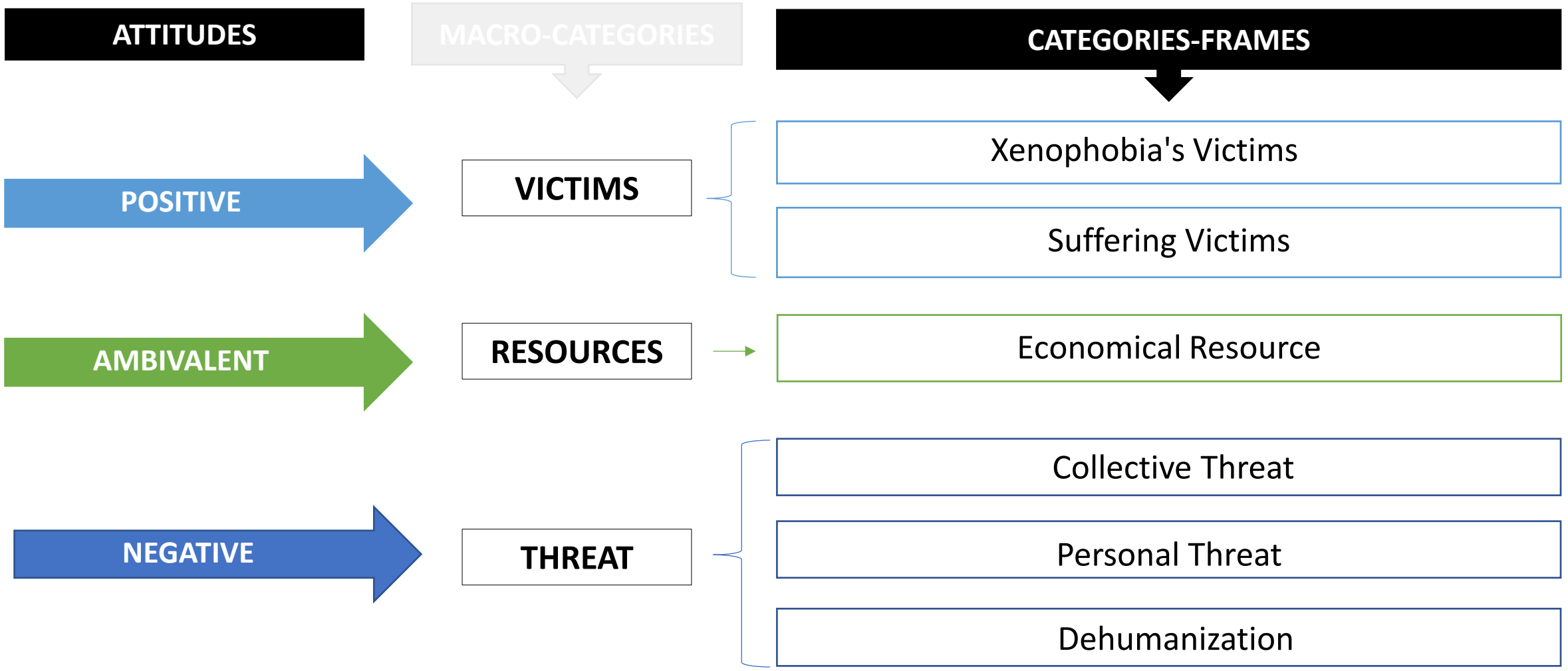
Oberaxe (Sept-Oct 2022): not only text



Hate Speech

- Related concepts and shared tasks
- Misogyny identification
- Spanish observatory on racism and xenophobia
- **Implicit HS: stereotypes, irony/sarcasm, hurtful humour**
- HS in memes
- Strategies against HS

Implicit HS: stereotype identification



Sánchez-Junquera J., Chulvi B., Rosso P., Ponzetto S. (2021) How do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. Applied Sciences, 11(8), 3610

StereoImmigrants dataset

- Texts on stereotypes about immigrants from the Spanish Congress of Deputies
- Annotation in collaboration with a **social psychologist** based on **frames** (i.e., different scenarios) where politicians place immigrants in their speeches

Label	Length	Texts	
Stereotype	45.62 ± 24.69	1673	3635
Non-stereotype	36.00 ± 21.17	1962	
Victims	48.93 ± 27.5	743	1479
Threat	45.84 ± 24.42	736	

Transformers and attention mechanism

- BETO: BERT  for Spanish

	S/N	V/T
Original text	0.82	0.79
Masking Technique with *AbsFreq*	0.79	0.75
Masking Technique with *RelFreq*	0.84	0.81
BETO	**0.86**	**0.83**

Victims:

BETO: hay una situación de desamparo en muchas personas a la que necesitamos dar una solución .

Threat:

BETO: el tiempo nos ha dado la razón , se ha convertido en un problema muy serio , en un problema muy importante tanto para europa como para españa .

Sánchez-Junquera J., Rosso P., Montes-y-Gómez M., Chulvi B. (2021) Masking and BERT-based Models for Stereotype Identification. In: Procesamiento del Lenguaje Natural (SEPLN), 67:83-94

Characteristics of HS and stereotypes

Analysis of the relevance of various **linguistic features** in texts that contain hate speech or stereotypes, reveals:

- both are featured by **negative emotions and feelings**: anger, awe, disgust, aggressiveness and fear
- **HS**: offenses are related to animals, physical disabilities or diversity and behaviours/morality
- **Stereotypes**: offenses are linked to economic and social issues, cognitive and ethnic sphere, even if in a more indirect way
- Both are characterized by specific **syntactic patterns**: negation and adverbial locutions
  - HS: mark some characteristics of **outgroup**, juxtaposing it sometimes with the ingroup
  - Stereotypes: increase the **intensity of some beliefs** (typical topics against minorities such as defects, morality, crimes and social advantages)

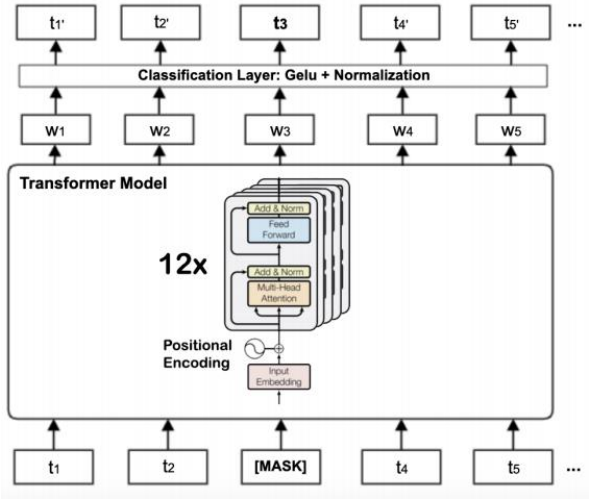
DETESTS@IberEval 2022 and 2024...

DETEction and classification of racial Stereotypes in Spanish: https://detestsiberlef.wixsite.com/detests

Implicit HS: sarcasm

Stiamo consegnando l'Italia ai stranieri..... GrazieStato
We're handing Italy over to foreigners..... ThanksState

Dai ragazzi, e Natale! Portiamo un po' di calore al campo nomadi. Io penso alla benzina, voi portate i fiammiferi?
Come on guys, it's Christmas! Let's bring some warmth to the nomad's camp. I'll take care of the gasoline, will you bring the matches?



Frenda S., Cignarella A., Basile V., Bosco C., Patti V., Rosso P. (2022) The Unbearable Hurtfulness of Sarcasm. In: Expert Systems with Applications (ESWA), vol. 193

Stylistic and syntactic features

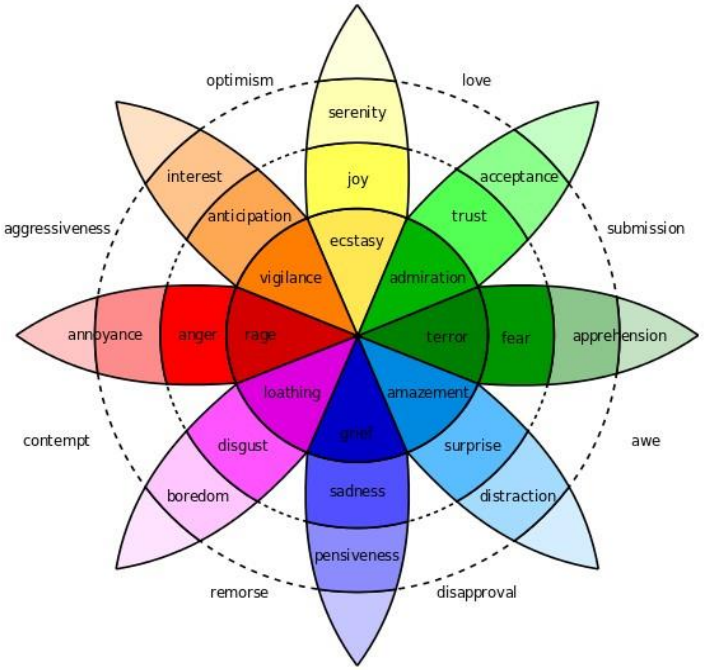
- Stylistic Features
  - punctuation
  - negation

- Syntactic Features
  - adverbial locutions
  - intensifiers
  - discourse connections
  - mentions
  - nominal phrases

Semantic features

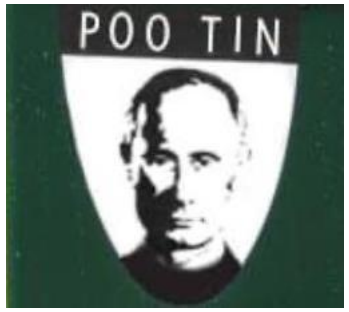
- Incongruities and Similarities
- Affective and Hurtful Language
  - the sentiment and the variation of sentiment using **Sentix**
  - categories of hurtful words employing **HurtLex**
  - emotions, feelings and their variation using **EmoLex**

Category	Description
PS	Ethnic Slurs
RCI	Location and Demonyms
PA	Profession and Occupation
DDP	Physical Disabilities and Diversity
DDF	Cognitive Disabilities and Diversity
DMC	Moral Behavior and Defect
IS	Words Related to Social and Economic advantages
OR	Words Related to Plants
AN	Words Related to Animals
ASM	Words Related to Male Genitalia
ASF	Words Related to Female Genitalia
PR	Words Related to Prostitution
OM	Words Related to Homosexuality
QAS	Descriptive Words with Potential Negative Connotations
CDS	Derogatory Words
RE	Felonies and Words Related to Crime and Immoral Behavior
SVP	Words Related to the Seven Deadly Sins of the Christian Tradition



(Hurtful) humour

Wordplay as homophony



Example of **wordplay** as **homophony**:
identical sound but different spelling

Ermakova L., Miller T., Boccou J., Digue A., Damoy A., Campen P. (2022) Overview of the CLEF 2022 JOKER Task 2: Translate Wordplay in Named Entities. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, vol. 3180, pp. 1666-1680

Wordplay as portmanteau

Example of **wordplay** as **portmanteau** (word formed by merging the sounds and meaning of two different words) + **paronymy** in the case of the 2nd word (a slightly difference in both spelling and sound):

#HijoPutin = hijo (son) + puta (bitch)

HUHU@IberEval 2023

sites.google.com/view/huhuatiberlef23/huhu

HUHU · Evaluation · Important Dates · Organizers · Data

HUrtful HUmour (HUHU)

Detection of humour spreading prejudice in Twitter

Shared Task at IberLEF 2023

Why **HUHU**: Everybody hurts, sometimes

- **Humour** as strategy to **spread prejudice**
- It **evades moral judgement**
- It **perpetuates stereotypes**
- It **justifies discriminatory acts**

Merlo L.I., Chulvi B., Ortega-Bueno R., Rosso P. (2023) **When Humour Hurts: Linguistic Features to Foster Explainability**. In: Procesamiento del Lenguaje Natural (SEPLN), num. 70, pp. 85-98

HUHU: Target groups of prejudice

- **Women and feminists** (G1)

Los hombres se fijan mucho en el físico de las mujeres. Ellas, en cambio, no se fijan en algo tan superficial, se fijan más en el valor del interior... de la billetera.

- **LGBTI+ community** (G2)

¿En qué se parece una sirena y un transexual? En que los dos son mujeres con cola.

- **Immigrants and racially discriminated people** (G3)

¿Cómo haces para que un estadio lleno de africanos haga una ola? Pásales una bolsa de pan Bimbo por encima

- **Overweight people** (G4)

Amor, ¿Estoy guapa? Tienes el cuerpo de un Dios ¿Ah, si? ¿El de cuál? El de Buda

HUHU: Subtasks

1. HUrtful Humour detection

2a. Prejudice target detection

2b. Degree of prejudice detection (from 1 to 5)

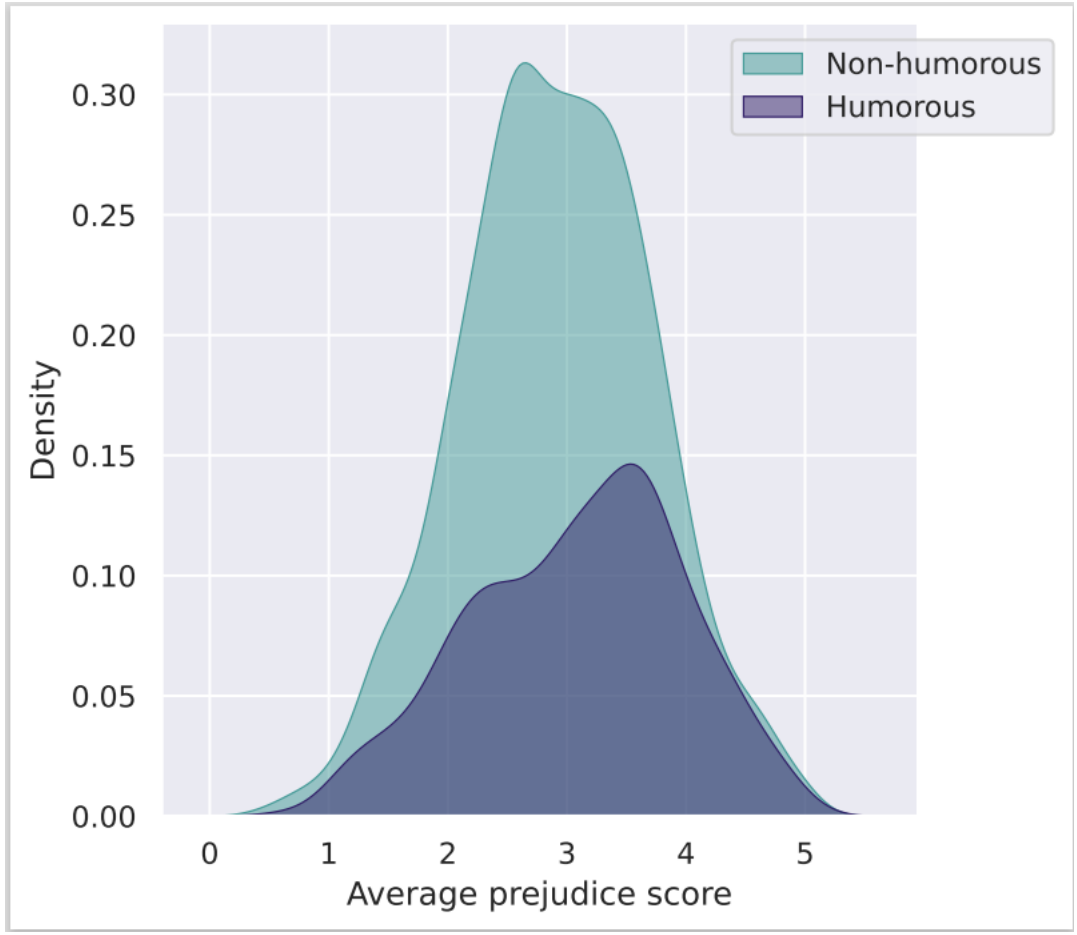
HUHU: Dataset

- 80 students of psychology manually tracked down approx., 900 **Twitter** accounts of users that spread hate speech using humour in **Spanish**
- Removal of duplicated tweets
- 3 annotators (one male and two females UPV students via ServiPoli foundation) to annotate: (i) if the tweet expresses **prejudice** and (ii) if with **humour**
- **5 annotators** (two males and three female UPV students via ServiPoli foundation) to annotate the **prejudice degree**

HUHU dataset: Statistics and degree of prejudice

Source	😆	😐	G1	G2	G3	G4
Crawled	607	2323	1652	791	753	169
HAHA	518	1	328	66	89	100
Total	1125	2324	1980	857	842	269

Source	😆	😐	G1	G2	G3	G4
Train	869	1802	1292	607	664	214
Test	256	522	688	250	178	55



Humorous tweets more hurtful

HUHU participation (46 teams: 130 runs) and results

Team	run	F_1-score ↑
RETUYT-INCO	1	0.820
BERT 4EVER	2	0.799
CISHUHUC	1	0.796
BLOOM-1b1		*0.789*
MosquitosBiased	1	0.784
HUHU-RMA-2023	1	0.782
amateur37	1	0.781
MJR	1	0.779
JPK	2	0.778
INGEOTEC	1	0.775
CAVIROS	2	0.774
JUJUNLP	1	0.772
mesichiquito	1	0.766
LaVellaPremium	2	0.764
BETO		*0.759*
SVM-3gram-char		*0.679*
allTrue		*0.492*

Subtask 1

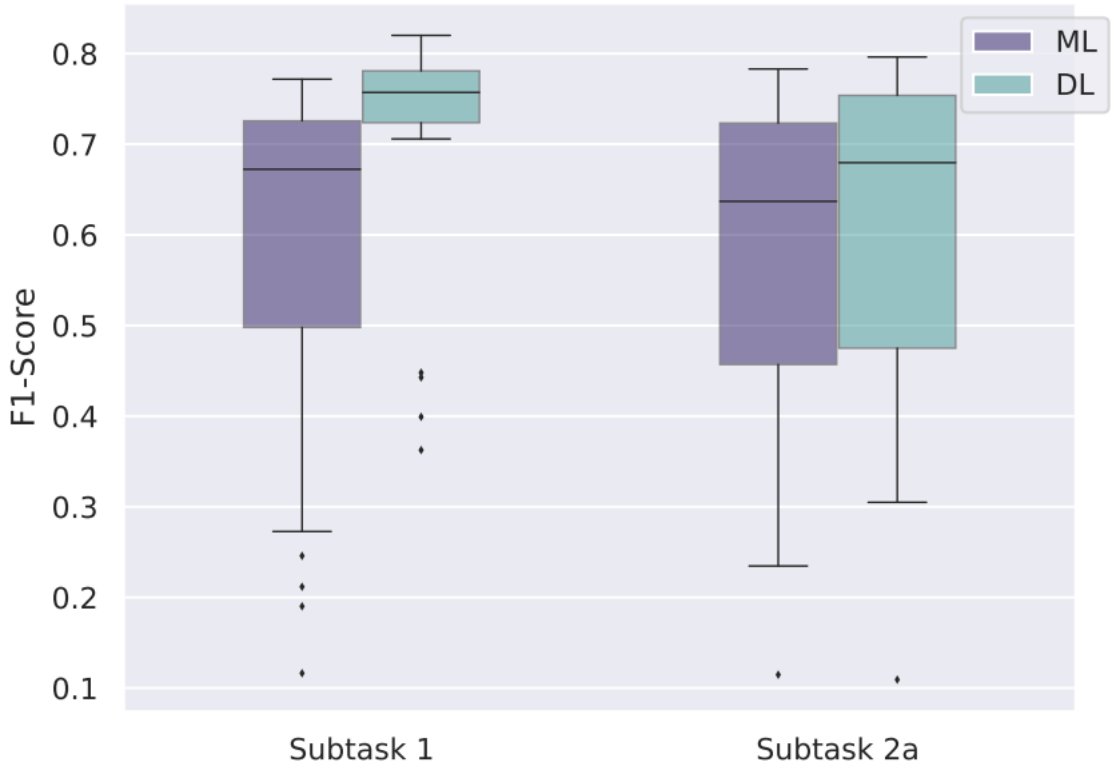
Team	run	Macro-F_1 ↑
JUJUNLP	1	0.796
Joe	1	0.783
Ratolins	1	0.778
RETUYT-INCO	1	0.773
BETO		*0.760*
BERT 4EVER	2	0.758
LaVellaPremium	1	0.753
MosquitosBiased	1	0.746
FENRIRFENIX	1	0.741
amateur37	1	0.739
Patata	2	0.732
mesichiquito	1	0.729
CAVIROS	2	0.727
Chincheta	1	0.722
SVM-3gram-char		*0.603*
allTrue		*0.482*

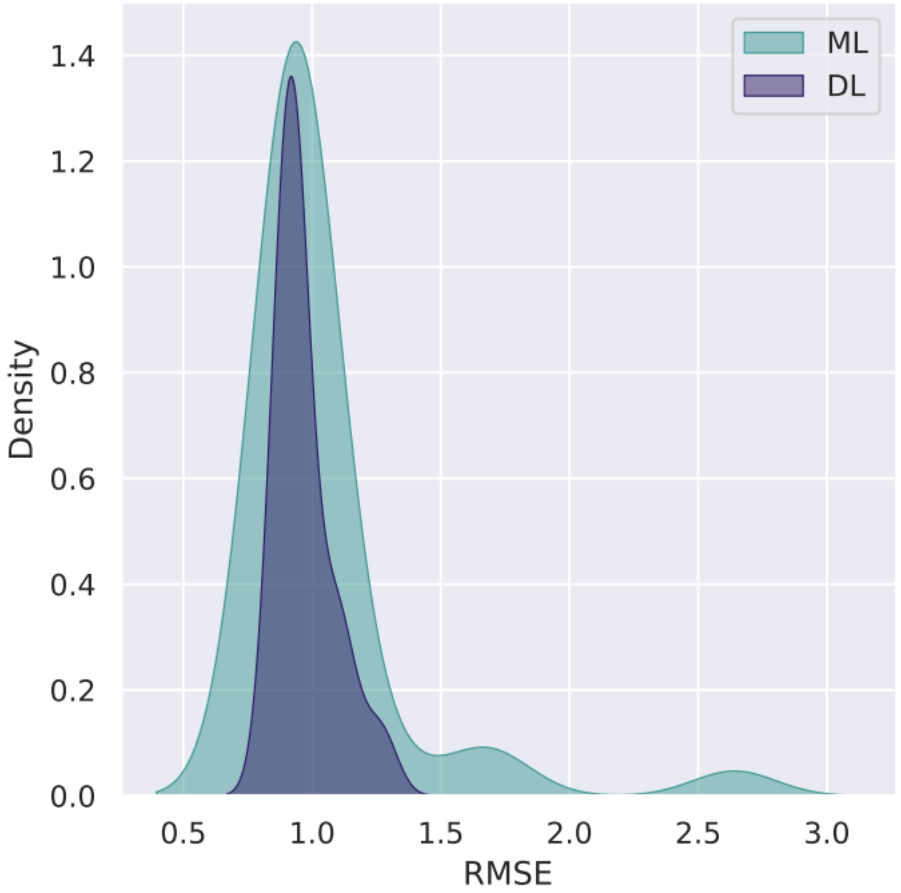
Subtask 2a
(top-ranked systems)

Team	run	RMSE ↓
M&C	1	0.855
Huhuligans	1	0.874
BETO		*0.874*
MosquitosBiased	1	0.881
Zeroimagination	1	0.881
CIC-NLP	1	0.881
ByteMeIfYouCan	1	0.887
cocalao	1	0.890
mesichiquito	1	0.891
MJR	1	0.893
MJR	2	0.893
FENRIRFENIX	1	0.895
LaVellaPremium	2	0.898
Climent	1	0.899
SVM-3gram-char		*0.907*
BLOOM-1b1		*0.915*

Subtask 2b

HUHU approaches: ML & DL





Fairness

FairTransNLP-Stereotypes. Fairness and Transparency for equitable NLP applications in social media: **Identifying stereotypes and prejudices** and developing equitable systems (PID2021-124361OB-C31)

Everybody hurts, sometimes

https://www.youtube.com/shorts/QeHP3nd_zW0

Hate Speech

- Related concepts and shared tasks
- Misogyny identification
- Spanish observatory on racism and xenophobia
- Implicit HS: stereotypes, irony/sarcasm, hurtful humour
- **HS in memes**
- Strategies against HS

HS in memes

Multi-modality: images, youtube videos, advertising

@ SemEval 2020

Text

Image

SUBTASK A

1. Sentiment Classification
positive / negative / neutral

SUBTASK B

2. Humor Classification
- **Sarcastic**
- Humorous
- **Offensive**
- Motivational (a motivational meme develops a positive attitude in the receiver)

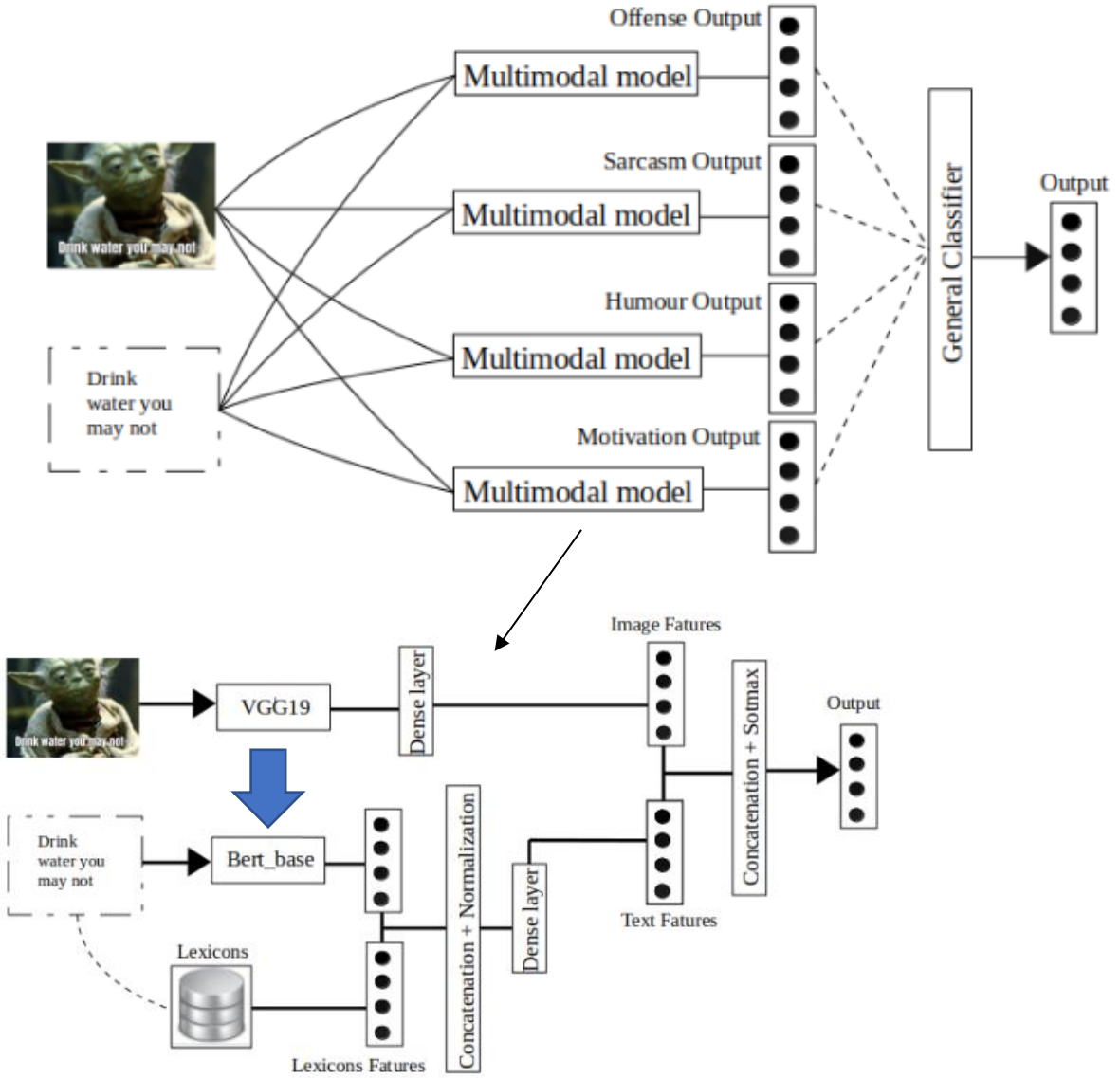
SUBTASK C

3. Scales of Semantic Classes (for each category in B)
not / slightly / mildly / very

C. Sharma, W. Paka, Scott, D. Bhageria, A. Das, S.a Poria, T. Chakraborty, B. Gabmbäck (2020). **Task report: Memotion analysis 1.0 @SemEval 2020: The visuo-lingual metaphor!** Proc. SemEval 2020

Our proposal

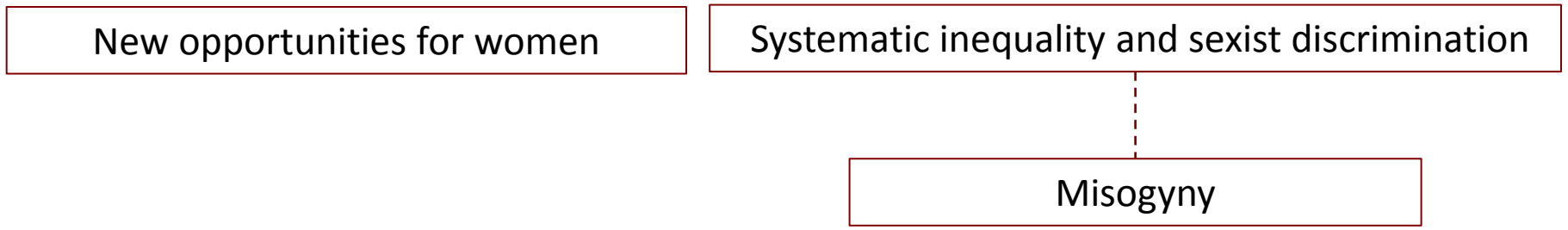
Model	Subtask A	Subtask B	Subtask C
Best system	0.3546	0.5183	0.3224
Proposal	0.3355	0.5093	0.3143
Last system	0.2477	0.4002	0.1267
Baseline	0.2176	0.5002	0.3008



G. De La Peña, P. Rosso, A. Giachanou (2020). **PRHLT-UPV at SemEval-2020 task 8: Study of multimodal techniques for memes analysis**. Proc. SemEval 2020

AMI -> MAMI: motivation

- Online platforms have become crucial in our society as instruments to define our identity
- **Women have a strong presence online**, particularly in **image based social media such as Facebook and Instagram**:
  - 75% females *vs* 64% males per day



https://themanifest.com/social-media/blog/how-people-use-social-media

AMI -> MAMI: motivation

Amnesty International Report (2017)

- 4000 women between the ages of 18 and 55 in 8 countries (*) were interviewed about misogyny
- **46% of them experienced misogynistic online abuse**

CONSEQUENCE

women involved in misogynous contents (e.g. text, images and video) decided to stop their posting activity in order to keep themselves safe

(*) Danmark, Italy, New Zealand, Poland, Spain, Sweden, UK, US

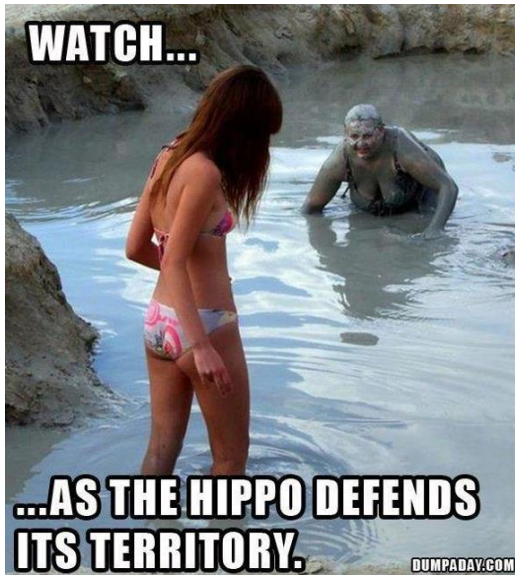
MAMI @ SemEval 2022

The Multimedia Automatic Misogyny Identification (MAMI) consists in the identification of **misogynous memes**, taking advantage of both text and images available as source of information. The task is organized around two main sub-tasks:

- **Sub-task A**: a basic task about ***misogynous meme identification***, where a meme should be categorized either as misogynous or not misogynous;
- **Sub-task B**: an advanced task, where the ***type of misogyny*** should be recognized among potential overlapping categories such as stereotype, shaming, objectification and violence.

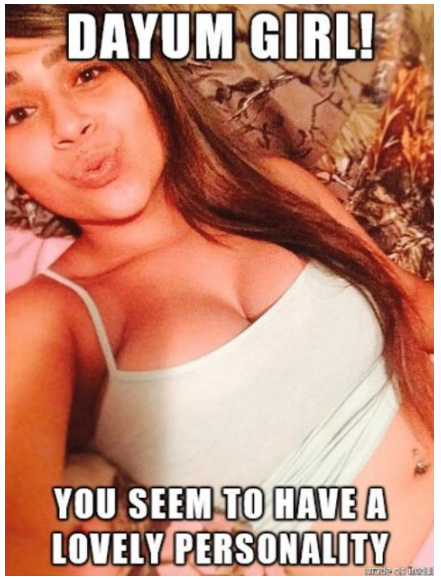
(*) co-organised with **University Milano-Bicocca & Google Jigsaw NY**

The MAMI dataset



Shaming

Stereotype



Objectification

Violence

10k MEMEs: **dataset available for training**

Collection of memes

- **Social media platforms** (Facebook, Twitter, Instagram, Reddit)
  - Searching for **threads dedicated to recent and popular memes**
  - Searching for **threads dedicated to anti-women or feminists supporters**
  - Exploring **discussions on sexism related to recent popular events** (e.g. Kavanaugh Case 3)

- **Websites** (9gag, Me.me, etc…)
  - Browsing by **hashtags** (#girl, #girlfriend, #women, #feminist)
  - Searching for **collections of meme's variations** with famous female characters

The MEME dataset

10k memes labelled memes for training (already released):

1. The first step is about collecting **socio-demographic** information:
   - Gender: female, male, unspecified
   - Age: age range between 18-15, 25-35, 35-45, 45-60, over-60
   - Location: country of birth

1. The second step is about Misogyny Labelling. Each labeller has to decide if a MEME is misogynous or not. If a MEME is labelled as misogynous, then two other questions will be provided:
   - **Type of Misogyny**: labellers should indicate (**multiple choice**) if the meme represents shaming, stereotype, objectification and/or violence.
   - **Misogyny Rating**: labellers should provide a rating about **how much the meme is misogynous** using stars, i.e. ☆, ☆☆ or ☆☆☆.

Hate Speech

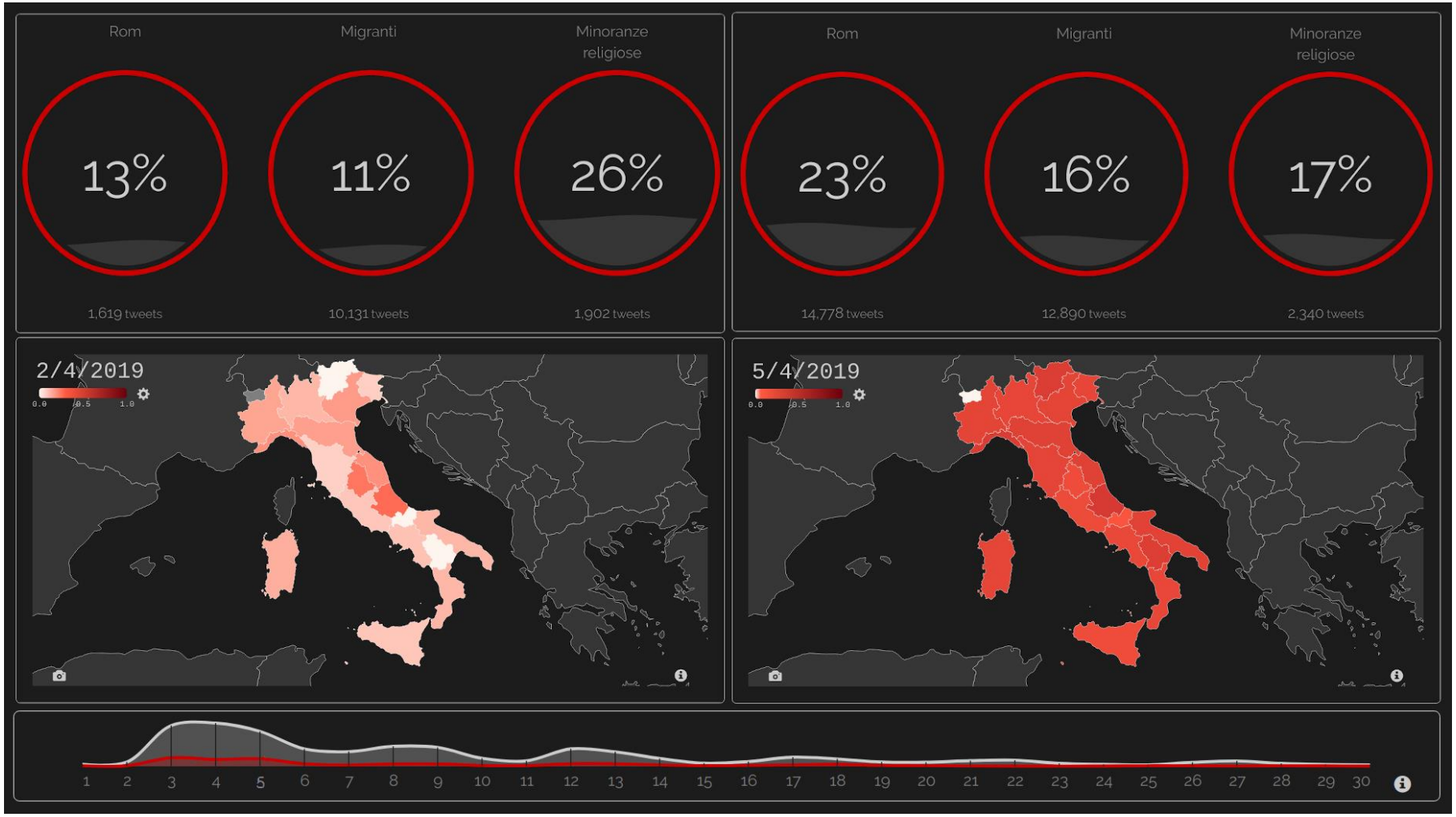
- Related concepts and shared tasks
- Misogyny identification
- Profiling haters @ PAN
- Spanish observatory on racism and xenophobia I
- Implicit HS: stereotypes, irony/sarcasm, hurtful humour
- HS in memes
- **Strategies against HS**

Strategies against HS

- **Monitoring** HS content in social media

Monitoring: contro l'odio

HATE SPEECH AND SOCIAL MEDIA



A. T. E. Capozzi et al. (2019). **Computational linguistics against hate: hate speech detection and visualization on social media in the "Contro L'Odio" project**. In Proc. CLiC-it 2019, ceur-ws.org, vol. 2481

Other projects

- **MANDOLA**: http://mandola-project.eu/ on reporting of illegal hate-related speech
- **CREEP**: http://creep-project.eu/ on monitoring cyberbullying online
- **Geography of Hate in the US**: https://www.bloomberg.com/news/articles/2016-12-15/the-geography-of-hate-in-the-u-s
- **Hatred Barometer**: https://www.amnesty.it/barometro-odio/ coordinated by the Italian section of Amnesty International
- **HateMeter**: http://hatemeter.eu/ special focus on anti-Muslim and opposing hate content with counter-narratives

Strategies against HS

- **Monitoring** HS content in social media

- **Pressurising** social media to deny posting extremely intolerant content

GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARÍA DE ESTADO DE MIGRACIONES
oberaxe
Observatorio Español del Racismo y la Xenofobia

Monitorización del discurso de odio en Internet en España

Reports, behaviour code, calls for projects

- Pew research center **online harassment report** (2017):

  41% of people have been target of HS
- EC signed a **behaviour code** with Twitter, Google, Facebook, and Microsoft to fight HS (2016)
- Twitter wants to forbid **dehumanising speech** (26/09/2018)
- EC call on **Monitor, prevent, and counter** HS online (2018)

Controversial humour…??

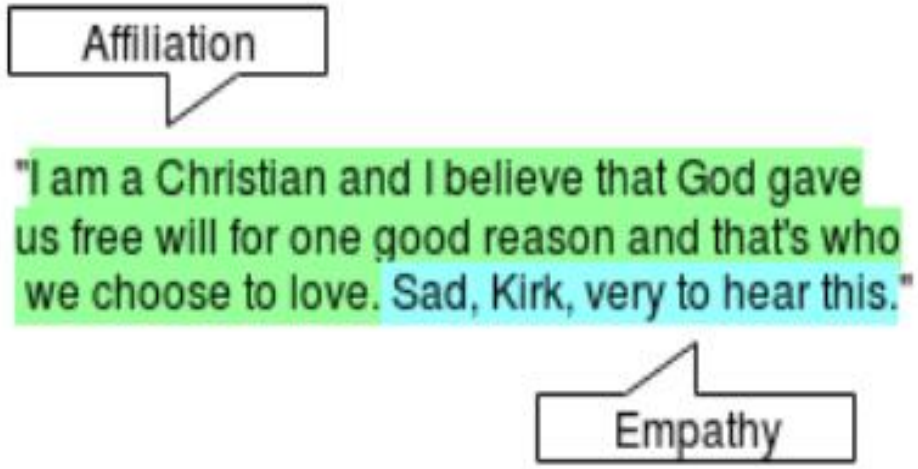
"...We do, however, allow clear attempts at **humor or satire** that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: **jokes**, stand-up comedy, popular song lyrics, etc.)."

What does Facebook consider to be hate speech? Retrieved from https://www.facebook.com/help/135402139904490

Strategies against HS

- **Monitoring** HS content in social media
- **Pressurising** social media to deny posting extremely intolerant content
- **Countering** HS: education of people wrt **stereotypes** and **prejudice**
  - Change perceptions and attitudes
  - Education and training
  - Campaigns against online HS

Countering HS



Dataset for counterspeech:

- 13,924 manually annotated comments from Youtube
- Labels: Counterspeech comment / Not

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	526	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	618	1100
Positive tone	359	237	268	864
Hostile	712	946	1083	2741
Total	2582	2811	3726	9119

B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S.K. Maity, A. Mukherjee (2019). **Thou shalt not hate: Countering online hate speech.** In Proc. of the Int. AAAI Conf. on Web and Social Media, vol. 13, pp. 369-380

Countering: Be Positive!

Funding Institution:

Google

Be Positive!

funded in 2020-22 under the "Google.org Impact Challenge on Safety" call

Be Positive! is aimed at automatically collecting and identifying online Hate Speech in order to increase positive contents addressed to groups vulnerable to discriminations and promote their active presence on social media. Be Positive! involves:

the improvement of Hate maps developed within the project Contro l'odio
the creation of an Automatic Writing Assistant, a tool that automatically suggests positive contents against Hate Speech
the organization of training courses addressed to schools, journalists, communication experts, health workers, minorities, and activists

Years

2020 - 2022

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

No hate speech movement https://www.coe.int/en/web/no-hate-campaign

NO HATE SPEECH MOVEMENT

Youth – Jeunesse

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

DIRECT LINKS

- Compendium of Resources
- Videos
- Contact us

What is the No Hate Speech Movement?

The No Hate Speech Movement is a youth campaign led by the Council of Europe Youth Department seeking to mobilise young people to combat hate speech and promote human rig at the national and local levels through national campaigns in 45 count 2017 through the work of various national campaigns, online activists a

What will you find in this site?

This website provides information about the campaign and the resources developed to prevent, counter and produce alternative narratives to hate speech. Hate speech remains an issue of major concern for human rights in Europe; this website also provides information about the other work of the Council of Europe alongside the youth campaign and, we hope, inspiration for everyone concerned with upholding human rights online.