Computational Linguistics against Hate:
Resources, Models, and Evaluation to Monitor and Contrast Abusive Language Online

Valerio Basile

CISA Seminar

Edinburgh, December 2nd 2019

> whoami

University of Turin, Italy

PhD
Groningen 2015

PostDoc
Inria Sophia Antipolis

> whoami

University of Turin, Italy

PhD
Groningen 2015

PostDoc
Inria Sophia Antipolis

Outline

Hate Speech Detection:
definition and evaluation

More than hate:
abusive, offensive, hateful language and bias

AI and humans:
cultural background and polarization of opinion

PARENTAL

ADVISORY

EXPLICIT CONTENT

Hate Speech Monitoring Group

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

contro ~~l'odio~~

Viviana Patti
Cristina Bosco
...many more

https://hatespeech.di.unito.it

Hate Speech Monitoring Group

Italian Hate Speech Corpus

6.000 tweets annotated by experts on:
- Hate Speech (binary)
- Aggressiveness
- Offensiveness
- Stereotype
- Irony

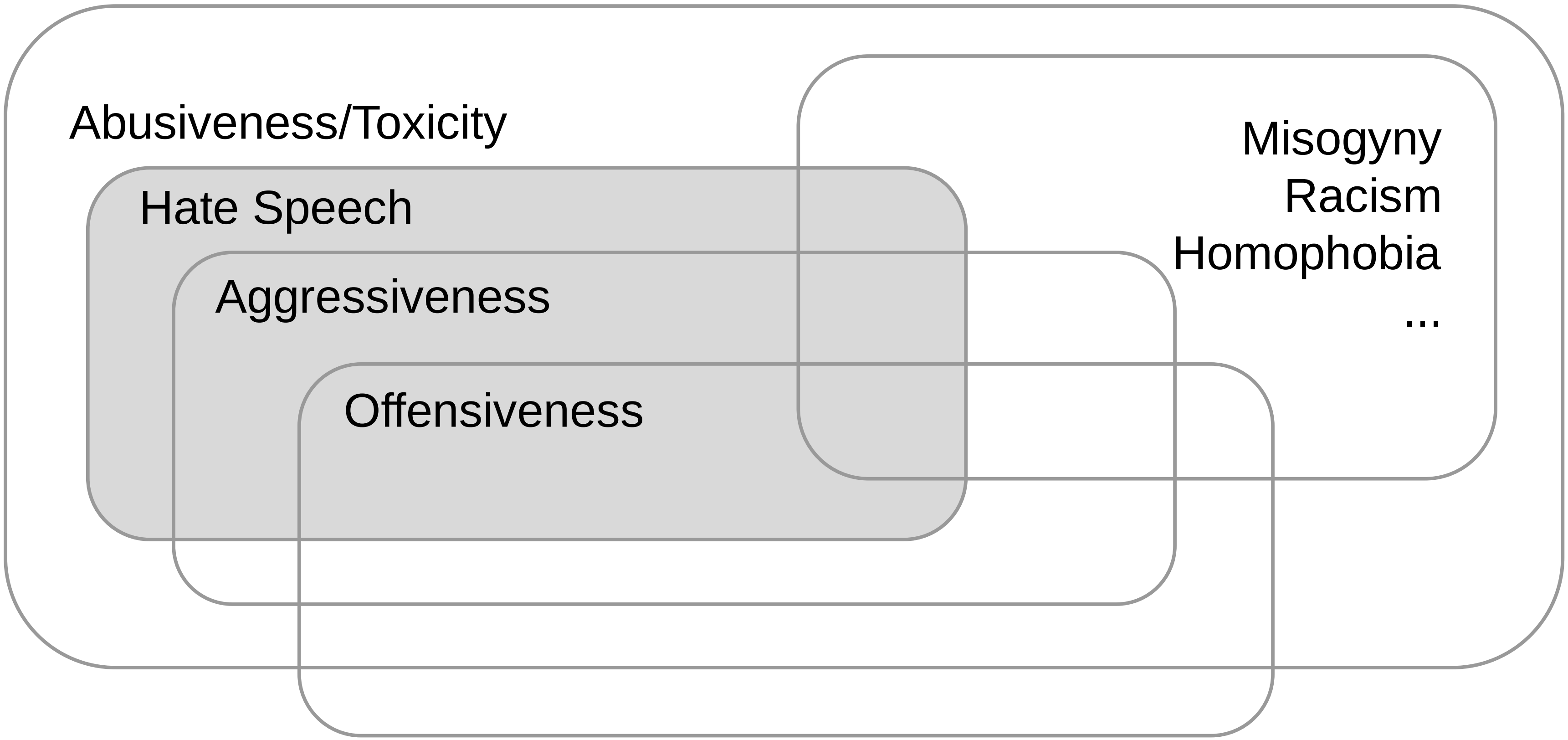
https://github.com/msang/hate-speech-corpus

HS Definitions

From Sanguinetti and Poletto:

- addressed, or just refer to, one of the minority groups identified as HS targets, or to an individual considered for its membership
- spreading, inciting, promoting or justifying violence against a target.

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]



Abusiveness/Toxicity
Hate Speech
Aggressiveness
Offensiveness
Misogyny
Racism
Homophobia
...

Evaluation Campaigns

EVALITA 2018
Two tasks on HS

AMI - Automatic Misogyny Identification

HaSpeeDe - Hate Speech Detection

http://www.evalita.it/2018/tasks

#1 on Spanish AMI (EN and ES)
Alessandra Cignarella and Endang Pamungkas

HatEval 2019

Shared task #5 at SemEval 2019

Hate Speech detection on tweets

Languages: English, Spanish

Targets: Immigrants, Women

108 runs from 74 teams

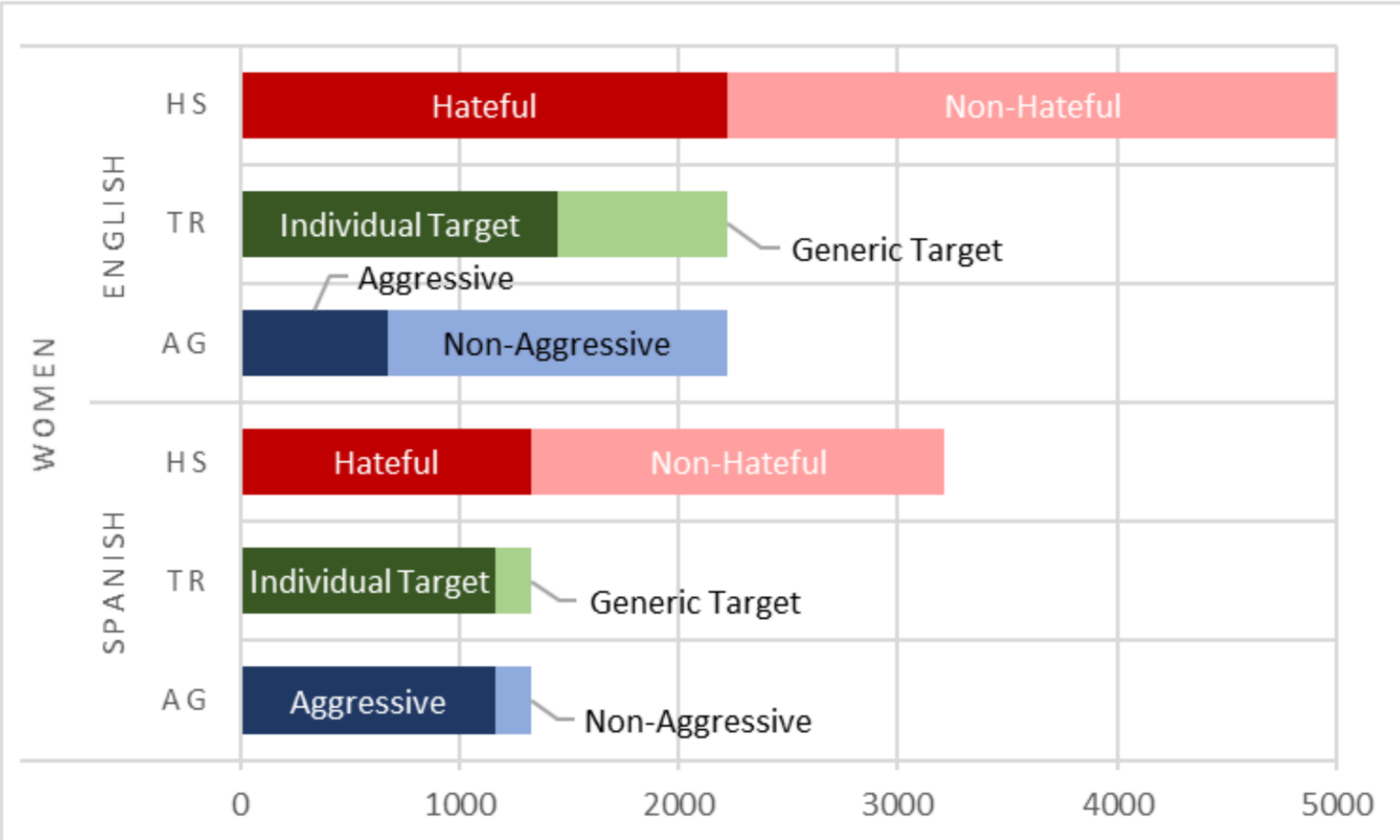
Together with C. Bosco, V. Patti, M. Sanguinetti, P. Rosso, F. Rangel, D.Nozza, E.Fersini

HatEval 2019: annotation

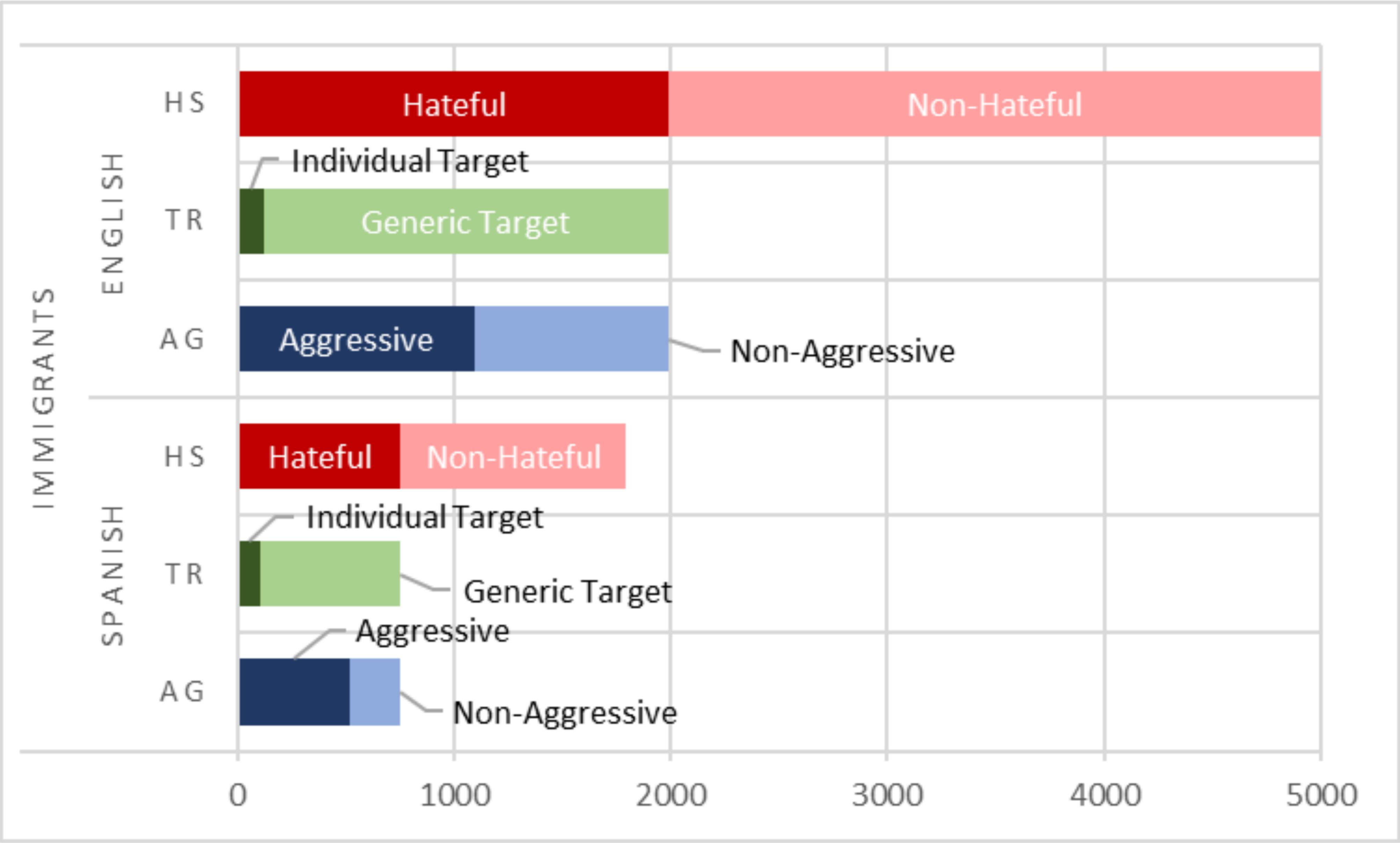
Two-level annotation

- HS - a binary value indicating if HS is occurring against one of the given targets
- Target Range - if HS occurs, a binary value indicating if the target is a generic group of people or a specific individual.
- Aggressiveness - if HS occurs, a binary value indicating if the tweeter is aggressive or not.

HatEval 2019: annotation



HatEval 2019: annotation



HatEval 2019: examples

[hateful]

[id: 32411] Cállate @ y la gran puta madre que te repario. Que le diste a la poltica...nada. Basura.

Shut up @ you motherfucker. What did you do for politics... nothing. Trash.[9]

[non-hateful]

[id: 33033] @ This is inhumane 😔 Karma is a bitch she ll get around these brainless heartless assholes!

[9]The target of the misogynistic hate here is Victoria Donda Prez, an Argentinian woman, human rights activist and member of the Argentine National Congress (mentioned in the at-mention of the original tweet).

HatEval 2019: examples

[Individual]:

[id: 4723] @ Come on box I show you my cock darling

[Generic]:

[id: 5823] Women are equal and deserve respect. Just kidding, they should suck my dick.

HatEval 2019: examples

[Aggressive]

[id:1890] Sick barstewards! This is what happens when we put up the refugees welcome signs! They not only rape our wives or girlfriends, our daughters but our ruddy mothers too!! https://t.co/XAYLr6FjNk

[Non-Aggressive]

[id: 945] @EmmanuelMacron Hello?? Stop groping my nation.Schneider: current migrant crisis represents a plan orchestrated and prepared for a long time by international powers to radically alter Christian and national identity of European peoples.http

HatEval 2019: evaluation

per-class

$$Accuracy = \frac{number\ of\ correctly\ predicted\ instances}{total\ number\ of\ instances}$$

$$Precision = \frac{number\ of\ correctly\ predicted\ instances}{number\ of\ predicted\ labels}$$

$$Recall = \frac{number\ of\ correctly\ predicted\ labels}{number\ labels\ in\ the\ gold\ standard}$$

$$F_1\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

macro

$$F_1\text{-score} = \frac{F_1(HS) + F_1(AG) + F_1(TR)}{3}$$

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Y_i, Z_i)$$

HatEval 2019: results

Best systems: RNNs (LSTM, GRU), Transformer

Spanish better than English (?)

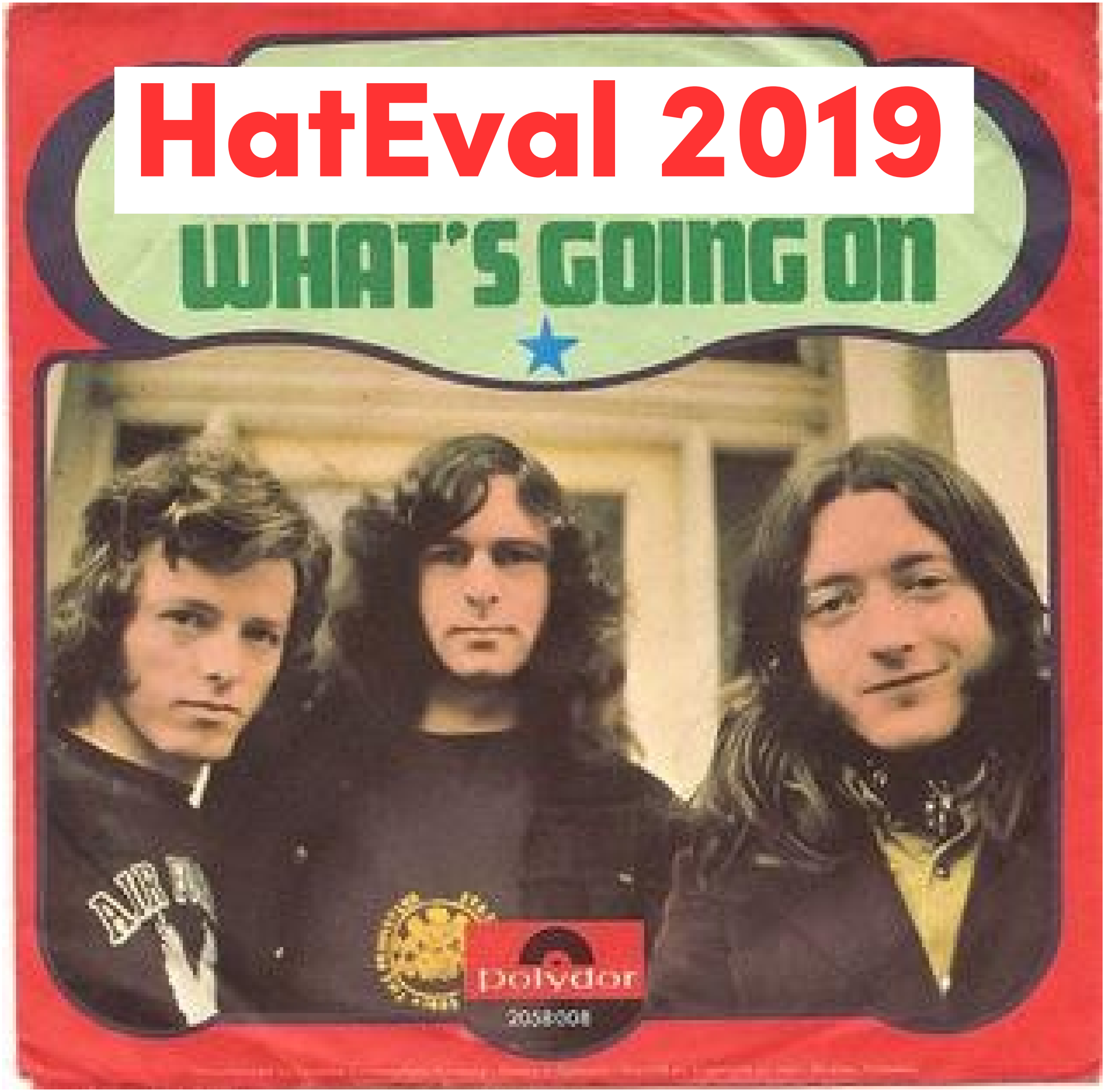
Best recall on EN ~= 0.5

Task B EN: all systems below MFC!

Complains of big drop in training→ test metrics

https://docs.google.com/spreadsheets/d/1wSFKh1hvwwQIoY8_XBVkhjxacDmwXFpkshYzLx4bw-0/edit#gid=503116726

HatEval 2019



Words and meanings

All in all it's just another brick in the **wall**

We’re going to build that **wall**

Weirdness Index

Given an general and a specific corpora

$$Weirdness(w) = \frac{w_s/t_s}{w_g/t_g}$$

w_s frequency of w in the specific corpus
w_g frequency of w in the general corpus
t_s total count of words in the specific corpus
t_g total count of words in the general corpus

Financial vs. BNC: *dollar, government, market*
From: Ahmad et al., 1999

Polarized Weirdness

Specific → subset determined by a class
General → its complement

Example
- Classes = {positive, negative}
- 100 instances: 50 positive and 50 negative
- 3,000 words in instances labeled positive
- 2,000 words in instances labeled negative
- *good* occurs 50 times in positive instances
- *good* occurs 5 times in negative instances

$$PW_{positive}(good) = 6.66$$
$$PW_{negative}(good) = 0.15$$

Weird HS words

Top 20 weird words in English HatEval
nodaca, enddaca, kag, womensuck, @hillaryclinton, americafirst, trump2020, taxpayers, buildthewallnow, illegals, @senatemajldr, dreamer, buildthewall, they, @potus, walkawayfromdemocrat, votedemsout, wethepeople, illegalalien, backtheblue.

Top 20 weird words in the Male GxG set
costituzionale, socialisto (socialist), Lecce, DALLA, utente, Samp, Sampdoria, Nera, allenatore, Orlando, Bp, ni, maresciallo, garanzia, cerare, voluto, pilotare, disco, caserma, From

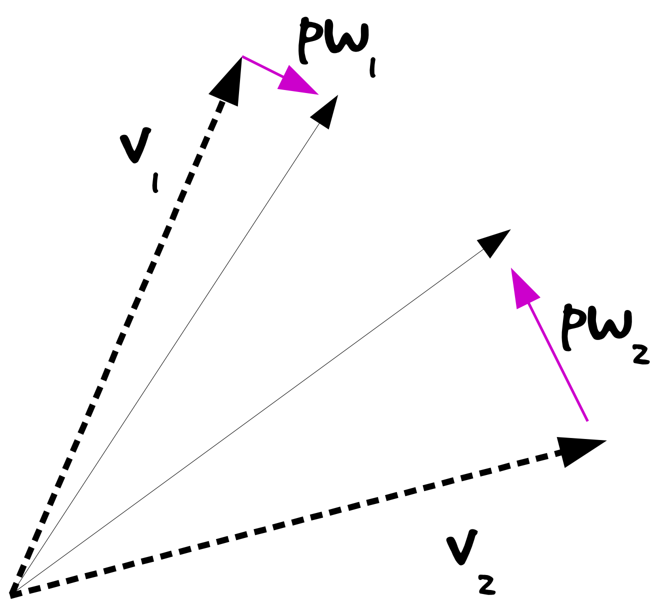
Word Embedding Adaptation

$$\vec{v}_1 = \frac{pw_1}{pw_1 + pw_2} \cdot \vec{v}_2 + \frac{pw_2}{pw_1 + pw_2} \cdot \vec{v}_1$$

$$\vec{v}_2 = \frac{pw_2}{pw_1 + pw_2} \cdot \vec{v}_1 + \frac{pw_1}{pw_1 + pw_2} \cdot \vec{v}_2$$

v_i vector representation of w_i

pw_i polarized weirdness of w_i wrt. Positive class (detection)



Experimental Evaluation

- CNN 64x8 hidden layer, ReLU activation, 4-size max pooling, ADAM optimization
- learning rate between 10^{-2} and 10^{-3}
  epochs between 10 and 25
- Keras (Python)
  mygithub://dnnsentenceclassification
- Polyglot word embeddings (Al-Rfou et al., 2013)
  64 dimensions, multilingual

Results I: Hate Speech Detection

Table 1: Results of the English Hate Speech Detection experiment.

		no-HS			HS			Avg.
Model	**Acc.**	**Pr.**	**R.**	**F1**	**Pr.**	**R.**	**F1**	**F1**
CNN	.528	.592	.595	.594	.437	.434	.436	.515
CNN+W	.527	.614	.497	.549	.450	.568	.502	.527

Table 2: Results of the Spanish Hate Speech Detection experiment.

		no-HS			HS			Avg.
Model	**Acc.**	**Pr.**	**R.**	**F1**	**Pr.**	**R.**	**F1**	**F1**
CNN	.468	.567	.401	.470	.398	.564	.466	.468
CNN+W	.482	.588	.394	.472	.413	.608	.492	.482

Weird Explainability

Table 1: Examples of words from the HatEval datasets, showing how their vector representation moves to reflect the semantic shift. Particular words that are generally neutral get closer to offensive words in the hate speech context.

Word embeddings	Generic word	Offensive word	Semantic shift	Cosine distance
Polyglot EN	wall	fuck	yes	1.224
Polyglot EN + P.W.	wall	fuck	yes	0.444
Polyglot EN	car	fuck	no	1.279
Polyglot EN + P.W.	car	fuck	no	1.413
Polyglot ES	directora (*director (F)*)	puta (*whore*)	yes	1.271
Polyglot ES + P.W.	directora (*director (F)*)	puta (*whore*)	yes	1.222
Polyglot ES	director (*director (M)*)	puta (*whore*)	no	1.366
Polyglot ES + P.W.	director (*director (M)*)	puta (*whore*)	no	1.411

Wiegand et al. 2019

On bias in datasets and how to correct it.

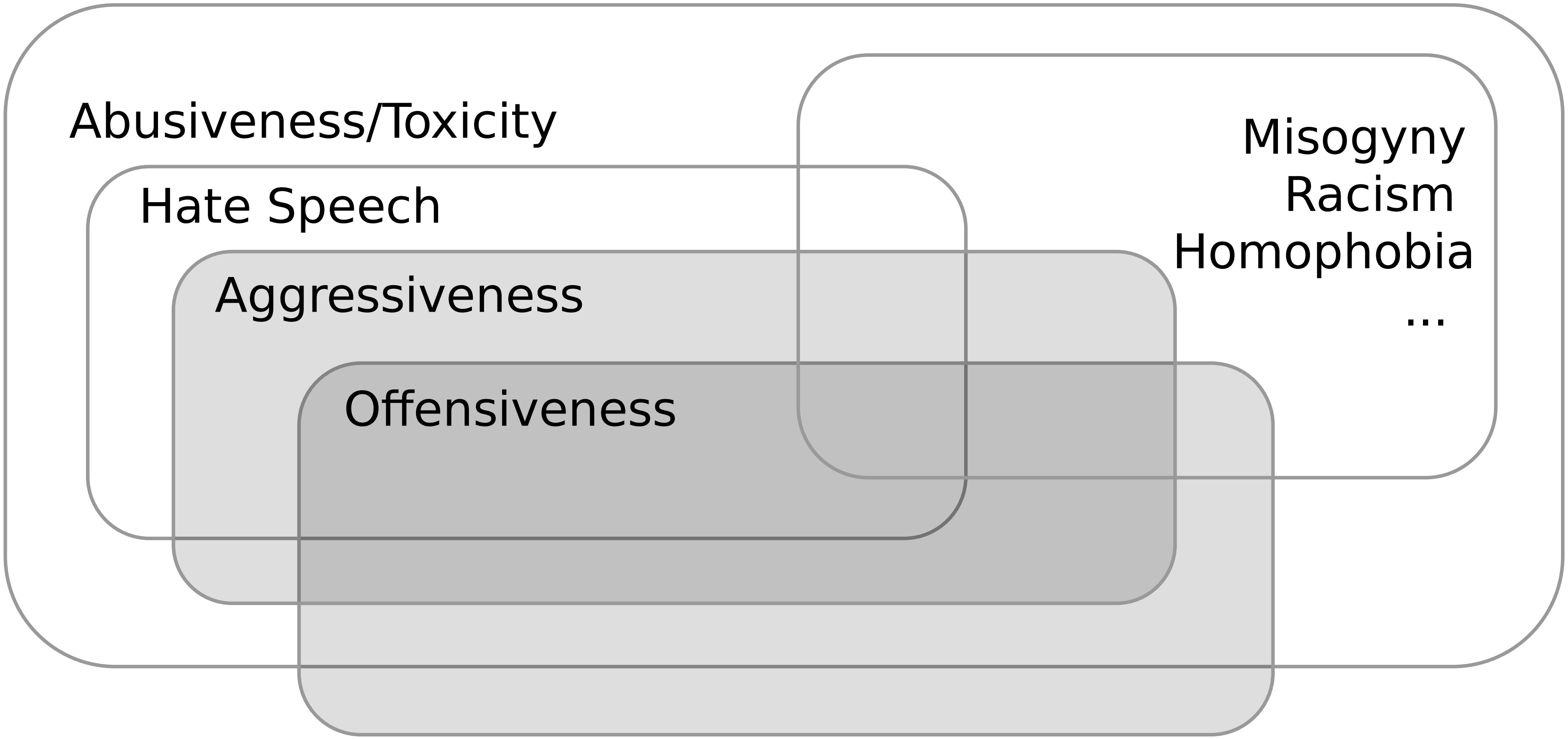
Explicit vs. Implicit abuse/hate

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).

Feature Set	Prec	Rec	F1
all words	80.91	80.08	80.49
(ii) query words removed	76.65	76.02	76.33
(i) topic words removed	75.07	74.41	74.72

Table 3: Impact of removing specific words from classifier trained and tested on *Waseem*.


Abusiveness/Toxicity
Hate Speech
Aggressiveness
Offensiveness
Misogyny
Racism
Homophobia
...

OLID and OffensEval

Offensive Language Identification Dataset (Zampieri et al. 2019)

Used for SemEval 2019 task 6: OffensEval

Table 1: OLID statistics per class: number of messages, average message length in tokens, average Offensive Prior. Asterisks mark statistical significance differences ($p < 0.05$). OFF = offensive; NOT = not offensive.

Class	Stats	Train	Test
OFF	# messages	4,400	240
	Avg. Length (token)	24.88*	25.91
	Offensive Prior (avg.)	0.2547*	0.2306*
NOT	# messages	8,840	620
	Avg. Length (token)	21.90	28.10
	Offensive Prior (avg.)	0.0614	0.0370

OLID lexicon analysis

Top keywords with TF-IDF

Mostly swear words
→ explicit

SWs in NOT class too

Joint work with T. Caselli and J. Mitrovic

Table 2: OLID top 10 keywords per class

Class	Train	Test
OFF	unepic	davidhogg
	sociopath	bitch
	shit	female
	witch	fuck
	pussy	clown
	omfg	oh
	silly	potus
	sucks	extremely
	monster	racist
	terrible	5k
NOT	woman	nickidagoat
	victim	dicks
	wtf	fucking
	weather	lack
	yesterday	smack
	way	better
	xx	revolting
	yo	literally
	vile	titty
	welcome	11

Dictionary-based classification

Reimplementation of Duluth approach
Based on lexicon by Wiegand et al. 2018

Table 3: OffensEval - Test: Evaluation of dictionary-based system and comparison against NULI, Duluth, and SVM baseline.

Approach	Class	P	R	F1 (macro)
Dictionary	NOT	.836	.872	.722
	OFF	.629	.558	
NULI	NOT	.902	.908	**.828**
	OFF	.758	.745	
Duluth	NOT	.832	.900	.735
	OFF	.673	.533	
SVM	NOT	.800	.920	.690
	OFF	.660	.430	

OffensEval ranking:
1) NULI (BERT)
...
6) Duluth
...

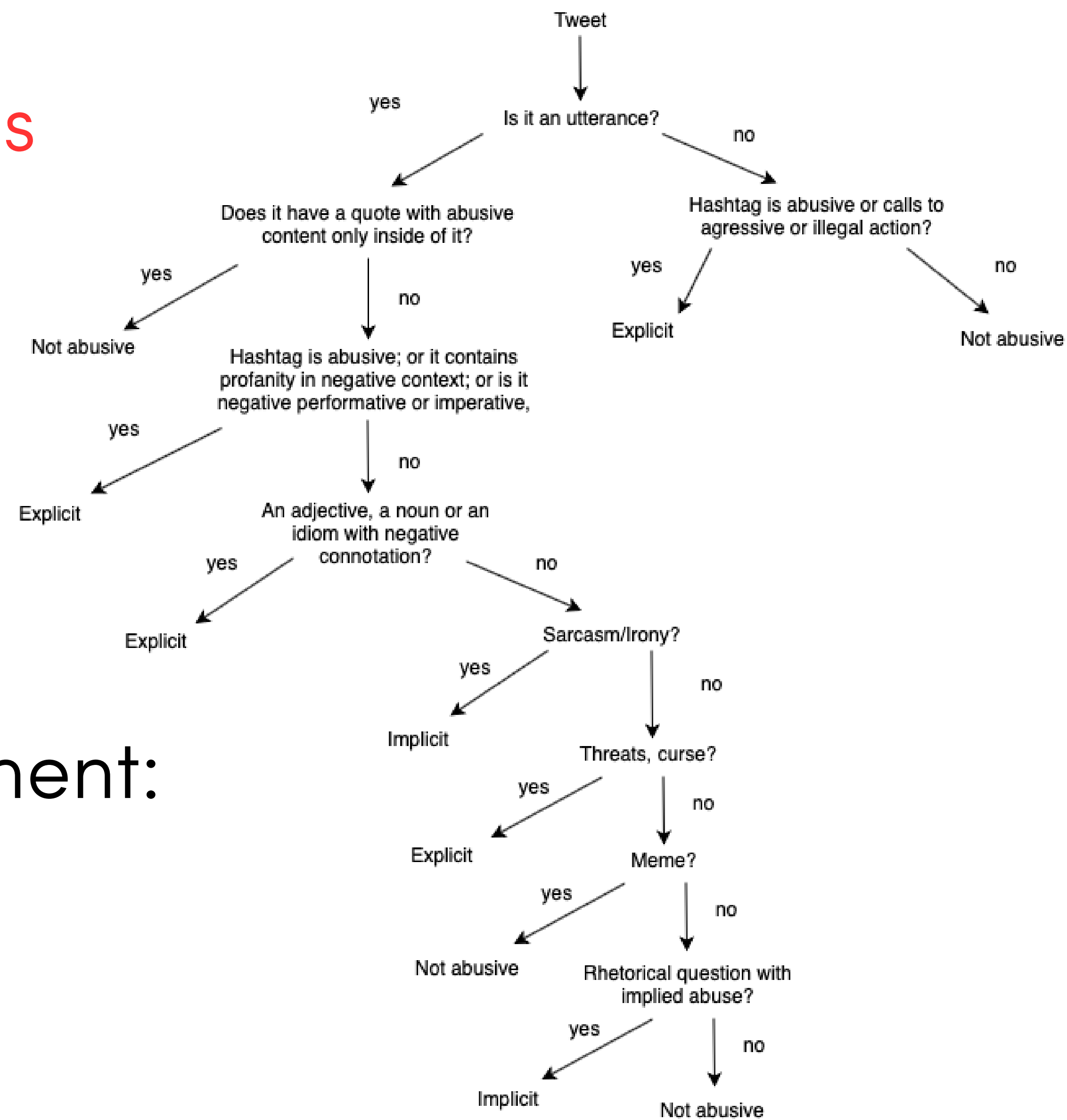
Explicit vs. Implicit

EXP/IMP Offensiveness

MSc Thesis by
Inga Kartoziya

EMP/IMP Abusiveness

Relatively high agreement:
Fleiss' K=0.61



Explicit vs. Implicit

Table 4: OffensEval: Explicit vs. Implicit offensive messages. EXP = EXPLICIT; IMP = IMPLICIT.

Data distribution	Class	Messages
Train	EXP	2901
	IMP	1499
Test	EXP	154
	IMP	86

- large overlap between OFF and EXP
- surprising amount of OFF NOTABU
- not negligible portion of abusive (EXP or IMP) untargeted

Table 5: AbuseEval v1.0: annotated data and annotation overlap with OLID/OffensEval. OLID/OffensEval labels: OFF = offensive; TIN = target; UTN = not targeted; NOT = not offensive. AbuseEval v1.0 labels: EXP = explicitly abusive; IMP = implicitly abusive; NOTABU = not abusive.

Data Distribution		OFF	TIN	UTN	NOT
Train	EXP	2,023	1,887	136	0
	IMP	726	668	58	0
	NOTABU	1,651	1,321	330	8,840
Test	EXP	106	103	3	0
	IMP	72	70	2	0
	NOTABU	62	40	22	620

Explicit vs. Implicit

BERT model fine-tuned on the Implicit/Explicit annotation of OffenseEval and AbuseEval (three-label classification)

Table 7: Results of the experiments on the Implicit vs. Explicit distinction.

Data set	Class	P	R	F1 (macro)
OffenseEval	NOT	$.868 \pm .023$	$.867 \pm .035$	$.614 \pm .157$
	IMP	$.240 \pm .059$	$.225 \pm .156$	
	EXP	$.637 \pm .029$	$.671 \pm .028$	
AbuseEval	NOTABU	$.864 \pm .019$	$.936 \pm .013$	$.535 \pm .023$
	IMP	$.234 \pm .086$	$.098 \pm .092$	
	EXP	$.640 \pm .060$	$.509 \pm .135$	

Offensive, Abusive, Hateful

So what is the relationship between these phenomena?

Offensive, Abusive, Hateful

So what is the relationship between these phenomena?

Experiment with “vanilla” pre-trained BERT

Table 8: Results of the cross-domain experiments.

Training set	Class	P	R	F1	
HatEval	NOT	.877 ± .021	.254 ± .053	.514 ± .033	~#7
	HS	.479 ± .012	.950 ± .022		
OffenseEval	NOT	.665 ± .068	.402 ± .091	.528 ± .016	
	HS	.462 ± .025	.712 ± .170		
AbuseEval	NOT	.661 ± .047	.672 ± .134	**.591**±.023	~#2
	HS	.531 ± .031	.510 ± .182		

Words matter

Phenomena matter

And the human?

The Human Factor

Datasets are made by humans.

Ethnicity and social background of the annotators may reflect their judgments in annotations.

Diverging opinions by annotators are valuable source of information for better training sets. (previous work: Aroyo and Welty; Checco et al.)

The Polarization Index

- Assuming a split into K groups of annotators
- P is high when
  - Intra-group agreements are high
  - Inter-group agreement is low

$$a(G_i) = 1 - \frac{\chi^2(G_i)}{|M|}$$

(a)

$$P(i) = \frac{1}{k} \sum_{1 \leq w \leq k} a(G_i^w)(1 - a(G_i))$$

(b)

Different from agreement!

PhD work of Sohail Akhtar

Polarization: Pilot Study

- HS Dataset on Brexit (119 tweets)

- 6 annotators in 2 groups:
  - Target: Immigrants, Muslims
  - Control: western background

Polarization: Pilot Study

Fleiss Kappa Measure for all Annotators			
Hate Speech	**Aggressiveness**	**Offensiveness**	**Stereotype**
0.35	0.21	0.30	0.20

Fleiss Kappa Measure for Group 1			
Hate Speech	**Aggressiveness**	**Offensiveness**	**Stereotype**
0.54	0.36	0.38	0.16

Control

Fleiss Kappa Measure for Group 2			
Hate Speech	**Aggressiveness**	**Offensiveness**	**Stereotype**
0.54	0.24	0.39	0.30

Target

Polarization: Pilot Study

Intra-group vs. inter-group agreement

	C2	C3	T1	T2	T3
C1	0.6	0.52	0.22	0.23	0.33
C2		0.52	0.16	0.18	0.26
C3			0.24	0.24	0.36
T1				0.69	0.52
T2					0.4

Data Augmentation Experiment

- Compute the P-index of every instance
- Instances with high polarization are filtered out
- Low polarization instances are replicated

Data from Waseem and Hovy 2017
+ new dataset ACCEPT

Table 1. Datasets used in the experiments with distribution of the labels.

Dataset	Positive class	Negative class	Total
Sexism	810	5,551	6,361
Racism	100	6,261	6,361
Homophobia	224	1,635	1,859

Data Augmentation Experiment

Table 2. Results of the prediction on Sexism dataset (1700 features).

Classifier	Accuracy	Precision	Recall	F1
SVM	95.11	**87.60**	71.60	78.74
SVM+P-max filter	95.13	86.40	73.01	79.11
SVM+replication	**95.27**	87.01	73.40	79.67
SVM+P-max filter+replication	**95.27**	86.60	**74.01**	**79.83**

Table 3. Results of the prediction on Racism dataset (1700 features).

Classifier	Accuracy	Precision	Recall	F1
SVM	98.55	55.40	11.01	18.40
SVM+P-max filter	98.58	59.01	12.01	19.88
SVM+replication	**98.61**	**70.01**	19.60	29.49
SVM+P-max filter+replication	**98.61**	69.80	**19.80**	**29.74**

Table 4. Results of the prediction on Homophobia dataset (3500 features).

Classifier	Accuracy	Precision	Recall	F1
SVM	**88.81**	61.01	11.40	19.02
SVM+P-max filter	**88.81**	**63.60**	13.60	22.30
SVM+replication	86.55	50.40	18.40	26.83
SVM+P-max filter+replication	87.63	47.90	**26.20**	**33.67**

Qualitative Analysis

By ranking the instances of a dataset by P-index, the most polarizing tweets emerge at the top

The vast majority of the tweets with P = 1 contain mixed remarks:

`@******* uh... did you watch the video? one of the women talked about how it's assumed she's angry because she's latina.`

Humour is highly polarizing

`Another #Arab car #terror attack in #Jerusalem #Israel. Will #Obama call it random traffic infringement? http://t.co/xxxxxxxx`

Topics in the ACCEPT data: gender theories and their education in school, family values

Conclusions (1)

Text classification alone is limited towards the understanding of these complex phenomena

Conclusions (2)

We need to start thinking about who is producing data for AI

What now?

Hate is a product of people

People are not islands