Iniciativas para identificar (algunos) contenidos maliciosos

Alberto Barrón-Cedeño
Università di Bologna
https://albarron.github.io
a.barron@unibo.it
@_albarron_

AI Tech Talk Valencia AI
13th October, 2022

Disclaimers

1. This is not a technical talk

2. I am biased, like everybody!

3. The higher my level of Italian, the lower my level of English, the less I remember the Iberian dialect, and the more I mix languages

About myself

Computing scientist working on

Natural Language Processing

Information Retrieval

Machine Learning

2022--20?? **Associate Professor**
2019-2022 **Senior Assistant Professor**
Università di Bologna

2014-2019 **Scientist**
Qatar Computing Research Institute

2012-2014 **Alain Bensoussan Fellow**
Universitat Politècnica de Catalunya

2012 **PhD in Computing Science**
Universitat Politècnica de València

Where am I?

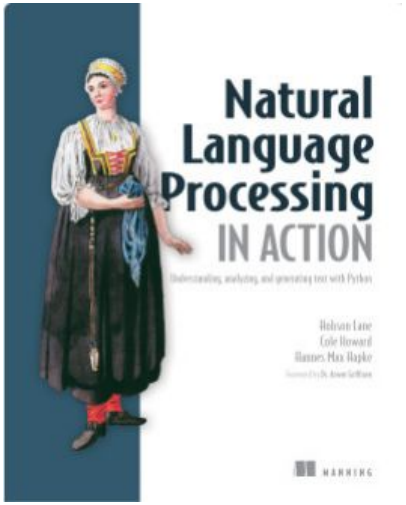
Università di Bologna

Campus di Forlì
Emilia-Romagna

My mission at UniBO

Bringing (real) computing to a D. of Interpreting and Translation

- Teaching in the MA Translation program:
  - Computational Linguistics
  - Computing Thinking and Coding
  - Software Localisation



- Supervising NLP PhD and MSc projects by students with a Linguistics/Translation background

Main collaborators involved

Arianna Muti

Katerina Korre

Giovanni
da San Martino

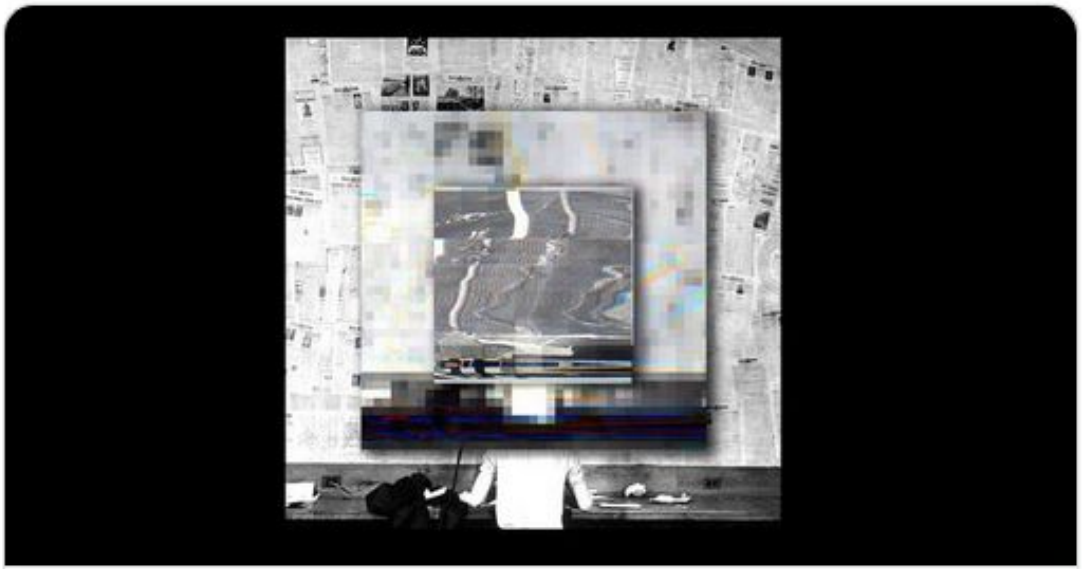
Preslav Nakov

Yifan Zhang

Contents

Yann LeCun
@ylecun

The supply of disinformation will soon be infinite, which will cause people to treat *all* information as disinformation, *unless* it comes with a traceable origin to a reliable and/or trusted source.
There are techs and standards to be deployed for this.



theatlantic.com
The Supply of Disinformation Will Soon Be Infinite
Disinformation campaigns used to require a lot of human effort, but artificial intelligence will take them to a whole new level.

5:55 PM · Oct 9, 2022 · Twitter for Android

153 Retweets **22** Quote Tweets **729** Likes

https://twitter.com/ylecun/status/1579138435398279170

Ramon Crespo (Decca) @rcr... · 8/9/22

Señora Pastor la unica manera de recuperar su credibilidad seria su dimision de la sexta y el divorcio del Sr Farreras. No se puede caer mas bajo en dignidad como ser humano, el suicidio tambien limpiaria su merecimiento como ciudadana. Un saludo.

https://twitter.com/_anapastor_/status/1580114140688834561

Errors from poor research (and writing)

Piadinas aprobadas por la IGP. MIRIAM GARCÍA

COCINA ITALIANA

PIADINAS DE EMILIA-ROMAÑA

El País (Spain), 27 May, 2021

https://elcomidista.elpais.com/elcomidista/2021/05/10/receta/1620635145_993526.html

Mainstream outlets are not always accurate

COCINA ITALIANA

PIADINAS DE EMILIA-ROMAÑA

"Las piadinas o *piadine* son un pan plano típico de la región de Emilia-Romaña [...] emparentadas con las tortas planas de todo el planeta, que van desde las tortillas mejicanas [sic] a los chapati indios

piadina

tortilla

chapati

Mainstream outlets are not always accurate

COCINA ITALIANA

PIADINAS DE EMILIA-ROMAÑA

"Las piadinas o *piadine* son un pan plano típico de la región de Emilia-Romaña [...] emparentadas con las tortas planas de todo el planeta, que van desde las tortillas mejicanas [sic] a los chapati indios

Nada más primigenio que hacer una pasta con harina y agua, aplastarla en forma de torta y cocerla en cualquier superficie caliente, un tipo de elaboración nacida en muchas zonas del mundo de forma independiente (por pura necesidad).

Sensationalising (and exaggeration)

EXCELSIOR
www.excelsior.com.mx

Mexico; 10 May, 2011

A Mexican makes history in Spain with a plagiarism detector

The researcher developed a new method that detects if a text has been copied from another language and posted on the internet

04/05/2011 09:44 EFE

la Repubblica

Italy; 31 March, 2021

Arianna's algorithm: "This is how I hunt posts against women in Twitter"

di Emanuela Giampaoli

Propaganda

Problematic journalism

- Ongoing (and uncorrected) errors that arise from poor research
- Sloppy verification
- Sensationalising that exaggerates for effect
- Hyper-partisan selection of facts at the expense of fairness

UNESCO (2018). Journalism, 'Fake news' & Disinformation. Handbook for Journalism Education and Training

Propaganda in the current affairs

- Politicians and artists do not need the media to reach their audience
  They have social media
- Journalists are struggling to produce news and attract both readers and sponsors
  - They have to act quick for it to be breaking news
  - They need to grab our attention
- Some newspapers are struggling to survive. They publish hidden marketing under request
- People do not read news (articles); they scroll social media

Propaganda techniques

"[...] based on social psychological research are used to generate propaganda. Many of these same techniques can be classified as logical fallacies, since propagandists use arguments that, while sometimes convincing, are not necessarily valid."

https://en.wikipedia.org/wiki/Propaganda_techniques

Loaded language	Causal oversimplification
Name calling or labeling	Slogans
Repetition	Appeal to authority
Exaggeration or minimization	Black-and-white fallacy, dictatorship
Doubt	Thought-terminating cliché
Appeal to fear/prejudice	Whataboutism, straw man, red herring
Flag-waving	Bandwagon, reductio ad hitlerum

Propaganda techniques (examples)

We are in the beginning of a mass extinction and all you can talk about is money and fairy tales of eternal economic growth.

Greta Thunberg, UNGA 2019

Appeal to fear. Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative.

Propaganda techniques (examples)

Loaded language.
Words/phrases with strong emotional implications

Name calling.
Labeling with something the audience fears, hates, finds undesirable or loves, praises

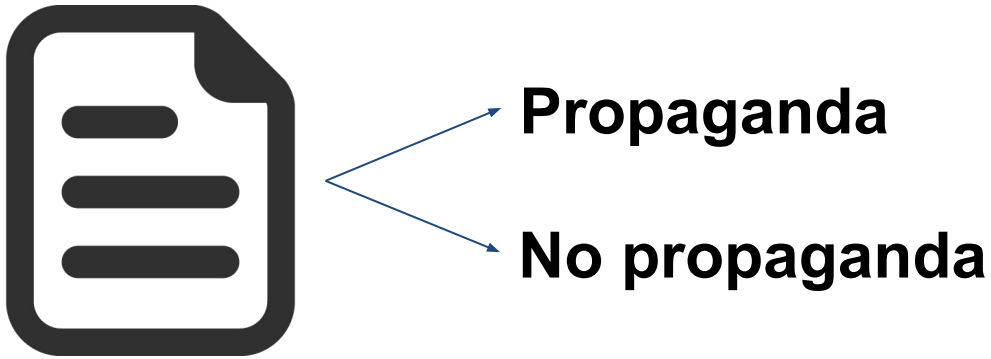
Doubt.
Questioning the credibility of someone/something

* Social media post, following the unfolding of a feminist pro-abortion demonstration on 28 September 2020, Mexico City

Objectives + Translation into NLP

- Developing tools to help readers notice propaganda
- Analysing the use of propaganda in media

Why? IMHO, it is one of the best ways to dim the effect of disinformation



A “simple” approach

ELSEVIER

Information Processing & Management

Volume 56, Issue 5, September 2019, Pages 1849-1864

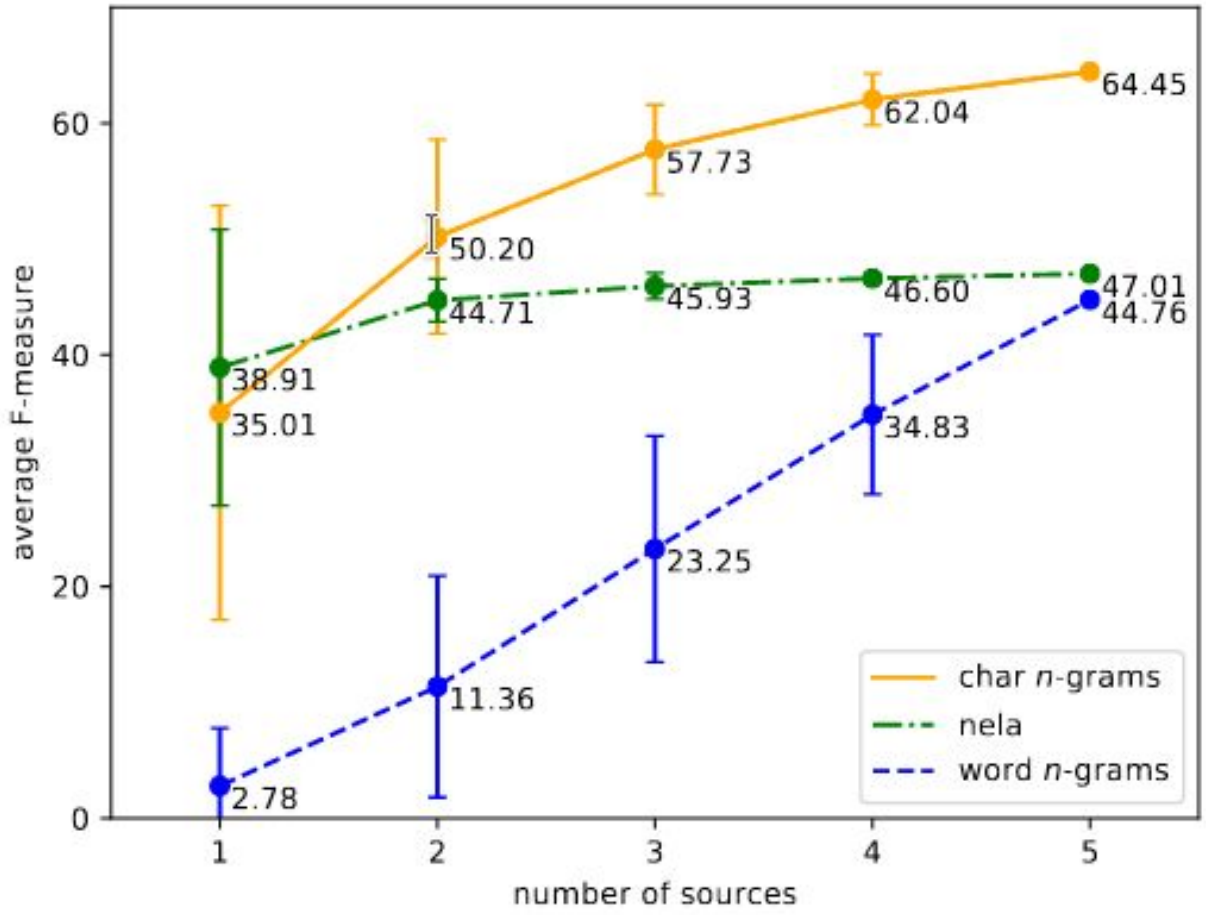
Proppy: Organizing the news based on their propagandistic content

Alberto Barrón-Cedeño [1, a] ✉, Israa Jaradat [1, b] ✉, Giovanni Da San Martino [c] ✉, Preslav Nakov [c] ✉

eds. Jorge, Campos, Jatowt, Nunes

- Supervised model to compute a propagandist index (F_1=80+)
- Qprop-18: corpus with 50k+ articles from 104 sources
- A bunch of experiments that show, among others, that both style and vocabulary (topic/domain independent) are crucial to spot propaganda

A “simple” solution



- It includes an analysis on the stability of different representations with respect to the number of propagandist and non-propagandist sources

Results summary

features	F_1
word n-grams	75.55
lexicon	44.87
voc. richness	29.72
readability	21.50
char n-grams	**82.13**
nela	50.98
char n-grams+lexicon	81.94
char n-grams+nela	**82.75**
readability+nela	76.83


Identifying propaganda techniques

PRTA: We need to go beyond

From binary document classification to multiclass sequence labelling

- To highlight the trick to the reader
- To *explain* why an article is propaganda
- To justify the flagging of an article as propagandist

PRTA: Propaganda techniques

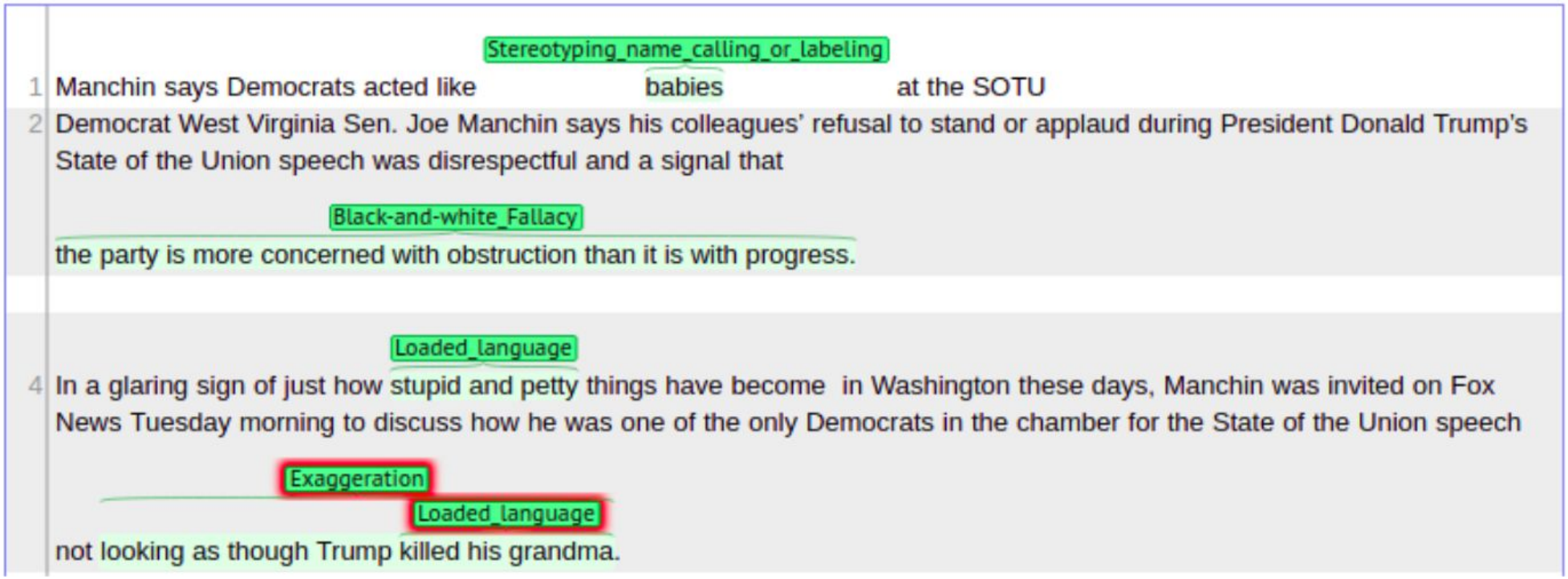
We identified 18 propaganda techniques which…

- Are likely to be used in news articles

- Can be spotted by an educated eye, without having to consult external information sources

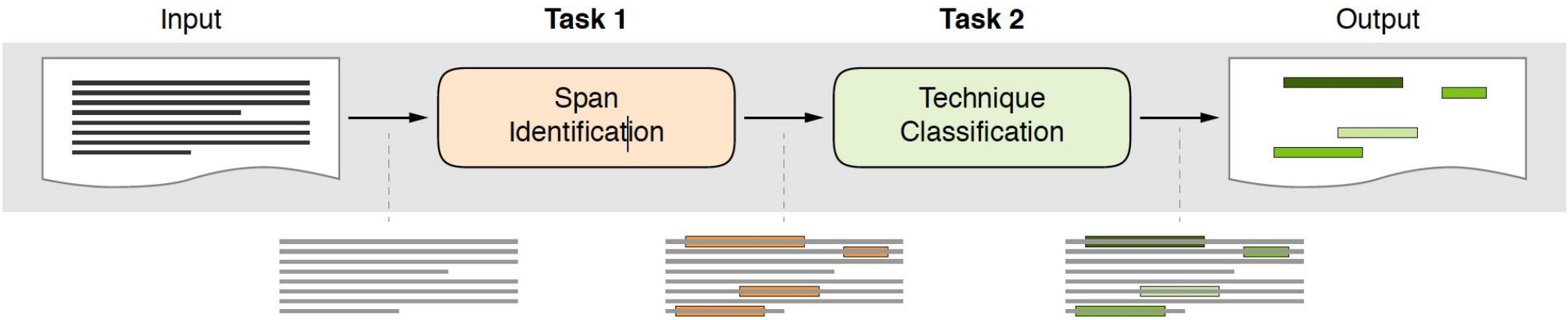
Da San Martino, Yu, Barrón-Cedeño, Petrov, and Nakov. **Fine-Grained Analysis of Propaganda in News Articles**. EMNLP-IJCNLP 2019

PRTA: PTC-SemEval20 Corpus



- 6+ professional annotators (hard to crowdsource)
- 2 annotators/article; 3 during consolidation
- ~50 articles/week

PRTA: Translation into specific NLP tasks



T1. Span identification

← Plain-text document

→ Specific propagandist text snippets

T2. Technique classification

← Propagandist snippet
Document context

→ Propaganda technique

PRTA instantiation into SemEval 2020

https://propaganda.qcri.org/semeval2020-task11

Da San Martino, Barrón-Cedeño, Wachsmuth, Petrov, Nakov
SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles
SemEval 2020

PRTA: SemEval T11.1 Span identification

Rank. Team	Transformers									Learning Models								Representations											Misc		
	BERT	RoBERTa	XLNet	XLM	XLM RoBERTa	ALBERT	GPT-2	SpanBERT	LaserTagger	LSTM	CNN	SVM	Naïve Bayes	Boosting	Log regressor	Random forest	CRF	Embeddings	ELMo	NEs	Words/n-grams	Chars/n-grams	PoS	Trees	Sentiment	Subjectivity	Rhetorics	LIWC	Ensemble	Data augmentation	Post-processing
1. Hitachi	✓	✓	✓	✓	✓		✓				✓						✓	✓		✓			✓						✓		
2. ApplicaAI		✓															✓													✓	
3. aschern		✓															✓	✓				✓							✓		✓
4. LTIatCMU	✓									✓							✓	✓		✓	✓		✓	✓	✓	✓		✓			
5. UPB	✓									✓							✓	✓												✓	
7. NoPropaganda	✓								✓	✓							✓	✓				✓									
8. CyberWallE	✓									✓		✓											✓		✓		✓				
9. Transformers	✓	✓								✓	✓								✓										✓		
11. YNUtaoxin	✓	✓	✓																												

	Test			
Team	**Rnk**	**F_1**	**P**	**R**
Hitachi	1	**51.55**	56.54	47.37
ApplicaAI	2	49.15	59.95	41.65
aschern	3	49.10	53.23	45.56
LTIatCMU	4	47.66	50.97	44.76
UPB	5	46.06	58.61	37.94

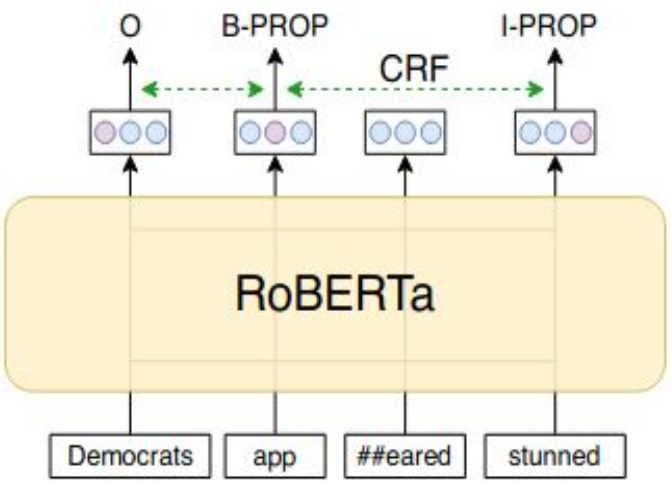
PRTA: SemEval T11.2 Technique classification

Rank. Team	Transformers										Learning Models											Representations										Misc		
	BERT	R BERT	RoBERTa	XLNet	XLM	XLM RoBERTa	ALBERT	GPT-2	SpanBERT	DistilBERT	LSTM	RNN	CNN	SVM	Naïve Bayes	Boosting	Log regressor	Random forest	Regression tree	CRF	XGBoost	Embeddings	ELMo	NEs	Words/n-grams	Chars/n-grams	PoS	Sentiment	Rhetorics	Lexicons	String matching	Topics	Ensemble	Data augmentation	Post-processing
1. ApplicaAI			✓																	✓													✓	✓	
2. aschern			✓																			✓											✓		✓
3. Hitachi	✓		✓	✓	✓	✓	✓						✓									✓		✓			✓						✓		
4. Solomon			✓											✓								✓									✓		✓		
5. newsSweeper	✓		✓					✓	✓																		✓								
6. NoPropaganda	✓	✓																																	
7. Inno	✓		✓	✓						✓	✓						✓						✓												
8. CyberWallE	✓																							✓				✓						✓
10. Duth	✓																							✓											
11. DiSaster	✓											✓					✓								✓								✓		

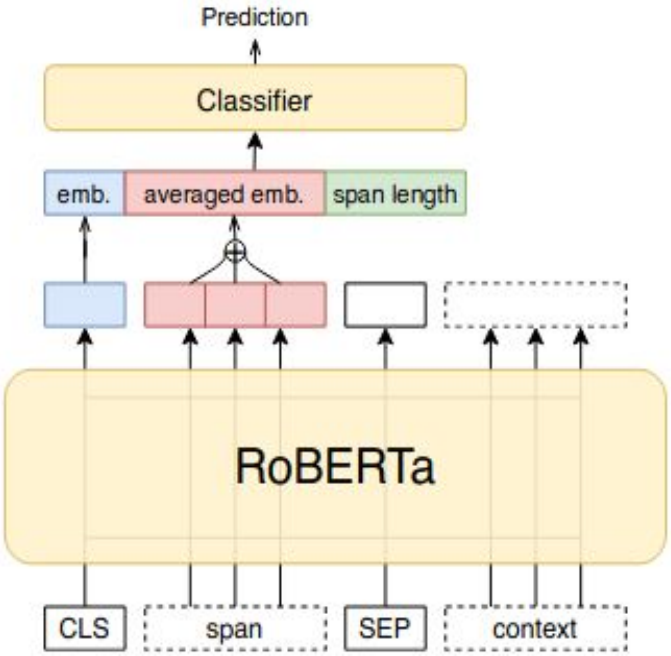
Rnk	Team	Overall
1	ApplicaAI	**62.07**
2	aschern	62.01
3	Hitachi	61.73
4	Solomon	58.94
5	newsSweeper	58.44

PRTA: Propaganda techniques

One of the top models available



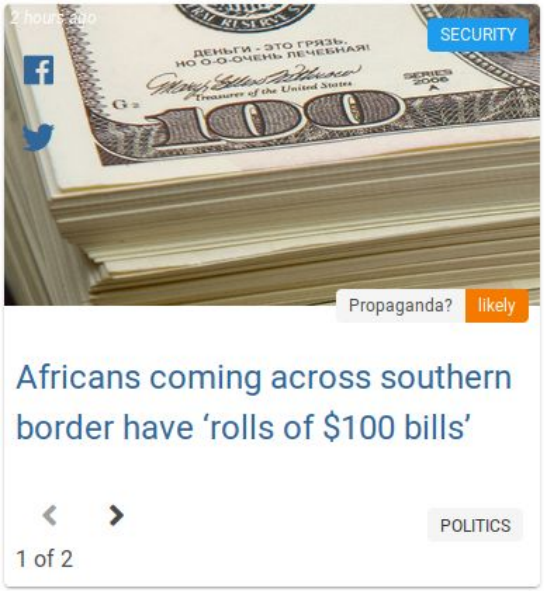
Span identification



Technique classification

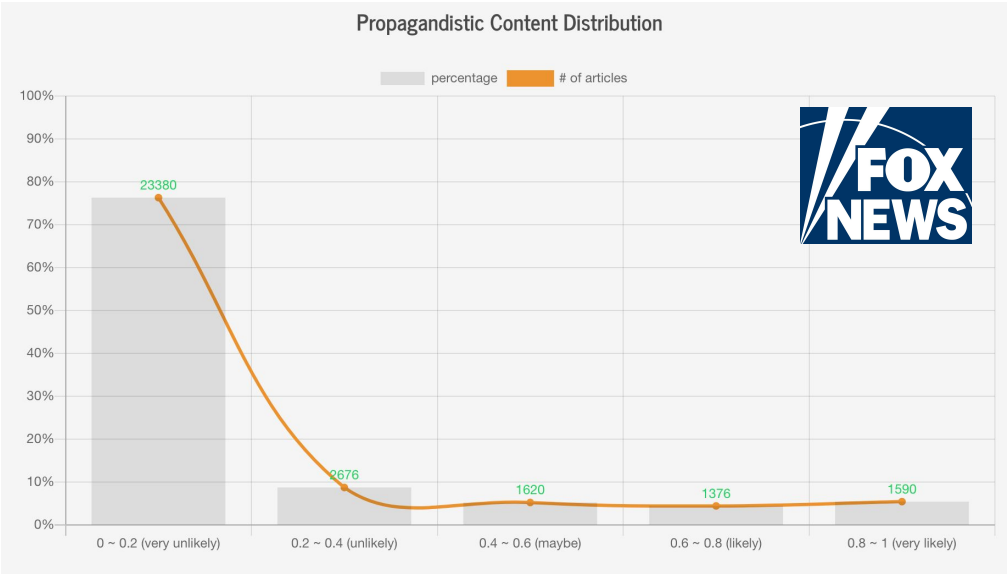
Chernyavskiy, Ilvovsky, Nakov.
Aschern at SemEval-2020 Task 11: It Takes Three to Tango: RoBERTa, CRF, and Transfer Learning
SemEval 2020

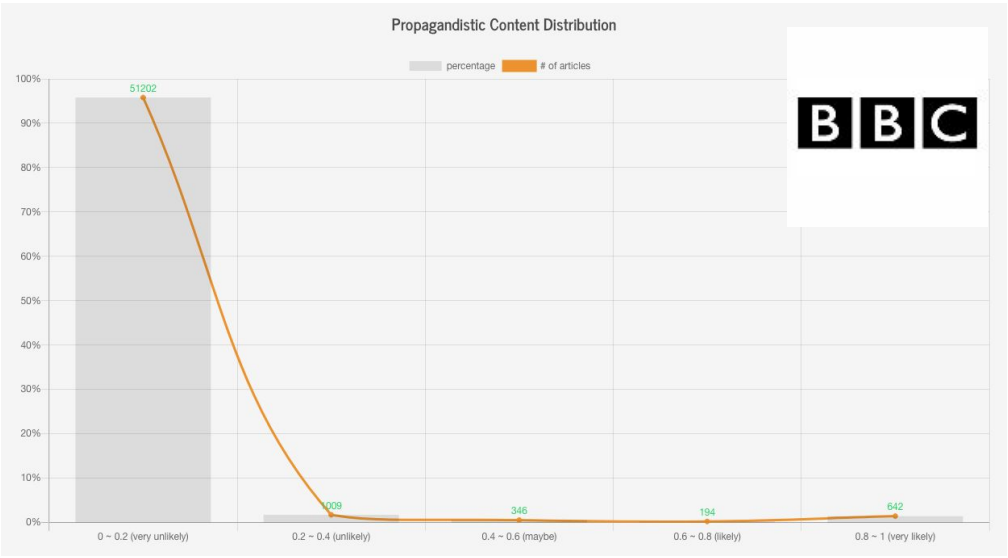
TANBIH أخبار تنبيه includes PPY and PRTA



Snapshot from

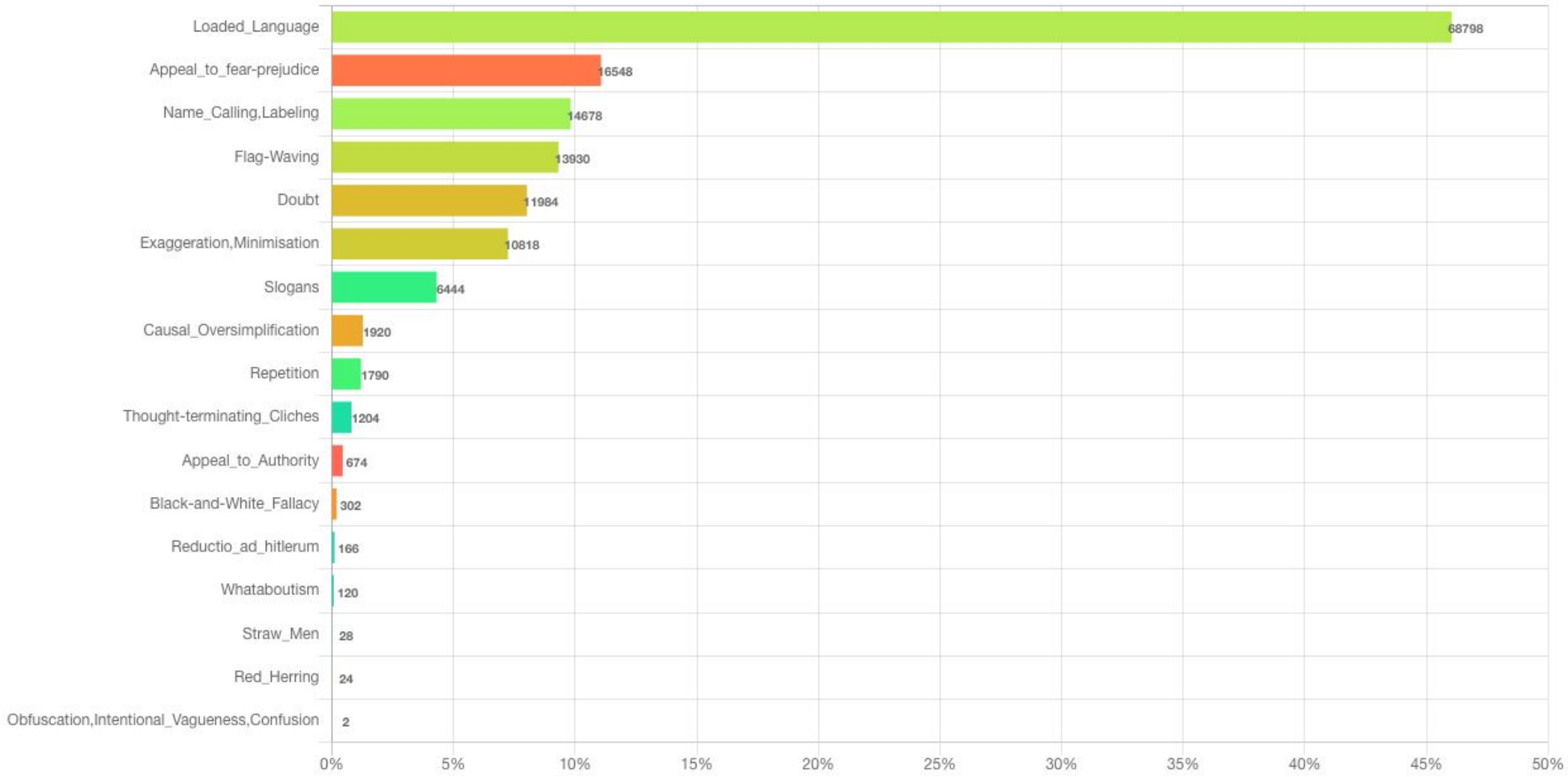
17 June, 2019





Stats as on 7 June, 2021

TANBIH أخبار تنبيه includes PRPPY and PRTA



Distribution of propaganda techniques in the news

(snapshot from 7 June, 2021)

TANBIH أخبار تنبيه includes PRPPY and PRTA

Can AI flag disease outbreaks faster than humans? Not quite

Global health authorities are increasingly using artificial intelligence to track emerging epidemics

☑ 6 - Doubt (?)
☑ 9 - Loaded Language (?)

- Can AI flag disease outbreaks faster than humans?[6] Not quite

- John Brownstein is co-founder of HealthMap, a system using artificial intelligence to monitor global disease outbreaks. John Brownstein is co-founder of HealthMap, a system using artificial intelligence to monitor global disease outbreaks. Photo: Steven Senne / Associated Press Photo: Steven Senne / Associated Press Image 1 of / 3 Caption Close Can AI[6] flag disease outbreaks faster than humans?[6] Not quite 1 / 3 Back to Gallery

- BOSTON — Did an artificial intelligence system beat human doctors in warning[6] the world of a severe[9] coronavirus outbreak in China?

- In a narrow sense, yes. But what the humans lacked in sheer[9] speed, they more than made up in finesse.

https://abcnews.go.com/Business/wireStory/ai-flag-disease-outbreaks-faster-humans-69094703

TANBIH أخبار تنبيه includes PRPPY and PRTA

- For the Democrats, the[5] question to Bork or not to Bork now stares them[5] in the face as plainly as [5] Adam Schiff's beady eyes.[9]

- To Bork[15] or Not to Bork,[15] a reference to Ted Kennedy's vilification of[9] Ronald Reagan's Supreme Court Nominee Robert Bork. Even as far back as 1987, the Democrats used fear tactics and public humiliation to[9] defeat an opponent. And now in the Age of Impeachment, that strategy is the Democratic Party's modus operandi.

- The Impeachment managers held the Senate hostage for a week. Hakeem Jeffries accused them of pushing conspiracies while detailing his own concocted Trump Derangement Syndrome[15] Madness.[9] While Zoe Lofgren declared that a global pandemic resulting in full blown chaos and[9] Economic Collapse are a back burner issue to the Democrats' rabid Borking[9] of our duly elected President —[10] especially now that there is a new case in the Bay area crap hole and[9] we are finding out that Coronavirus may be spread by feces.

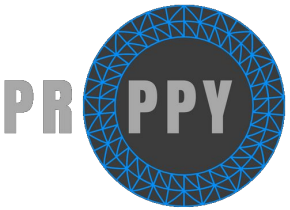
- Schiff's bold-faced lies[9] aggravated the Senators to the point that Lindsey Graham had given up on any sense of decorum. But when the vote to end the Democrats insistence for more witness came to its inevitable conclusion of failure, Senate Minority Leader Chuck Schumer demanded that Justice Roberts vote on the resolution, and was handed his lunch.

- Now, as the Brits[10] celebrate their long-fought battle for Brexit, America prepares to watch the revenge of[9] the Senate as the impeachment managers will endure their well-deserved scolding by [9] a captive and insulted Senate on[10] Monday and Tuesday. And President Trump will most certainly address the elephant in the room on[7] the eve of his acquittal.

By the way, people who know what's coming are taking advantage of our healthy & delicious storable food!

- [x] 5 - Causal/Oversimplification (?)
- [x] 7 - Exaggeration, Minimisation (?)
- [x] 9 - Loaded Language (?)
- [x] 10 - Name Calling, Labeling (?)
- [x] 15 - Slogans (?)

https://www.newswars.com/the-democrat-smackdown-in-the-senate/



/ PRTA

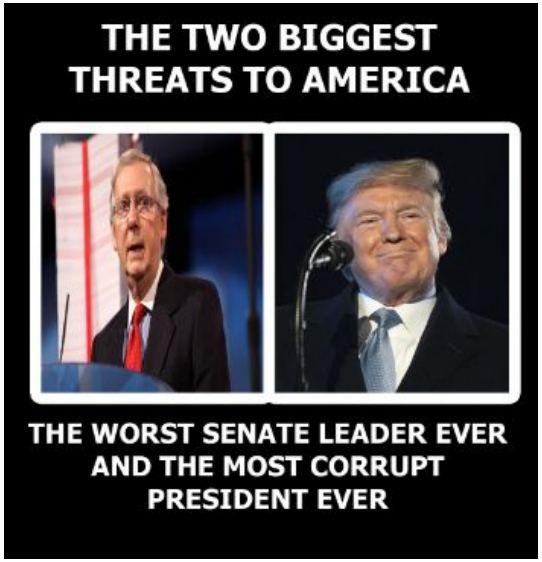
See proppy, prta and others at 

https://www.tanbih.org

https://www.tanbih.org/propaganda

Into memes

Appeal to fear

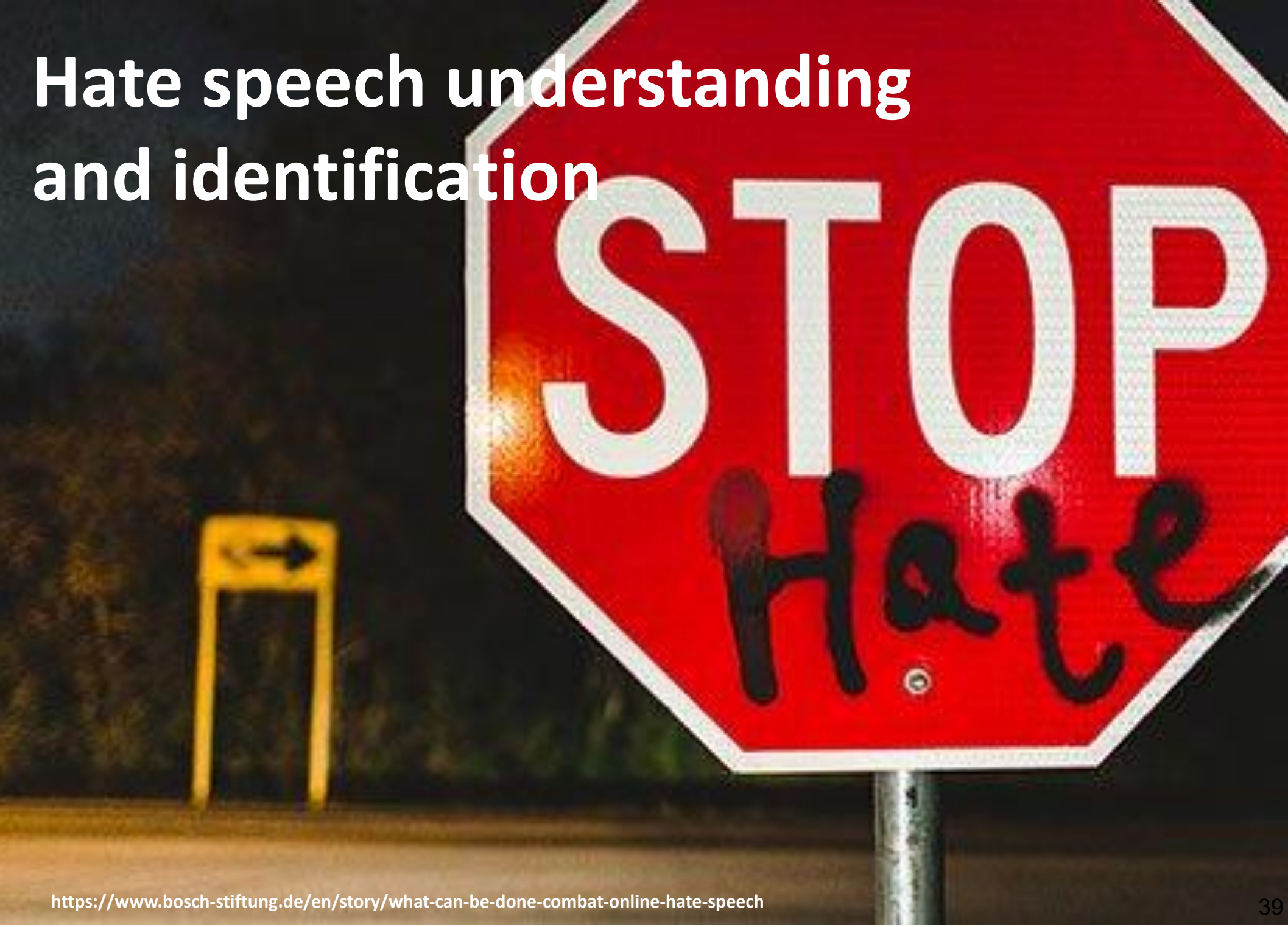


Name calling

SEMEVAL 2021 TASK 6 ON "DETECTION OF PERSUASION TECHNIQUES IN TEXTS AND IMAGES"

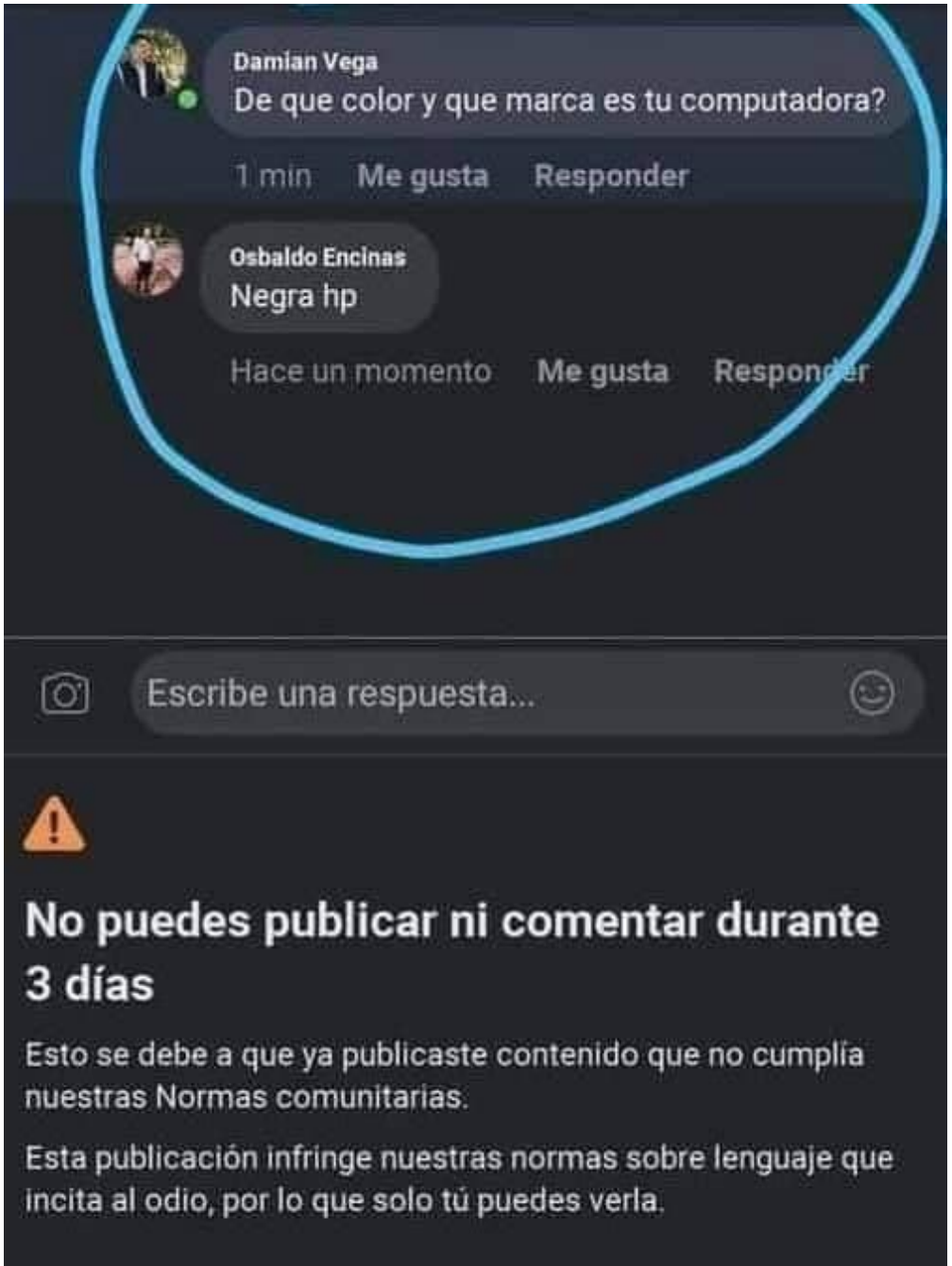
Dimitrov et al.
SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images
SemEval 2021

Hate speech understanding and identification



https://www.bosch-stiftung.de/en/story/what-can-be-done-combat-online-hate-speech

Hate speech filters don’t work (yet?)



(from twitter)

What is hate speech is not clear

http://www.quickmeme.com/meme/35knay

- it refers to how feminists, who for some reason look like hippies (generalisation), are manipulative and consider only what is to their advantage

- it refers to feminists being hysterical and not being able to hold a conversation

misogyny

What is hate speech is not clear

https://twitter.com/ww2girl1944/status/1236756839234183170

- it imposes a certain appropriate image on women
- it tells you that one is good and one is bad
- In the end women should be able to choose whatever they want to be!

(internalised) misogyny

Is hate speech annotation reliable?

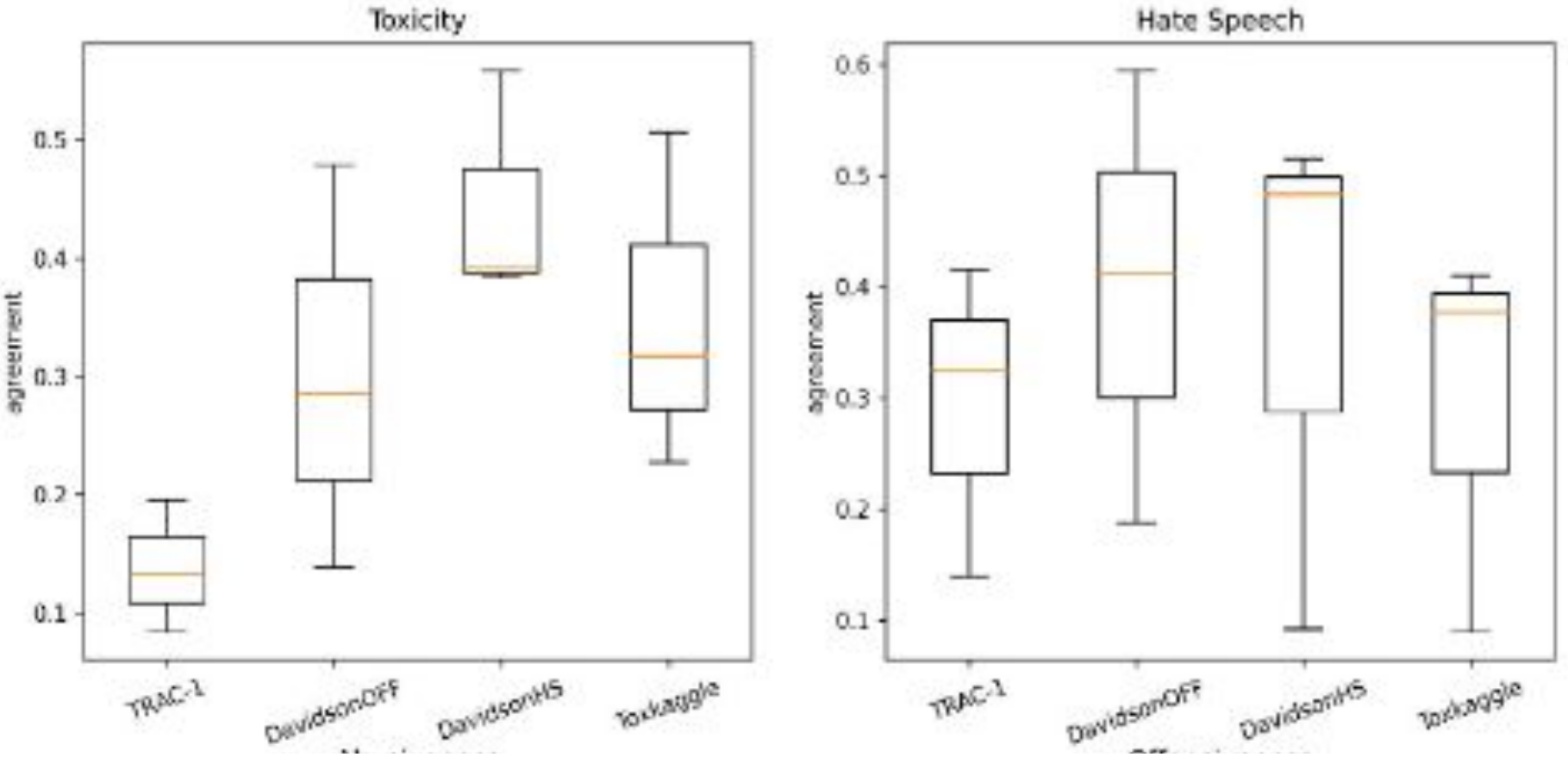
1. Take datasets previously annotated for toxicity, hate speech, abusiveness, offensiveness by explicit crowdsourcing

2. Re-annotate them by explicit crowdsourcing

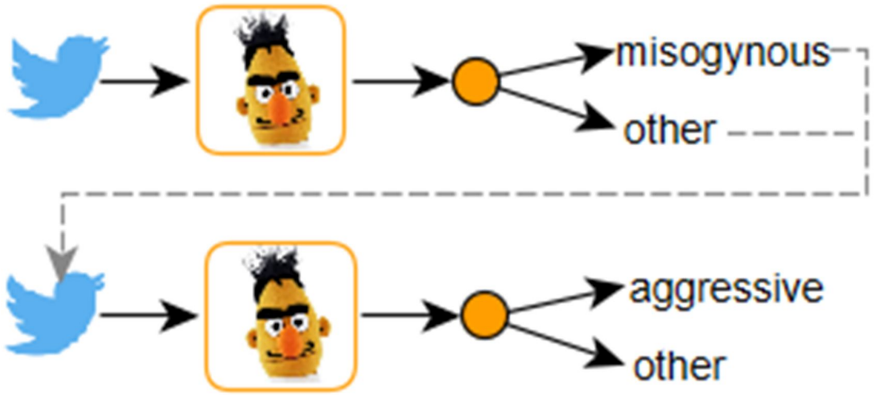
3. Evaluate the correlation between original and new annotation

Is hate speech annotation reliable?

Current outcomes



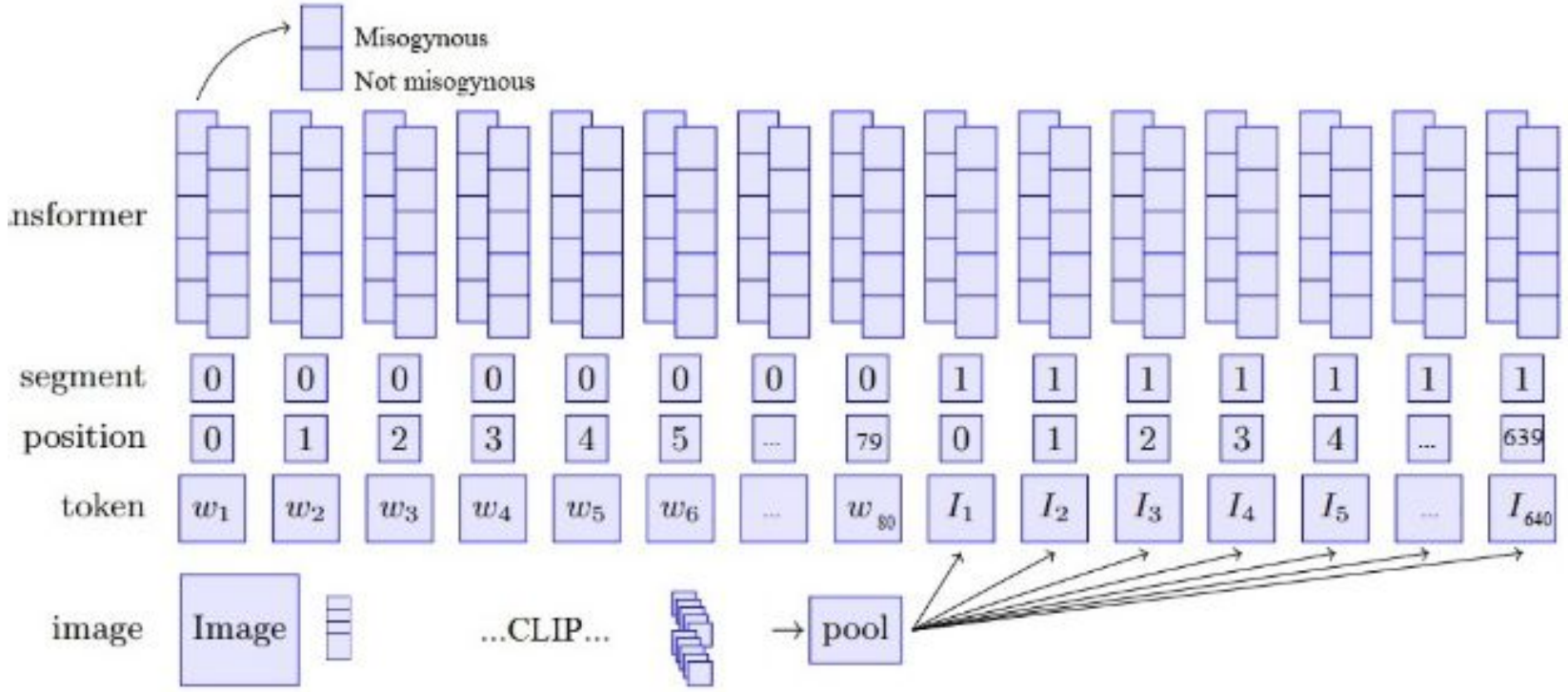
Identifying misogyny in tweets



Top-performing approach in Evalita 2020 (Italian)

Muti and Barrón-Cedeño
UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo
Evalita 2020

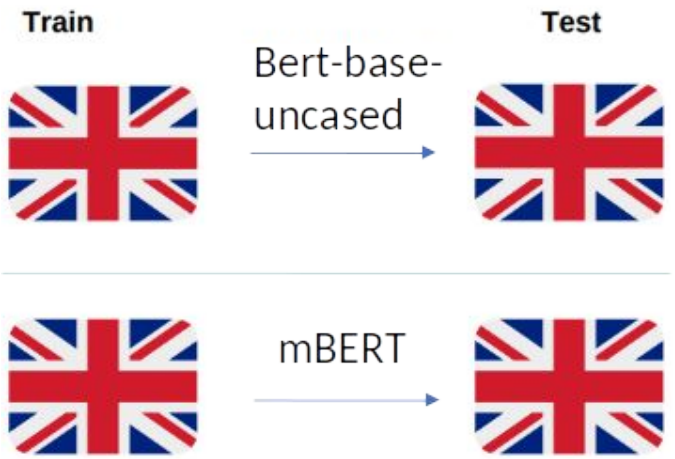
Identifying misogyny in memes

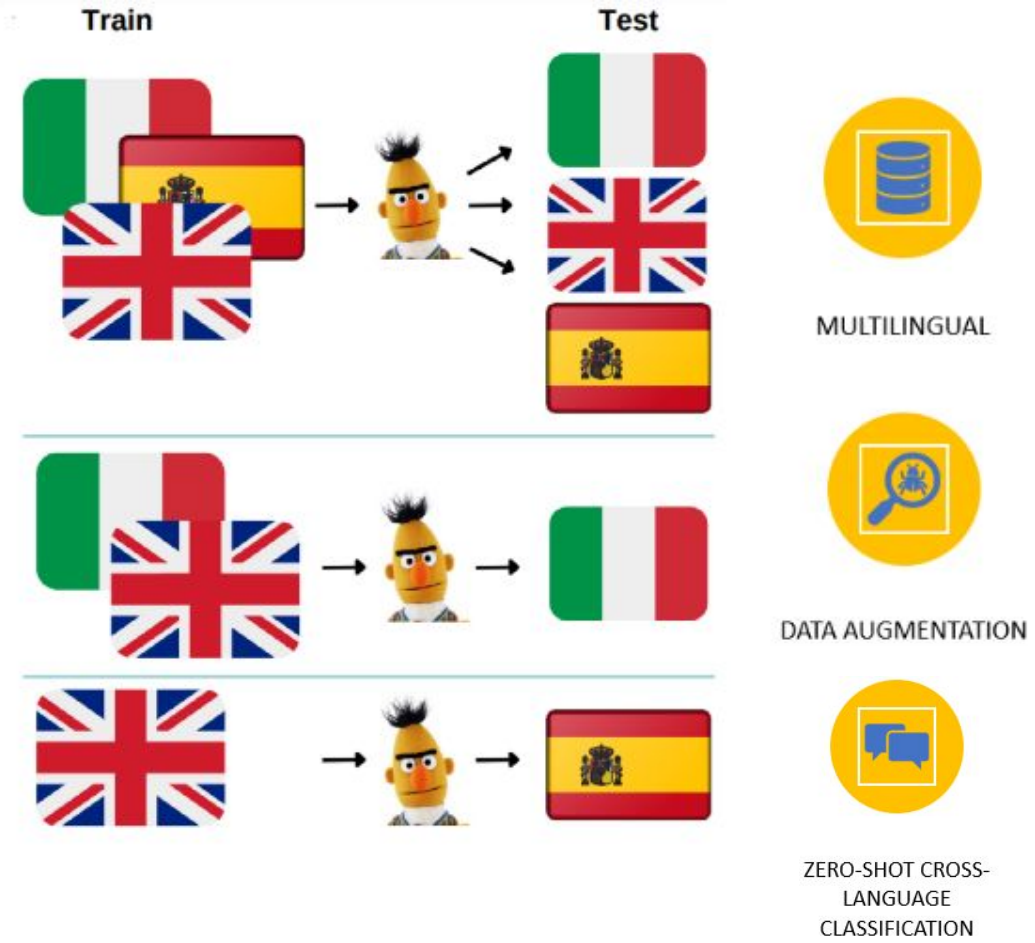


Top-5 out of 80+

Muti, Korre and Barrón-Cedeño
UniBO at SemEval-2022 Task 5: A Multimodal bi-Transformer Approach to the Binary and Fine-grained Identification of Misogyny in Memes
SemEval 2022

Identifying misogyny in multiple languages





Muti and Barrón-Cedeño
A Checkpoint on Multilingual Misogyny Identification
ACL 2022: Student Research Workshop

Identifying misogyny in multiple languages

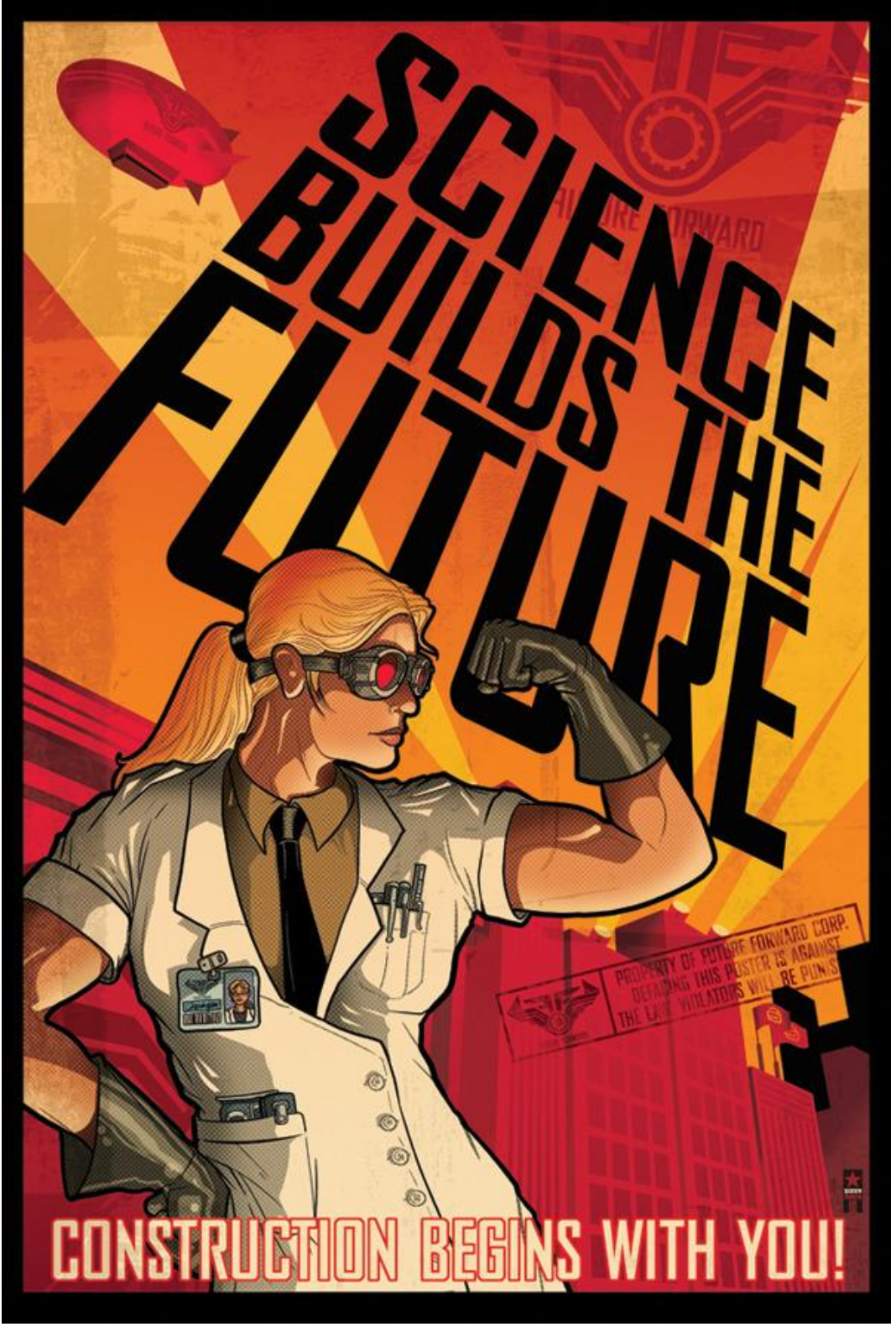
train	en	es	it	all
BERT en	**0.71**	–	–	–
BERT es	–	**0.85**	–	–
BERT it	–	–	**0.87**	–
mBERT en	0.65	0.14	0.17	–
mBERT es	0.62	0.81	0.50	–
mBERT it	0.47	0.63	0.87	–
mBERT en-es	0.67	0.83	–	0.75
mBERT en–it	0.66	–	0.86	0.77
mBERT es–it	–	0.80	0.86	0.84
mBERT en–es–it	0.68	0.82	0.86	0.78
best-AMI	0.70	0.81	0.84	–

Monolingual models trained with monolingual data perform better;

Multilingual models trained with monolingual data perform slightly worse

Adding data in one or two other languages impacts negatively wrt BERT, but positively wrt mBERT

Muti and Barrón-Cedeño
A Checkpoint on Multilingual Misogyny Identification
ACL 2022: Student Research Workshop



https://24.media.tumblr.com/932d312edd0fed799381c78fc38bfdfb/tumblr_miydsqINQv1qae9seo1_1280.jpg

Stuff we are doing/wondering now

Is ~~fake news~~ hate speech propaganda?

https://twitter.com/ww2girl1944/status/1236756839234183170

Black-and-white fallacy

Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an the extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship).

Fake news is today what spam was 20 years ago?

Spam	Fake news/propaganda
Hits when one person out of thousands clicks	Hits when turned viral (and reaches thousands)
90% accuracy (terrible false positives)	80% accuracy (much lower for techniques)
Succeeds when people does not catch it	Succeeds when people does not catch it

None has been solved

Propaganda in spam email

propaganda > 0.90

Dear Valued Candidate,

You were recently nominated as a iographical candidate for the next edition of Who's Who In America. We are pleased to inform you that the first phase of your candidacy was approved! Your prompt response is needed to ensure your complete professional information is considered. The office of the Managing Director appoints individuals ased upon a candidate's current position, and usually with information otained from researched executive and professional listings. The Director thinks that you may make an interesting iographical suject, a sindividual achievement is what Professional Who Who is all aout.

Propaganda in spam email

propaganda > 0.90

Not sure if you knew this but your website untrouled.org has some problems that you might want to consider looking into. I spent 2-3 minutes looking around and found: I visited your website and figured out I'd reach out to you and let you know there's serious room for dead easy (and affordale) improvement. If you would like, we can send youl We can develop the wesite on a more advanced platform at an affordale price. That price also includes making itcomplete moile responsive which will support all modern devices including all ranges of screen sizes.AAP-S: This is one time email and you may ask us to ""REMOVE"" you from our mailing list.

Propaganda in spam email

Analyse the use of propaganda in spam

Propaganda in images (perhaps not too much NLP)

© Ansa

Depp-Heard, cosa fanno dopo il processo? Johnny spende 60mila euro al ristorante indiano, Amber farà la mamma tutta l'estate

https://www.corrieredellosport.it/news/attualit/2022/06/07-93593298/depp-heard_dopo_il_processo_lui_spende_60mila_euro_al_ristorante

In-document factuality-based re-ranking

Adding complementary information to the title of an article

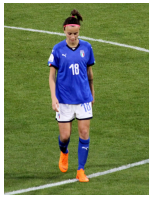
Zooming into “Arianna’s algorithm”

L’algoritmo di Arianna: “Così su Twitter do la caccia ai post contro le donne”

← Catchy title (+ journalist signature)

BOLOGNA. L’ultima vittima è stata **Barbara Bonansea**, la giocatrice della Juventus e della nazionale entrata nella top XI Fifa del 2020, tra le migliori al mondo. Che dopo l’annuncio, si è vista arrivare in rete una serie feroce di commenti sessisti. La colpa? Quella di saper tirare in porta. Come i maschi, anzi, meglio di gran parte di loro. Ennesima conferma che l’odio corra sul web e che spesso abbia tra le predestinate le donne. Che fare? Nell’attesa che gli odiatori si estinguano (potrebbe volerci parecchio tempo), la soluzione l’ha scovata **Arianna Muti**, fresca di laurea magistrale in Language, Society and Communication all’Alma Mater.

← Barbara Bonansea (Juventus player), victim of online misogyny; *Arianna* found the solution!



La 26enne di Osimo, residente sotto le Torri, ha infatti ottenuto la miglior performance nella sezione “Automatic Misogyny Identification” dell’ultima edizione di Evalita, iniziativa dedicata allo sviluppo di sistemi NLP (Natural Language Processing, ovvero di elaborazione del linguaggio) per la lingua italiana. E lo ha fatto ideando un algoritmo in grado di identificare e rimuovere tweet con contenuti violenti e offensivi contro le donne.
«Nei miei studi di linguistica con il professor **Fabio Tamburini** ho approfondito le applicazioni dell’Intelligenza Artificiale al riconoscimento del linguaggio, sono poi da sempre impegnata contro il sessismo, quindi l’invito di Evalita era perfetto per me. Ma ciò che più mi ha convinto è stata la pubblicazione annuale della Mappa dell’intolleranza da parte dell’osservatorio Vox Diritti. Secondo i dati, nel 2020, l’anno della pandemia, mentre in generale il cosiddetto hate speech su Twitter è diminuito rispetto al 2019, i tweet misogini sono aumentati del 90%, finendo per

← Best model at the Evalita 2020 shared task on misogyny identification: “able to identify and delete tweets with violent or offensive contents against women”

essere il 49,91% del totale». In pratica, un tweet negativo su due è contro una donna. «A questo si aggiunge che Twitter - spiega la neolaureata - fa affidamento sulle segnalazioni degli utenti per rimuovere i messaggi d’odio, che in molti casi non vengono intercettati e forse non è neppure tanto nell’interesse del social. Per questo è importante trovare soluzioni che possano riconoscerli in automatico».

← *Arianna*’s background and motivations

Lei per arginare il fenomeno in team con il ricercatore dell’Unibo **Alberto Barrón-Cedeño** ha preso come modello di riferimento Bert, l’algoritmo di Google in grado di capire cosa cerca la gente (e dunque di profilare i nostri gusti), adattandone il funzionamento per captare l’accanimento contro le donne sul web. «A partire da un database di 5.000 tweet, il mio sistema è stato in grado di individuare in automatico quelli che contenevano messaggi aggressivi nei confronti delle donne. Vengono letti e analizzati dall’algoritmo e in seguito classificati in tre categorie: non misogino, misogino e non aggressivo, misogino e aggressivo». Certo esiste un margine di errore, ma nel 77% dei casi l’algoritmo di Arianna ci ha visto giusto, consentendole di stilare anche l’elenco dei lemmi più usati in senso sessista. Con al primo posto - tra quelli dicibili - schiaffi, gola, acida. Nel lessico intollerante, molte le frasi legate al sesso, ma c’è pure il body shaming e l’astio contro le donne che lavorano.

← Model description; incl, BERT, data, accuracy, and an error analysis (from the paper)

Per ora l’algoritmo è studiato per Twitter e parla solo italiano, dal momento che la brevità dei post rende più semplice l’applicazione, ma le possibilità sono innumerevoli. «Si può fare con altri social e lingue. Io sarei pronta anche a regalarlo a Twitter, che naturalmente ha ingegneri e programmatori migliori di me. Diciamo che volevo mostrare che l’odio in rete si può fermare. Anzi, si deve».

← Future potential; incl. languages or giving the model to twitter: “now the algorithm is under study by Twitter” (!)

https://bologna.repubblica.it/cronaca/2021/03/31/news/l_algoritmo_di_arianna_cosi_su_twitter_do_la_caccia_ai_post_contro_le_donne_-294565109/

Work well, so that comrade Stalin will say **thank you!** (1949)