Capturing Human Values during Controversies

IC^{2}S^2 Tutorial

July 17, 2023 @ Copenhagen, Denmark

Kyriaki Kalimeri
ISI Foundation
Ph.D. Brain and Cognitive Sciences
@KyriakiKalimeri

Giovanni Da San Martino
University of Padova
Ph.D. Computer Science
@giodsm

Yelena Mejova
ISI Foundation
Ph.D. Computer Science
@YelenaMejova

Oscar Araque
Universidad Politécnica de Madrid
Ph.D. Telecommunication Engineering
@oaraquei

Outline

Talk

Human & Moral Values (Kyriaki)

Propaganda, persuasion & coordinated behaviour (Giovanni)

Controversy detection (Yelena)

Do

MoralStrength: extracting moral values from text (Oscar)

PRTA: detecting manipulation/persuasion techniques (Giovanni)

RWC: network analysis for polarization/controversy (Yelena)

Outline

Talk

Slides will be posted on slideshare and tutorial website:

https://propaganda.math.unipd.it/ic2s2-tutorial/

Do

Download code & sample data:

https://github.com/oaraque/human-values-tutorial-ic2s2-2023/

Human and Moral Values

What is a Value?

Schwartz's Values

UNIVERSALISM
UNDERSTANDING, APPRECIATION, TOLERANCE AND PROTECTION FOR THE WELFARE OF ALL PEOPLE AND FOR NATURE.

BENEVOLENCE
PRESERVATION AND ENHANCEMENT OF THE WELFARE OF PEOPLE WITH WHOM ONE IS IN FREQUENT PERSONAL CONTACT.

TRADITION
RESPECT, COMMITMENT AND ACCEPTANCE OF THE CUSTOMS AND IDEAS THAT TRADITIONAL CULTURE OR RELIGION PROVIDE THE SELF.

CONFORMITY
RESTRAINT OF ACTIONS, INCLINATIONS AND IMPULSES LIKELY TO UPSET OR HARM OTHERS AND VIOLATE SOCIAL EXPECTATIONS OR NORMS.

SECURITY
SAFETY, HARMONY, AND STABILITY OF SOCIETY, OF RELATIONSHIPS, AND OF SELF.

POWER
SOCIAL STATUS AND PRESTIGE, CONTROL OR DOMINANCE OVER PEOPLE AND RESOURCES.

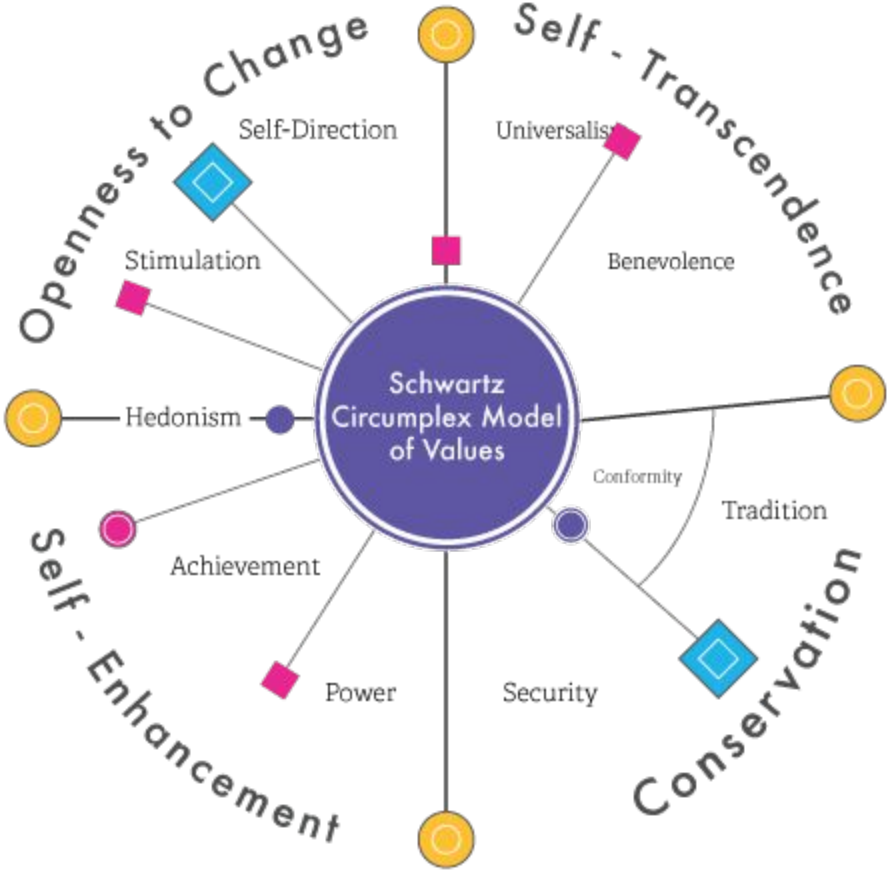
ACHIEVEMENT
PERSONAL SUCCESS THROUGH DEMONSTRATING COMPETENCE ACCORDING TO SOCIAL STANDARDS.

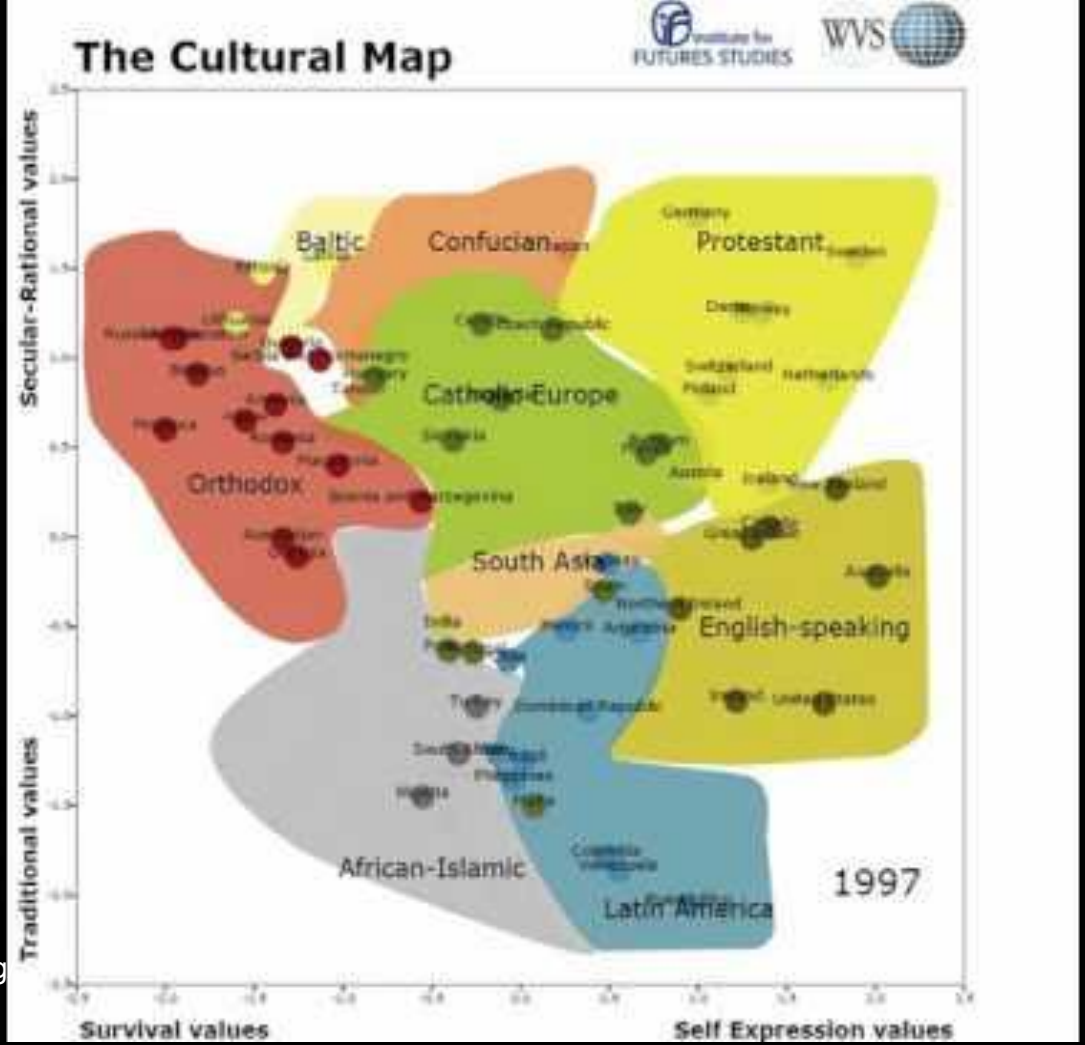
HEDONISM
PLEASURE AND SENSUOUS GRATIFICATION FOR ONESELF.

STIMULATION
EXCITEMENT, NOVELTY AND CHALLENGE IN LIFE.

SELF-DIRECTION
INDEPENDENT THOUGHT AND ACTION - CHOOSING, CREATING, EXPLORING.

“Values” are broad motivational goals!





http://www.worldvaluessurvey.org

Moral Values

Moral Foundations Theory

Care/Harm: virtues of caring and compassion.

Fairness/Cheating: unfair treatment, inequality, notions of justice.

Loyalty/Betrayal: obligations of group membership, loyalty, vigilance against betrayal.

Authority/Subversion: social order, obligations of hierarchical relationships such as obedience, respect

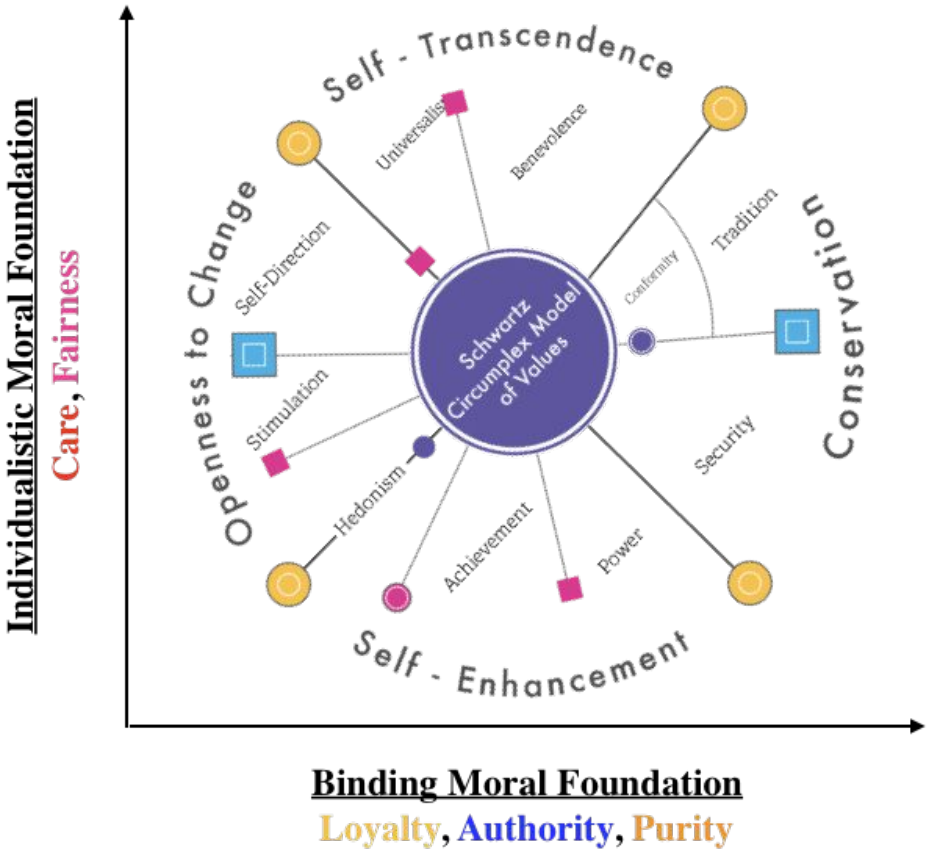
Purity/Degradation: physical and spiritual contagion, virtues of chastity, wholesomeness and control of desires.

Liberty/Oppression: feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty

Six key moral foundations

	 CARE/ HARM	 FAIRNESS/ CHEATING	 LOYALTY/ BETRAYAL	 AUTHORITY/ SUBVERSION	 SANCTITY/ DEGRADATION	 LIBERTY/ OPPRESSION
Adaptive challenge	Protect and care for children	Reap benefits of two way partnerships	Form cohesive coalitions	Forge beneficial relationships within hierarchies	Avoid contaminants	Keeping dominant individuals in the group 'in check'
Original triggers	Distress or neediness expressed by child	Cheating, co-operation, deception	Threat of challenge to group	Signs of dominance and submission	Waste products, diseased people	Bullying and constraining others
Key emotions	Compassion	Anger, gratitude, guilt	Group pride, rage against traitors	Respect, fear	Disgust	Anger at oppression
Relevant virtues	Caring, kindness	Fairness, justice, trustworthiness	Loyalty, patriotism, self-sacrifice	Obedience, deference	Temperance, chasity, piety, cleanliness	Freedom and self determination, protection of victims

Source: Johnathan Haidt The Righteous Mind


Individualistic Moral Foundation
Care, Fairness
Self - Transcendence
Universalism
Benevolence
Conservation
Tradition
Conformity
Openness to Change
Self-Direction
Stimulation
Schwartz Circumplex Model of Values
Hedonism
Achievement
Power
Security
Self - Enhancement
Binding Moral Foundation
Loyalty, Authority, Purity

Personality & Values

Values could be considered a higher level with respect to *Personality Traits.*

Moral values determine *how and when* dispositions and attitudes towards interpersonal and intergroup processes relate with our *life stories and narratives.*

Personality alone does not suffice to explain our judgments.

McAdams, D., & Pals, J. (2006). A new big five: Fundamental principles for an integrative science of personality. American Psychologist , 61 , 204

Attitudes Towards

Environmental: Climate Change

Healthcare: Vaccination

Philanthropy: Charitable Giving

Consumer Choices: Music Preferences

Ethics: AI and moral decision-making

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors

Christopher Wolsko *, Hector Ariceaga, Jesse Seiden

Oregon State University — Cascades, 2600 NW College Way, Cascades Hall, Bend, OR 97701, United States

Moral Narratives Around the Vaccination Debate on Facebook

Mariano Gastón Beiró, Universidad de Buenos Aires. Facultad de Ingeniería, Paseo Colón 850, C1063ACV, Argentina and CONICET, Universidad de Buenos Aires, INTECIN, Paseo Colón 850, C1063ACV, Argentina, **mbeiro@fi.uba.ar**

Jacopo D'Ignazi, ISI Foundation, Italy, **jacopo.dignazi@isi.it**

Victoria Perez Bustos, Universidad de Buenos Aires. Facultad de Ingeniería, Paseo Colón 850, C1063ACV, Argentina, **vperez@fi.uba.ar**

María Florencia Prado, Universidad de Buenos Aires. Facultad de Ingeniería, Paseo Colón 850, C1063 Buenos Aires, Argentina, Argentina, **mfprado@fi.uba.ar**

Kyriaki Kalimeri, ISI Foundation, Italy, **kkalimeri@gmail.com**

DOI: https://doi.org/10.1145/3543507.3583865

WWW '23: Proceedings of the ACM Web Conference 2023, Austin, TX, USA, April 2023

ISMIR 2022 BENGALURU

Schedule Papers Music LBDs Industry Jobs

P6-16: "More than words": Linking Music Preferences and Moral Values through Lyrics

Preniqi, Vjosa*, Kalimeri, Kyriaki, Saitis, Charalampos

Subjects (starting with primary): Human-centered MIR -> personalization ; Human-centered MIR -> user modeling ; Human-centered MIR ; Applications -> music recommendation and playlist generation ; MIR fundamentals and methodology -> lyrics and other textual data ; Human-centered MIR -> user behavior analysis and mining

nature

International journal of science

Article | Published: 24 October 2018

The Moral Machine experiment

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan

Nature **563**, 59–64 (2018) | Download Citation

ORIGINAL RESEARCH REPORT

Moral Framing and Charitable Donation: Integrating Exploratory Social Media Analyses and Confirmatory Experimentation

Joe Hoover*, Kate Johnson*, Reihane Boghrati†, Jesse Graham* and Morteza Dehghani‡

Moral Values Detection

Psychometric Surveys

Digital Traces

- Mobile Phone data (Browsing History, Location Data, Messaging, App, Email)
- Social Media Activity (Text, Music, Photos, Videos)

Values in the Wild

Why would you wear a mask during an epidemic?

https://www.theguardian.com/lifeandstyle/2021/aug/19/anti-masker-unlikely-friendship

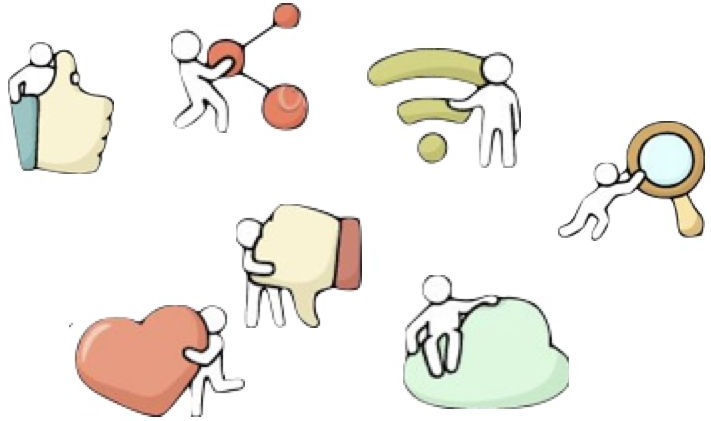
- What role do values play in the formation of our opinions?
- Can these values be manipulated to influence our opinions?
- What societal structures may correspond to the discussions around values?

Computers in Human Behavior
Volume 92, March 2019, Pages 428-445

Full length article

Predicting demographics, moral foundations, and human values from digital behaviours

Kyriaki Kalimeri [a], Mariano G. Beiró [b, 1], Matteo Delfino [a], Robert Raleigh [c], Ciro Cattuto [a]



7,000 participants
demographically representative
US sample
surveys & digital data

Mobile Apps & Web Searches are greatly informative of demographics

Moral & Human Values are much harder to predict, STILL, we get cool insights!

SnapChat

Moral Values

- Care

Schwartz Values

- Conservation
- Tradition
+ Openness
+ Self-Enhancement

Demographics

Low income

Low wealth

Not a Parent

Single

18 to 24 years old

Not a smoker

Morals reflected in Digital Non-verbal Behaviours

music preferences (Preniqi, Saitis, & Kalimeri 2022)

Facebook Likes & Questionnaires

well-being and behavioural interventions (Mejova and Kalimeri 2019)

(Questionnaires & Digital Mobile Data)

vaccine hesitancy (Kalimeri, Beiro, Urbinati, Cattuto 2019)

(Questionnaires & Facebook Pro/Anti Vax Pages Descriptions)

Moral Values Detection in Text

Lexicon Based Approaches

MFD - Liberals and Conservatives Rely on Different Sets of Moral Foundations

MoralStrength - Exploiting a moral lexicon and embedding similarity for moral foundations prediction (code: https://pypi.org/project/moralstrength/)

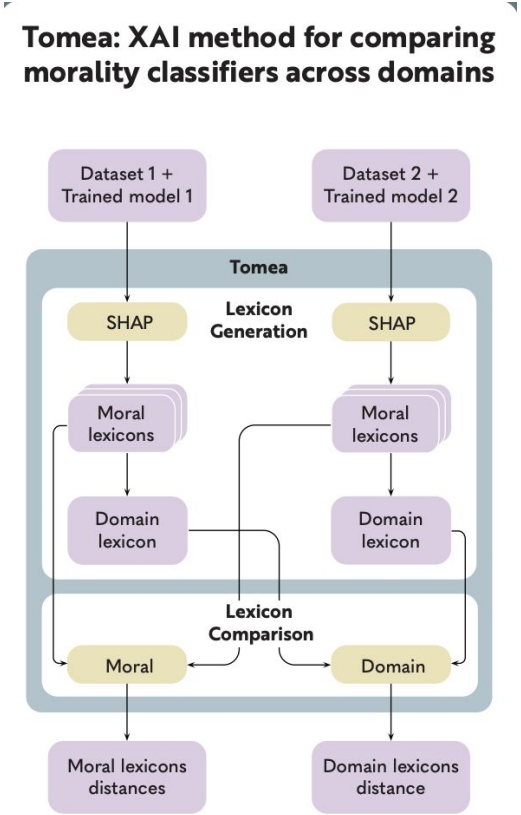
LibertyMFD - A Lexicon to Assess the Moral Foundation of Liberty.

eMFD - The Extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text

Fine-tuned Language Models

MoralBERT: Detecting Moral Values in Social Discourse

Domain Comparison of Moral Rhetoric



All Lives Matter (ALM), and Black Lives Matter (BLM) from Davidson et al. Automated Hate Speech Detection and the Problem of Offensive Language.

Tomea can shed light on how domain-specific language conveys morality,

'brotherhood' has a low impact on betrayal moral the ALM domain

but a considerably higher impact in BLM!

Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C., Kalimeri, K. and Murukannaiah, P.K., 2023, July. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14113-14132).

Lexicon Approaches

✔Interpretable
✔Efficiency
✔Domain-specific Adaptation

❌Limited Contextual Understanding
❌Difficulty in Handling Sarcasm
❌Lexicon Incompleteness

Large Language Models

✔Contextual Understanding
✔Ability to Handle Sarcasm and Irony
✔Generalization across Domains

❌Computational Requirements
❌Lack of Transparency
❌Ethical Concern

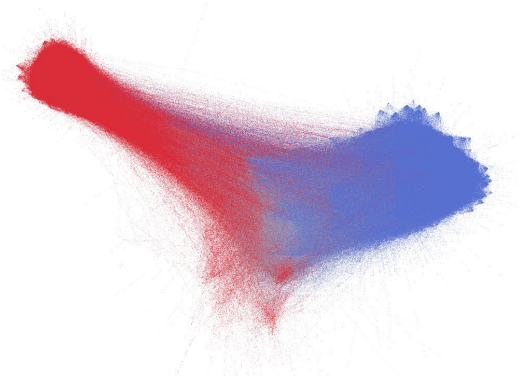
Decision making on Controversial Social Issues

Who is cooperating rather than competing in response to a crisis?

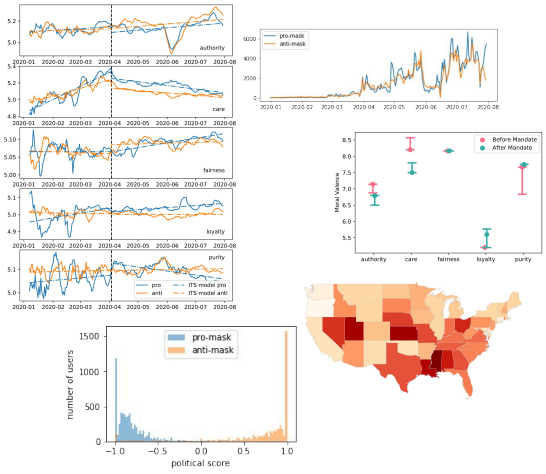
What role do our values play in the formation of our opinions?

Values in the Wild

→

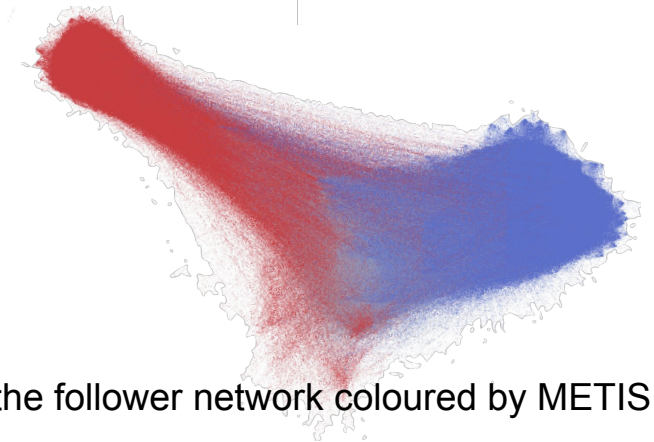


→

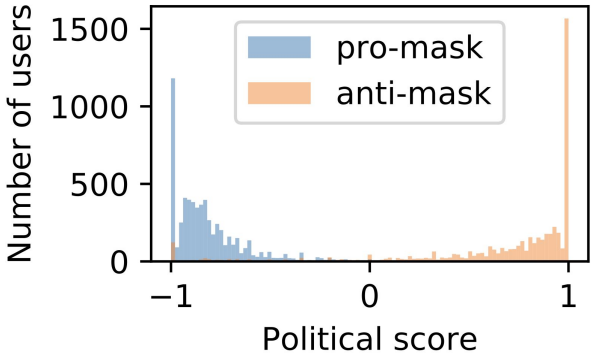


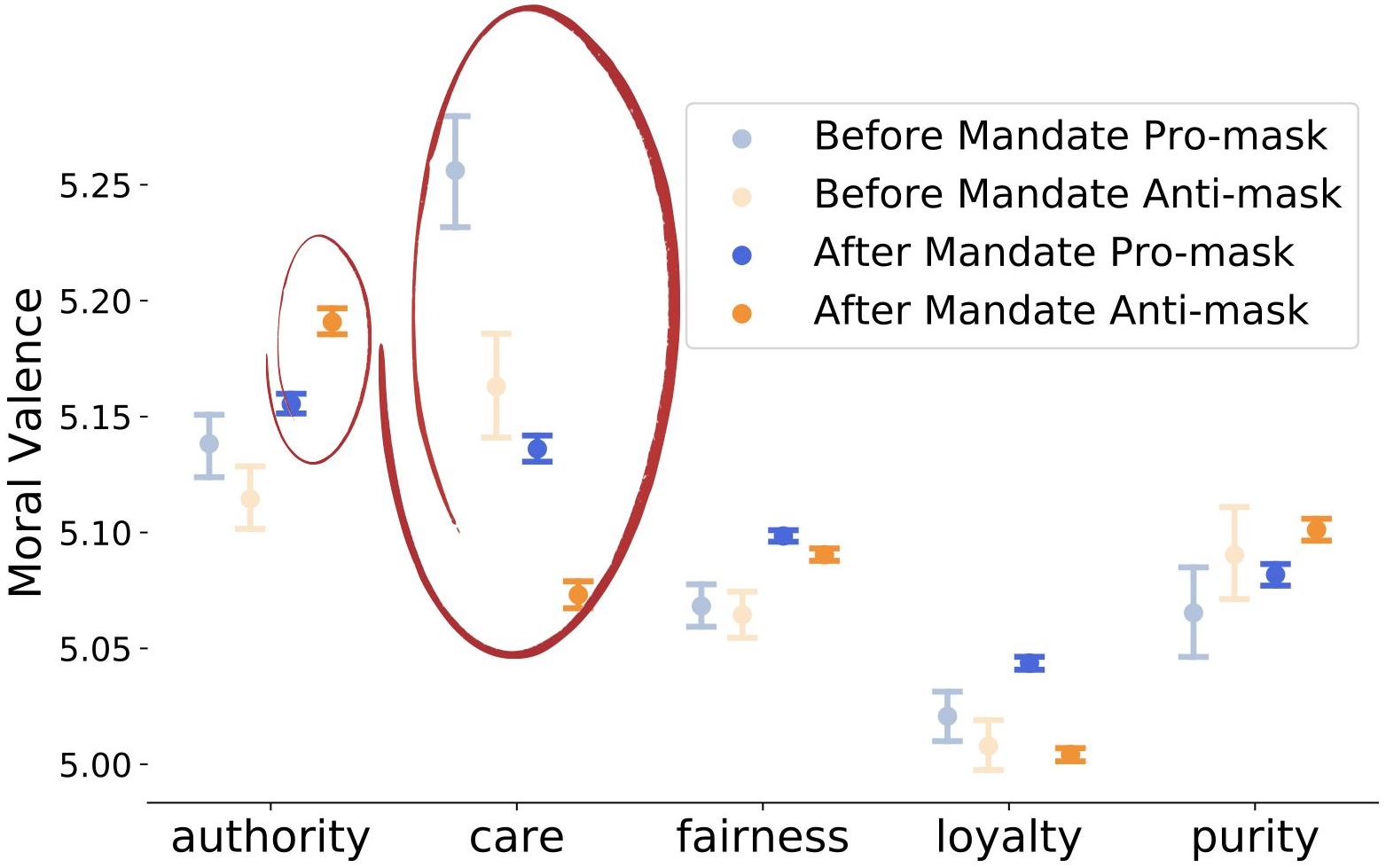
How can we analyze human values using computational tools applied to user-generated & news content?

Dataset stage	N tweets	N users
Keyword-based collection	18 245 298	5 935 103
Geo-location filter	5 685 866	1 383 729
Engagement & network filter	3 614 343	598 792



GCC of the follower network coloured by METIS score

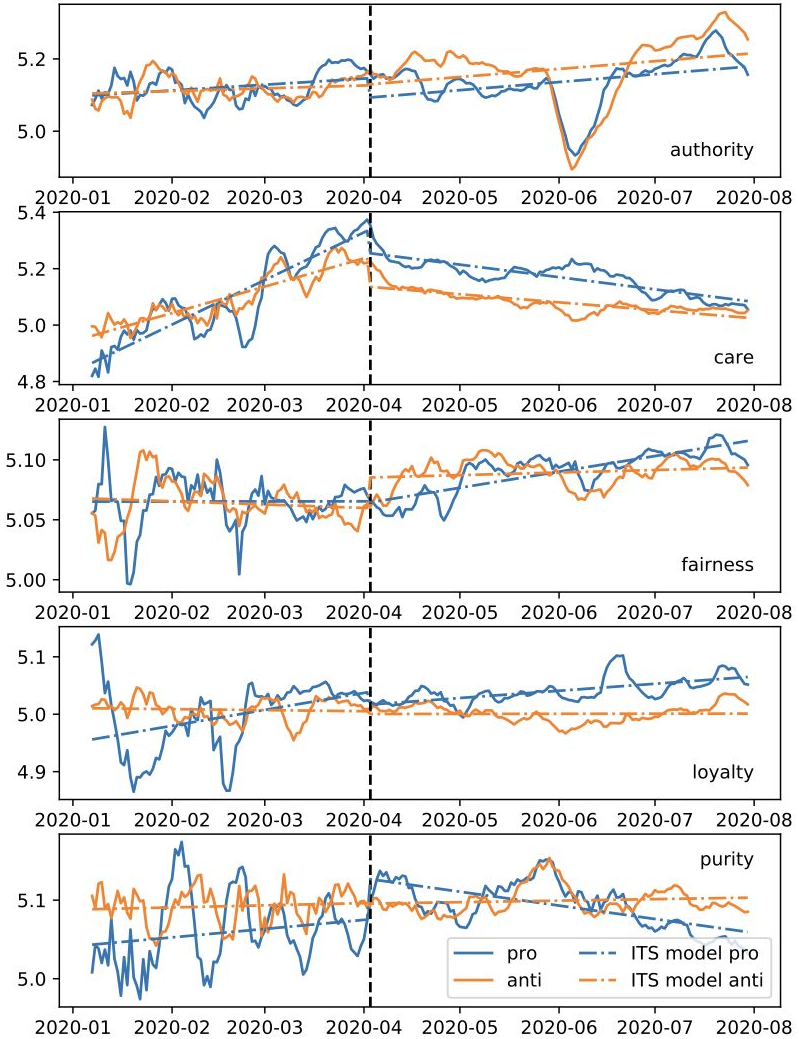



Moral Valence
5.25
5.20
5.15
5.10
5.05
5.00
authority
care
fairness
loyalty
purity
Before Mandate Pro-mask
Before Mandate Anti-mask
After Mandate Pro-mask
After Mandate Anti-mask

Evolution of moral values

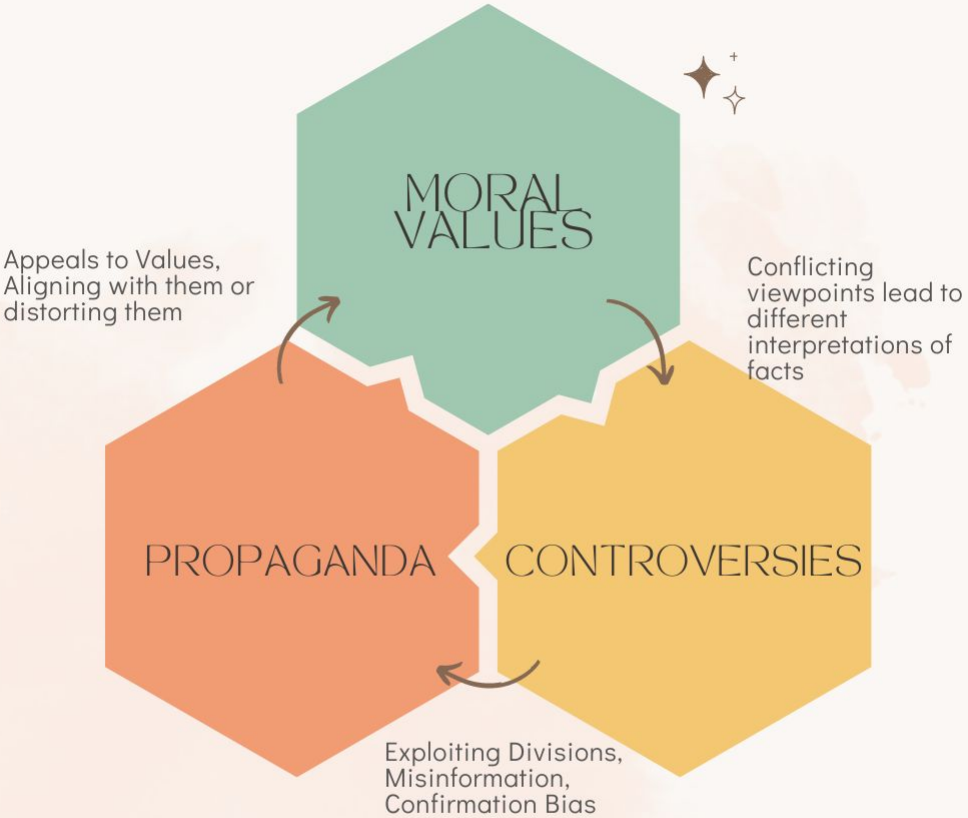
Interrupted Time Series Analysis

Loyalty progressively diverging



Who is more susceptible to disinformation and conspiracy theories?

Which moral values are expressed by those who proliferate conspiracy theories?


MORAL VALUES
Appeals to Values, Aligning with them or distorting them
Conflicting viewpoints lead to different interpretations of facts
PROPAGANDA
CONTROVERSIES
Exploiting Divisions, Misinformation, Confirmation Bias

Moral Values used in Propaganda

Propaganda often seeks to **appeal to** individuals' **moral values** and beliefs to influence their opinions and actions.

By **aligning with or distorting** moral values, propaganda can gain credibility, emotional appeal, and a sense of moral righteousness.

Propaganda may manipulate moral values by **selectively presenting information**, distorting facts, or framing issues in a way that aligns with a specific agenda.

Moral Values, Propaganda and Controversies

Propaganda frequently plays a role in **creating or intensifying controversies** by disseminating biased or misleading information.

Controversies arise when different groups or individuals hold conflicting viewpoints, often **fueled by differing interpretations of facts and values.**

Propaganda can exploit controversies by spreading disinformation, exaggerating divisions, or **demonizing opposing viewpoints,** thereby influencing public opinion and exacerbating conflicts.

Moral value assessment in natural language

Knowledge-Based Systems
Volume 191, 5 March 2020, 105184

MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction

Oscar Araque, Lorenzo Gatti, Kyriaki Kalimeri
Intelligent Systems Group, Universidad Politecnica de Madrid, Madrid, Spain
Human Media Interaction Lab, University of Twente, Enschede, The Netherlands
ISI Foundation, Turin, Italy

Research Question Can we predict the moral rhetoric in user-generated text?

MoralStrength Dictionary

(i)contains 5 times more lemmas with respect to the MFD (~1000)
(ii) expansion via WordNet including common use words
(iii) human annotations of "strength" in a Likert-Scale for all lemmas

Evaluation

We evaluated our framework on the **benchmark dataset Moral Foundations Twitter Corpus** which consists of 7 datasets of various topics and contains approximately **35,000 annotated tweets.**

We propose three approaches of increasing complexity which employ the MoralStrength lexicon to predict the moral rhetoric:

- **Moral Freq**: frequency counts of the lemmas
- **Moral Stats**: statistical summary of the lemmas
- **SIMON**: word embedding similarity based representations

Tutorial notebook

- The full code for the tutorial lives here:

https://github.com/oaraque/human-values-tutorial-ic2s2-2023

- Code for moral value assessment in natural language:

MoralValues/Moral-Value-Estimation.ipynb

MoralStrength

- **MoralStrength** is a Python module that allows us to assess moral values using the MoralStrength lexicon.

- We can install MoralStrength through pip:

  pip install moralstrength

- Available at pypi:
  - https://pypi.org/project/moralstrength/

Oscar Araque, Lorenzo Gatti, Kyriaki Kalimeri, MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction, Knowledge-Based Systems, Volume 191, 2020, 105184, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2019.105184.

0. Importing libraries

In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.sparse import hstack
from sklearn.preprocessing import minmax_scale
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.model_selection import cross_val_predict
from sklearn.linear_model import SGDClassifier
from sklearn.feature_extraction.text import CountVectorizer

from tqdm.notebook import tqdm

import moralstrength
from moralstrength.moralstrength import estimate_morals
```

1. Read Dataset

- We use the **Moral Foundations Twitter Corpus (MFTC)**, a dataset composed of Twitter messages, **manually annotated** with moral values
- This dataset contains several categories of data, corresponding to different campaigns or movements:
    - Hurricane Sandy
    - Baltimore Protest
    - All Lives Matter
    - **Black Lives Matter (BLM)**
    - 2016 Presidential Election

Read the dataset in JSON format

In [2]:
```
df = pd.read_json("BLM.json", orient="records")
df.head(10)
```

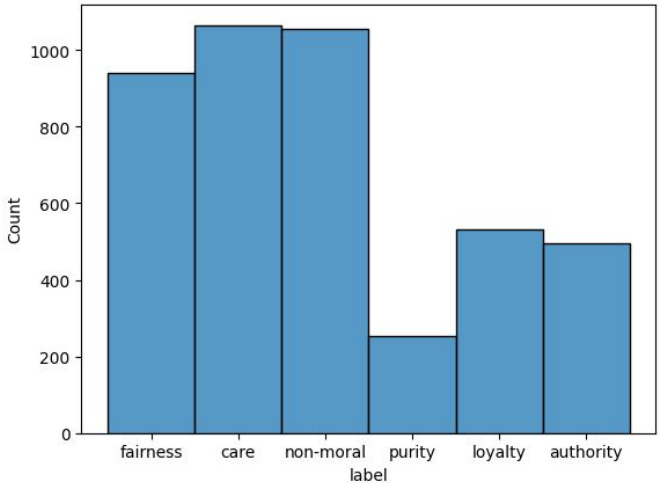
Out[2]:

	text	label
0	The courage to be impatient with evil and pati...	fairness
1	#NotAllCops but OMFG already. 😡 Protect and se...	care
2	stop shaving, it's your manly dignity #blackje...	non-moral
3	ARABS MORTAL HATRED AND ENSLAVEMENT OF THE BLA...	care
4	"@Babbsgirl2: #SheriffDavidClarke is my hero! ...	non-moral
5	Inciting Racial Fear, Hatred and Violence\n#Bl...	care
6	These killings show: 1.) racism 2.) a desensit...	care
7	Police try kindness #blacklivesmatter http://t...	care
8	@S_T_O_P_TERROR @DailyMirror #BLM GLOBALIST CO...	non-moral
9	@GrooveSDC @CaffeineAndHate #ISaluteWhitePeopl...	fairness

Some basic characteristics of the data

- Number of instances: **4,340 documents**
- Distribution of labels

```
In [5]:
1 sns.histplot(df["label"])
2 plt.show()
```



2. Extract the Moral Values from natural language

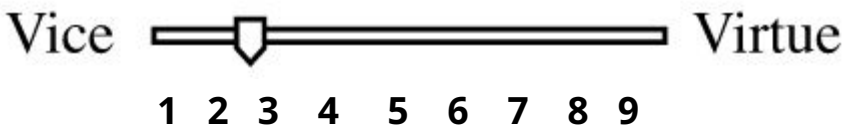
For each document, we want to extract the associated Moral Values. To do so, we use the estimate_morals function:

In [6]:
```
result = estimate_morals(df["text"], process=True)
result.head(10)
```

Out[6]:

	care	fairness	loyalty	authority	purity
0	4.000000	7.600000	NaN	NaN	1.857143
1	1.600000	NaN	NaN	4.8	7.750000
2	7.000000	NaN	NaN	NaN	NaN
3	NaN	3.666667	NaN	NaN	NaN
4	NaN	NaN	7.857143	NaN	NaN
5	1.666667	NaN	NaN	NaN	NaN
6	1.666667	4.000000	NaN	NaN	NaN
7	7.000000	NaN	NaN	NaN	NaN
8	NaN	NaN	3.250000	NaN	NaN
9	NaN	NaN	NaN	NaN	NaN

MoralStrength assessments are encoded following a Likert scale



The encoding range is **[1, 9]**, being a value of 5 interpreted as morally neutral.

We adapt the original scale using the following formula:

$$e' = | e - 5 |$$

For example,

The courage to be impatient with evil and patient with people, the courage to fight for social justice. @CornelWest #blacklivesmatter

has an associated Fairness score of **7.6**. In this case, e'=2.6.

Another example,

#CNN #Obama #Aclu #Ap #UN #BlackLivesMatter #Chicago Declare war on these racist delusional Killers Pure Hate and Lies #FBI #DOJ

has e=1.667 and e'=3.333

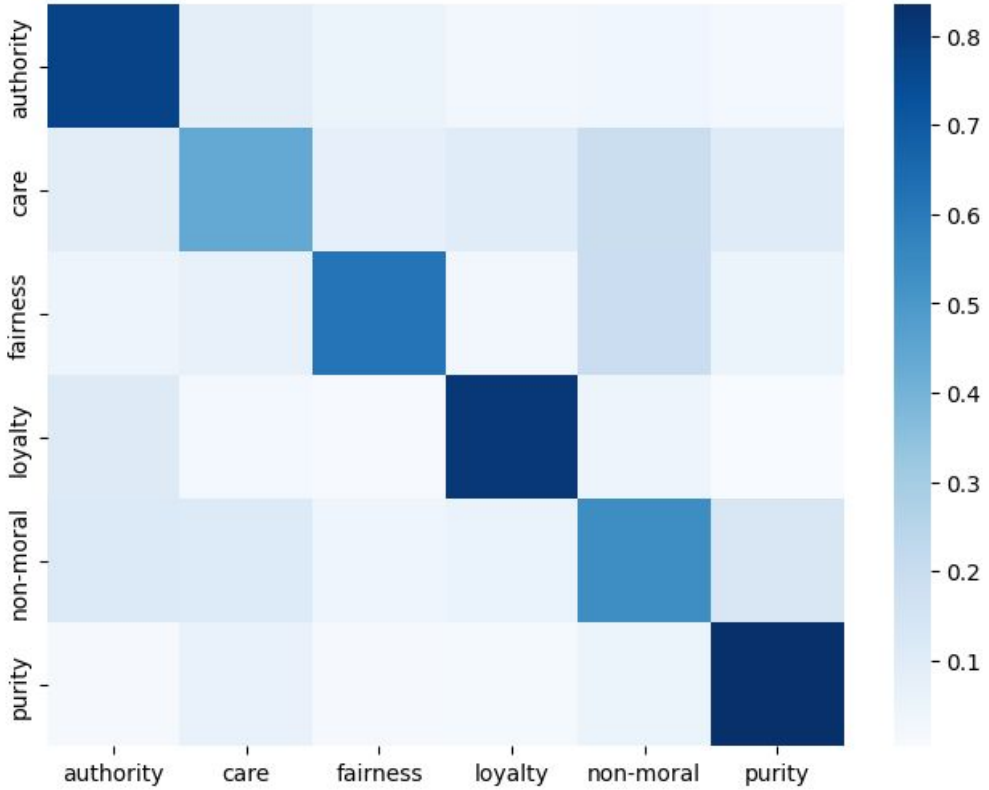
In [7]:
```
unsup_predictions = np.abs(result - 5).idxmax(axis=1).fillna("non-moral")
```

Now, we can use the adapted values as relevance for the moral foundation. Modelling the task as **presence classification** for each of the moral foundations.

In [8]:
```
print(
    classification_report(df["label"], unsup_predictions)
)
```

	precision	recall	f1-score	support
authority	0.54	0.78	0.63	494
care	0.65	0.44	0.52	1065
fairness	0.78	0.61	0.69	940
loyalty	0.67	0.81	0.73	531
non-moral	0.56	0.53	0.55	1056
purity	0.41	0.83	0.55	254
accuracy			0.61	4340
macro avg	0.60	0.67	0.61	4340
weighted avg	0.63	0.61	0.60	4340

Results for the **unsupervised** prediction of moral values



Confusion matrix for the unsupervised evaluation

3. Supervised classification of Moral Values

- Previously, we have shown how **MoralStrength** allows us to assess morality in natural language in an unsupervised manner.

- Now, we show how we can use these assessments to **enrich** text representations in supervised settings.

3.1 Pre-process text (simple)

In [10]:
```
1 texts_preprocessed = list()
2 for text in tqdm(moralstrength.nlp_reduced.pipe(df["text"])):
3     texts_preprocessed.append(text)
```

4340/? [00:02<00:00, 1800.39it/s]

- Uses **spacy** for preprocessing
- MoralStrength includes a pre-loaded reduced spacy model

3.2 Extract TF-IDF features

In [11]:
```
texts = list()
for doc in texts_preprocessed:
    text_i = list()
    for token in doc:
        text_i.append(token.lemma_)
    texts.append(" ".join(text_i))
```

In [12]:
```
unigram = CountVectorizer(max_features=10000)
unigram_features = unigram.fit_transform(texts)
unigram_features
```

Out[12]:
```
<4340x10000 sparse matrix of type '<class 'numpy.int64'>'
        with 54578 stored elements in Compressed Sparse Row format>
```

3.3 Evaluate the classifier

In [13]:

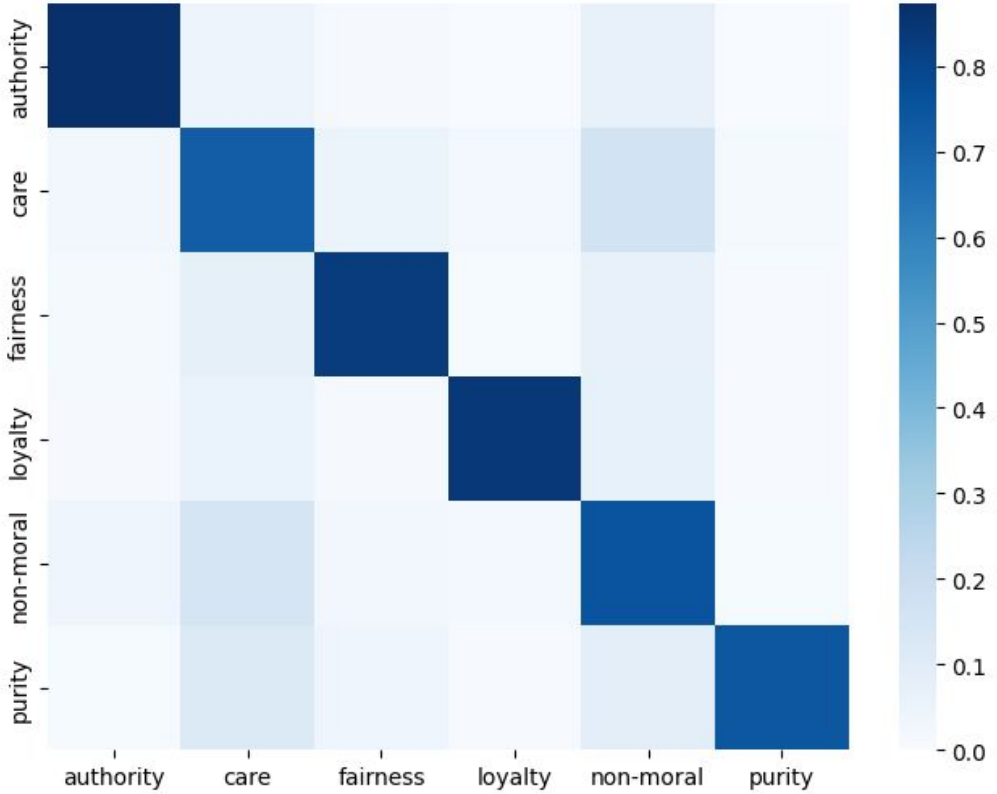
```
def get_classifier():
    return SGDClassifier(loss="hinge", random_state=42)

classifier = get_classifier()

sup_preds = cross_val_predict(
    classifier,
    unigram_features,
    df["label"],
    cv=10,
    n_jobs=-1
)

assert sup_preds.shape[0] == df["label"].shape[0]
```

In [14]:

```
print(
    classification_report(df["label"], sup_preds, digits=2)
)
```

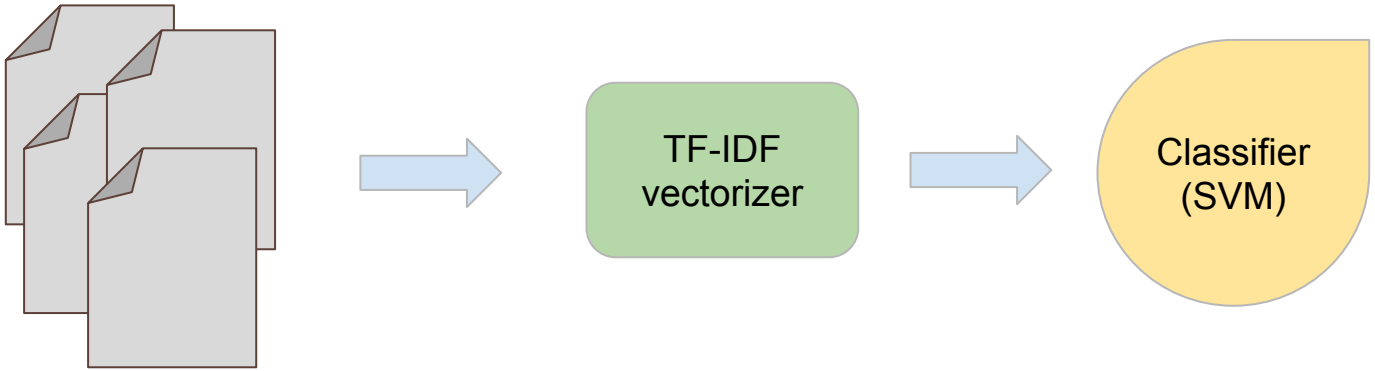
	precision	recall	f1-score	support
authority	0.83	0.87	0.85	494
care	0.71	0.72	0.72	1065
fairness	0.87	0.83	0.85	940
loyalty	0.90	0.84	0.87	531
non-moral	0.71	0.76	0.73	1056
purity	0.85	0.74	0.79	254
accuracy			0.79	4340
macro avg	0.81	0.79	0.80	4340
weighted avg	0.79	0.79	0.79	4340

Results for the baseline supervised classification



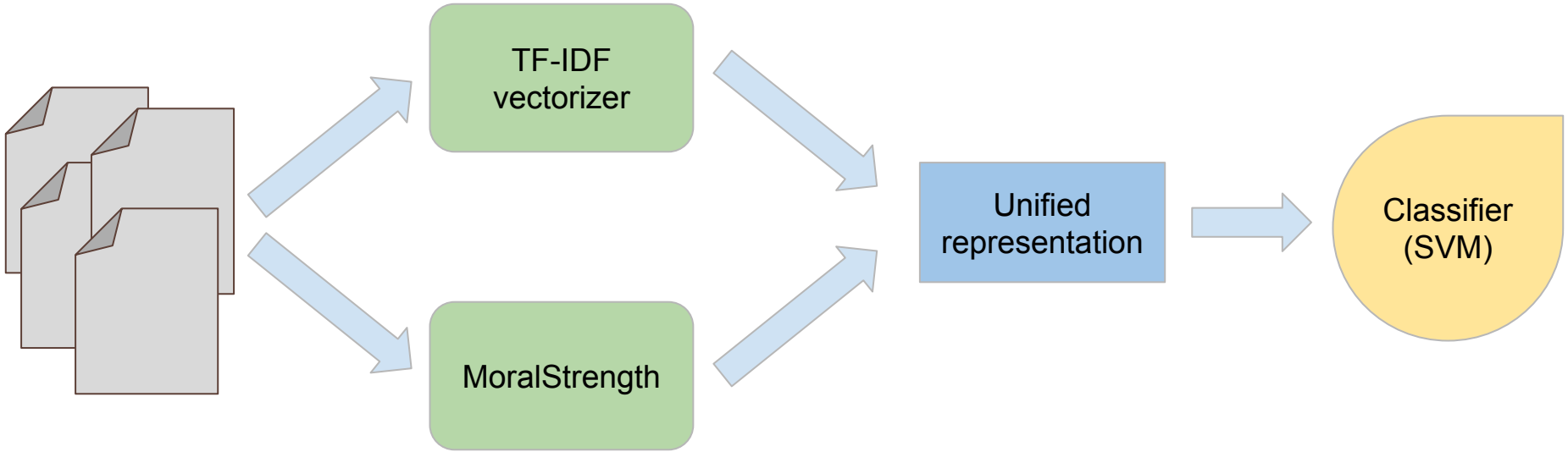
Confusion matrix for the baseline supervised classification

The defined learning model is as follows



It uses just **one mode** of information.

Now, let's evaluate a more complete model, combining the textual model with the extracted moral values



3.4 Including information from MoralStrength

In [16]:
```
1 moralstrength_features = minmax_scale((result.fillna(5.0) - 5).values)
2
3 combined_features = hstack([unigram_features, moralstrength_features])
4 combined_features
```

Out[16]:
```
<4340x10005 sparse matrix of type '<class 'numpy.float64'>'
        with 76246 stored elements in COOrdinate format>
```

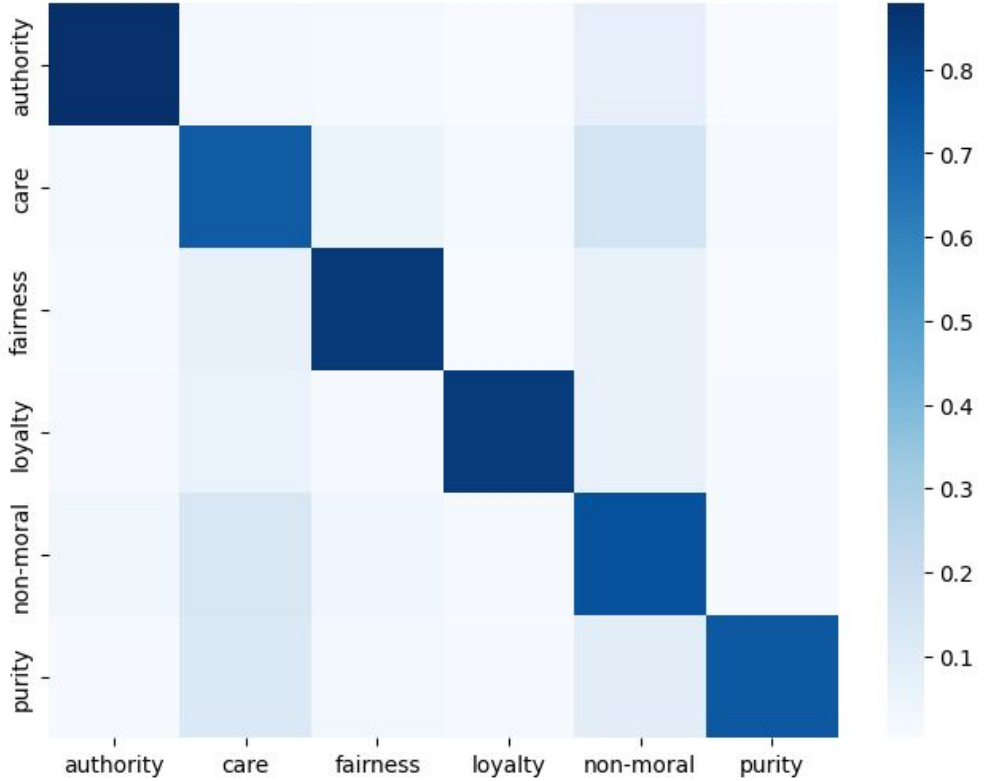
3.5 Evaluate the combined classifier

In [17]:

```
classifier = get_classifier()

sup_preds = cross_val_predict(
    classifier,
    combined_features,
    df["label"], cv=10, n_jobs=-1
)

assert sup_preds.shape[0] == df["label"].shape[0]
```

	precision	recall	f1-score	support
authority	0.83	0.88	0.86	494
care	0.73	0.73	0.73	1065
fairness	0.87	0.84	0.86	940
loyalty	0.92	0.84	0.88	531
non-moral	0.71	0.77	0.74	1056
purity	0.85	0.74	0.79	254
accuracy			0.80	4340
macro avg	0.82	0.80	0.81	4340
weighted avg	0.80	0.80	0.80	4340

Results for the combined supervised evaluation



Confusion matrix for the combined supervised evaluation

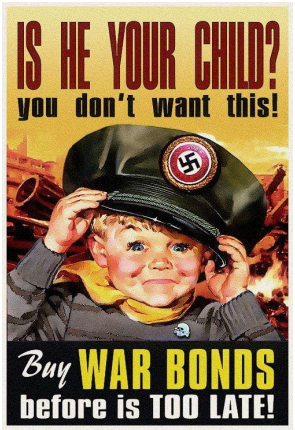
Conclusions

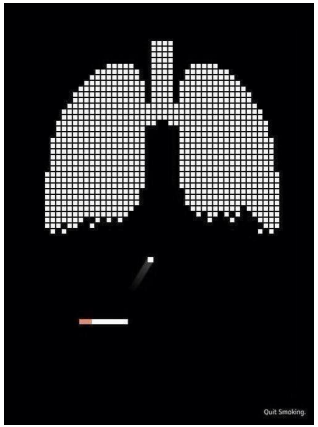
- **MoralStrength** allows us to perform unsupervised moral value analysis on textual data
  - Incorporating its own pre-processing mechanisms
- This information can be used to:
  - *Study morality* in an unsupervised manner, directly using the moral signals
  - *Enrich textual representations* in a machine learning system, adding them to additional information sources (e.g., sentiment or emotion)

Propaganda, persuasion & coordinated behaviour

Definitions: Propaganda

“Communications that deliberately misrepresent symbols, appealing to emotions and prejudices and bypassing rational thought, to influence its audience towards a specific goal”*





*definition re-elaborated from Institute for Propaganda Analysis (Ed.). (1938). How to Detect Propaganda. In Propaganda Analysis. Volume I of the Publications of the Institute for Propaganda Analysis (pp. 210–218).

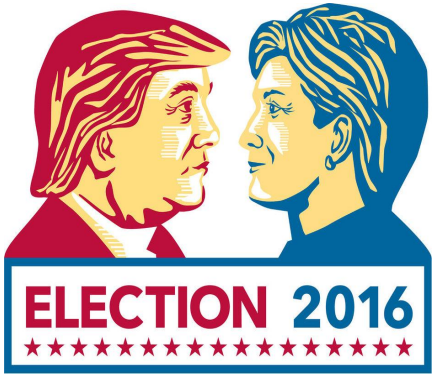
From pre-Internet Propaganda...

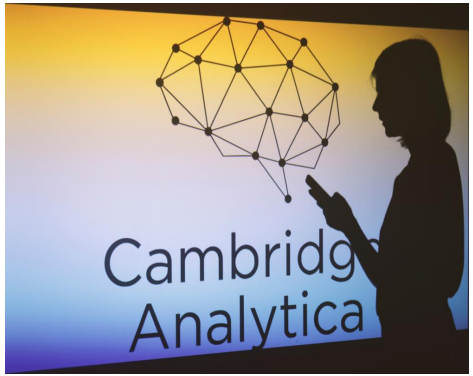
- Control of mass media
- Closed Borders (non-anonymous campaigns)
- Requiring massive resources

Bolsover, G., & Howard, P. (2017). Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda. Big Data, 5(4), 273–276.

...To Computational Propaganda

- "The rise of the **Internet** [...] has **opened** the creation and dissemination of **propaganda** messages, which were once the province of states and large institutions, **to a wide variety of individuals and groups**."





Bolsover, G., & Howard, P. (2017). Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda. Big Data, 5(4), 273–276.

Computational Propaganda Cookbook

- Different technical skills needed
  - Creating persuasive messages
  - Disseminating the messages (using bots)
  - Maximising audience reach
  - Microprofiling

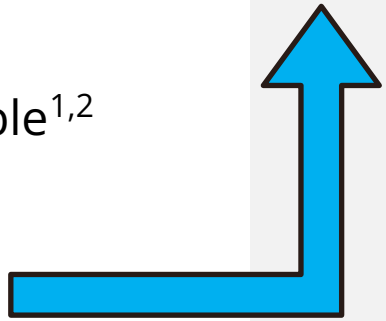
Propaganda: Document-Level Analysis

- Binary Classification Task: “Is a document propagandistic?” Yes, No
- Few datasets and Models available[1,2]
- Annotating documents might be controversial, requires experts (expensive)
- Lacks explainability

1 *Hannah Rashkin et al. Truth of varying shades: Analyzing language in fake news and political fact-checking. In EMNLP, pages 2931–2937, 2017*
2 *Alberto Barrón-Cedeño et al. Proppy: Organizing the news based on their propagandistic content. Inf. Process. Manag., 56(5):1849–1864, 2019*

Propaganda: Document-Level Analysis

- Binary Classification Task: “Is a document propagandistic?” Yes, No
- Few datasets and Models available[1,2]
- Annotating documents might be controversial, requires experts (expensive)
- Lacks explainability



Distant Supervision

- If a news source is propagandistic □ each of its articles is
- Risk of Modelling the source instead of the concept of Propaganda

1 *Hannah Rashkin et al. Truth of varying shades: Analyzing language in fake news and political fact-checking. In EMNLP, pages 2931–2937, 2017*
2 *Alberto Barrón-Cedeño et al. Proppy: Organizing the news based on their propagandistic content. Inf. Process. Manag., 56(5):1849–1864, 2019*

Persuasion Techniques Detection

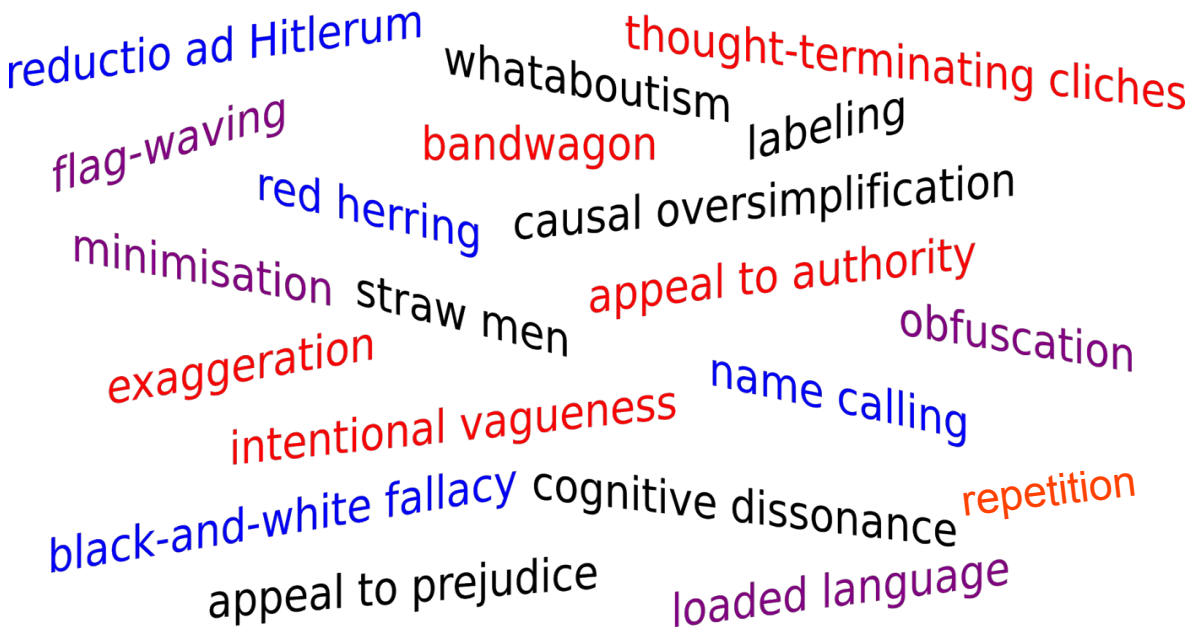
- Propaganda is conveyed through a series of rhetorical and psychological techniques

- The set of propaganda techniques differs between scholars[1], from

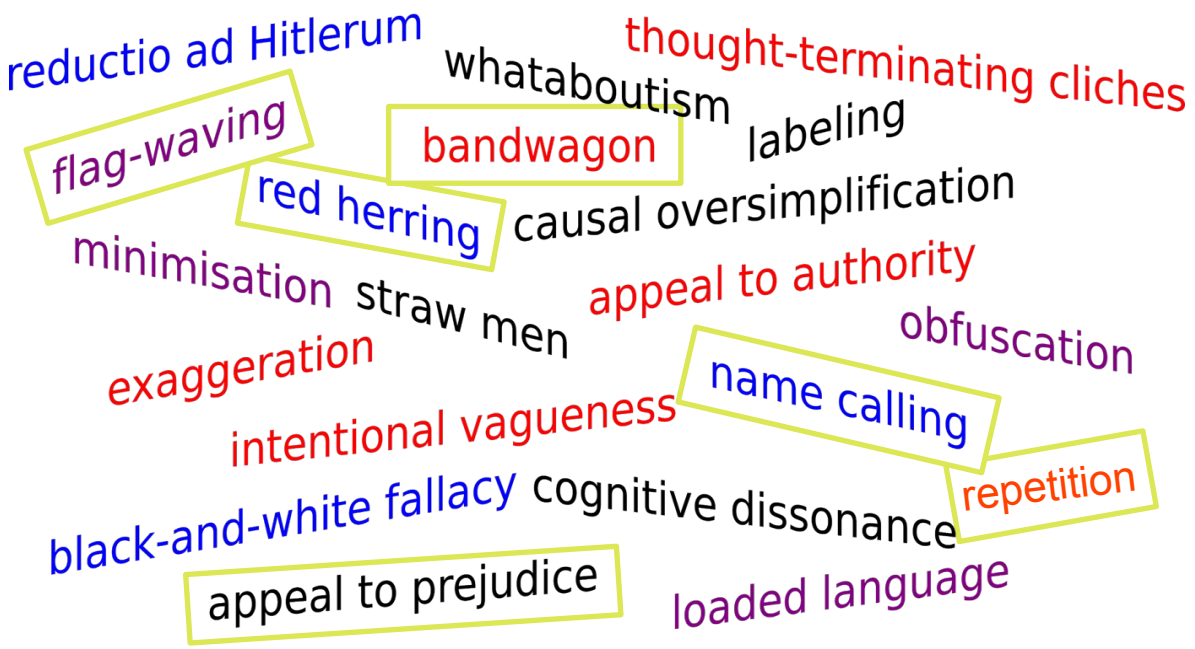
  - 7 of Miller[2] to
  - ~70 in Wikipedia[3]

1 *Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In Proceedings of the Australian Security and Intelligence Conference, pages 58–65, Perth, Australia.*
2 *Clyde R. Miller. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning.*
3 *http://en.wikipedia.org/wiki/Propaganda_techniques*

Persuasion Techniques



Persuasion Techniques




Weather: Sunny and warm, 79/62
SPORTS • FINAL
Monday, September 24, 2007
DAILY NEWS
50¢
New York's Hometown Newspaper
NYDailyNews.com
THE EVIL HAS LANDED
Mahmoud Ahmadinejad
STORIES ON PAGES 4-7
EDITORIAL PAGE 22

Name Calling


WOULD YOU VOTE FOR
CLINTON AS PRESIDENT?
57%
SAY YES

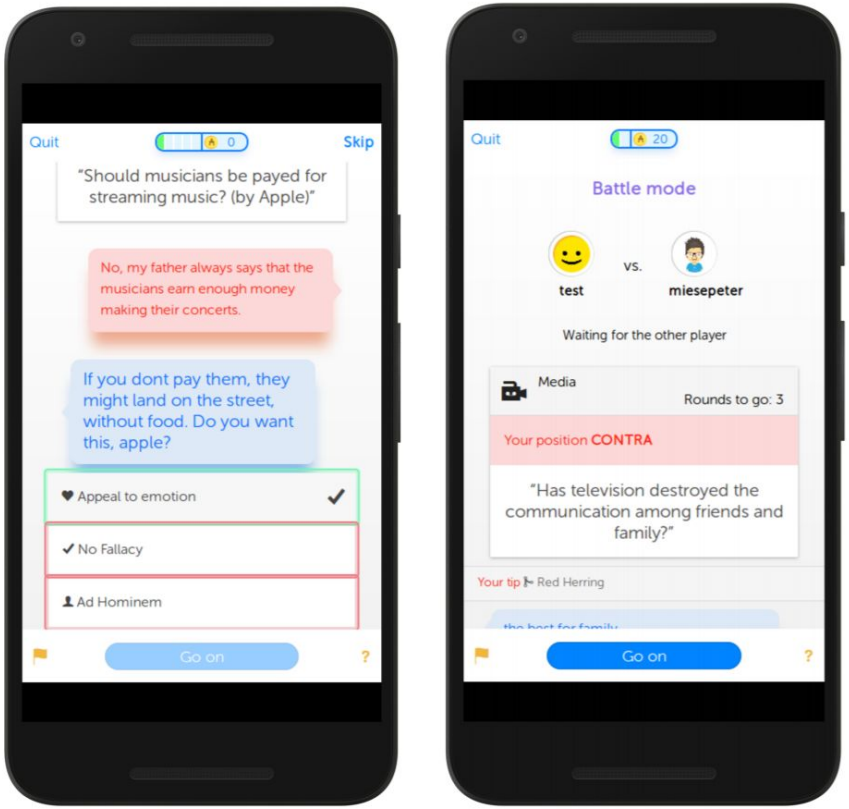
?

Bandwagon

Argotario

- A game to educate people to recognize and create fallacies. Users
  - recognise fallacies in others' arguments
  - write fallacious arguments

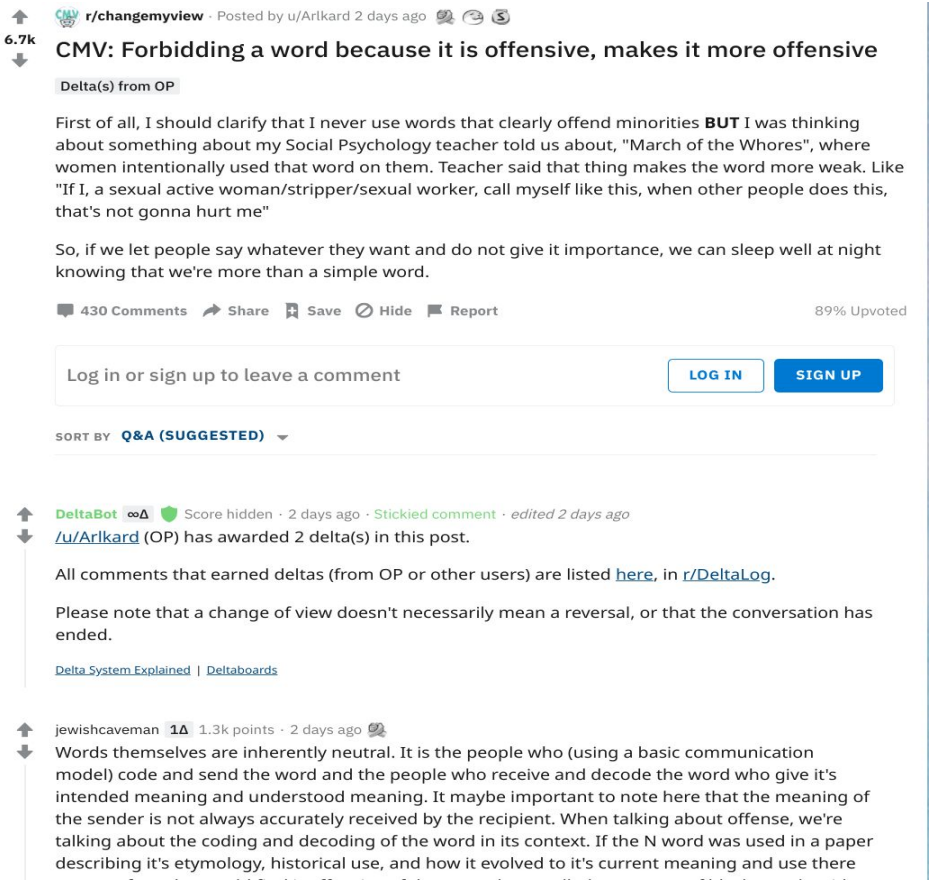
- A byproduct of Argotario is a corpus with 1.3k arguments annotated with five fallacies (including ad hominem, red herring)
  - in English and German



Change my View Corpus

- Change My View: online moderated platform for argumentation posted on reddit

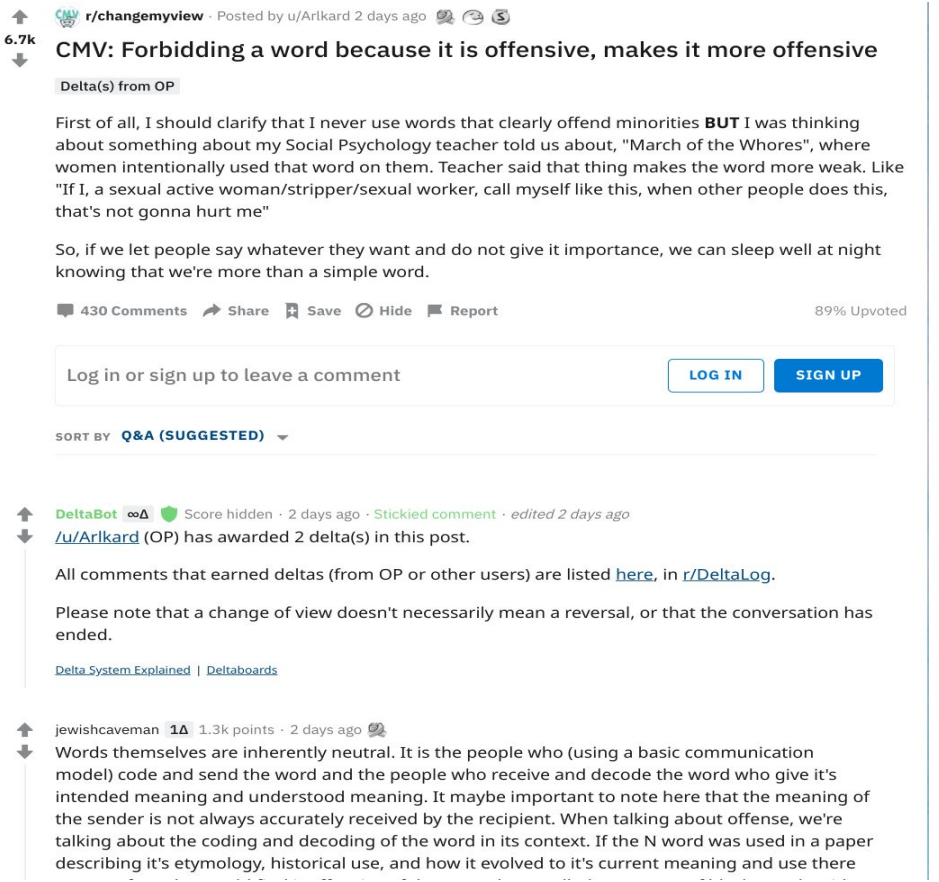
- A user posts an opinion
  - other users provide their arguments to change his/her point of view
  - the original poster award points to the convincing arguments



Change my View Corpus

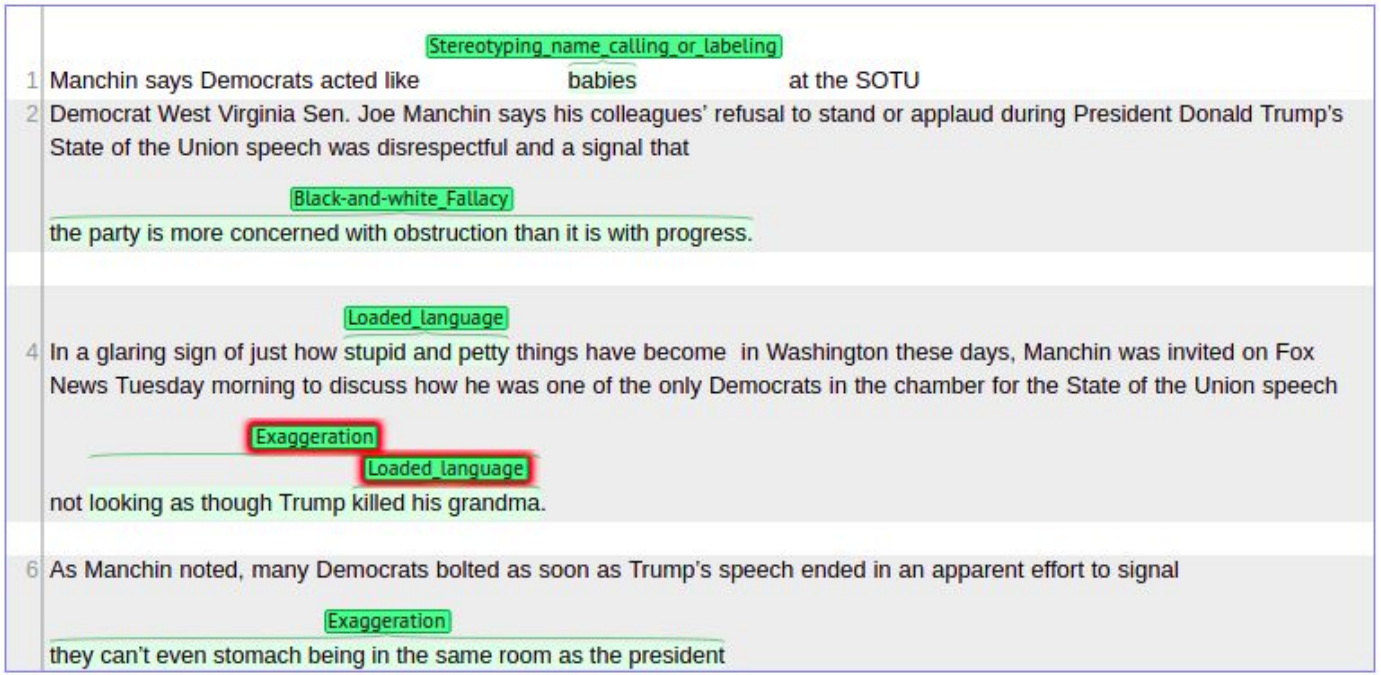
- Moderators remove all ad hominem (attack to the person) arguments
- collected 3,396 threads with 3,866 ad hominem in total

Model	Accuracy
Human upper bound estimate	0.878
2 Stacked Bi-LSTM	0.782
CNN	**0.810**



Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In Proceedings of NAACL-HLT '18, pages 386–396, New Orleans, LA, USA.

Propaganda Techniques Corpus (PTC)



G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, "Fine-Grained Analysis of Propaganda in News Articles", in EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019.

PTC Corpus

- 536 news articles in English from 48 sources (450k words) annotated at fragment level with 18 techniques (400 man hours of work)

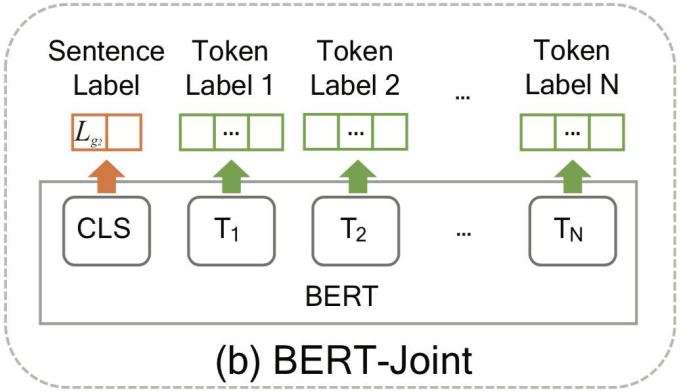
- Later extended to 2049 articles in 9 languages[1] (French, German, Italian, Polish, Russian, Greek, Spanish, Georgian)

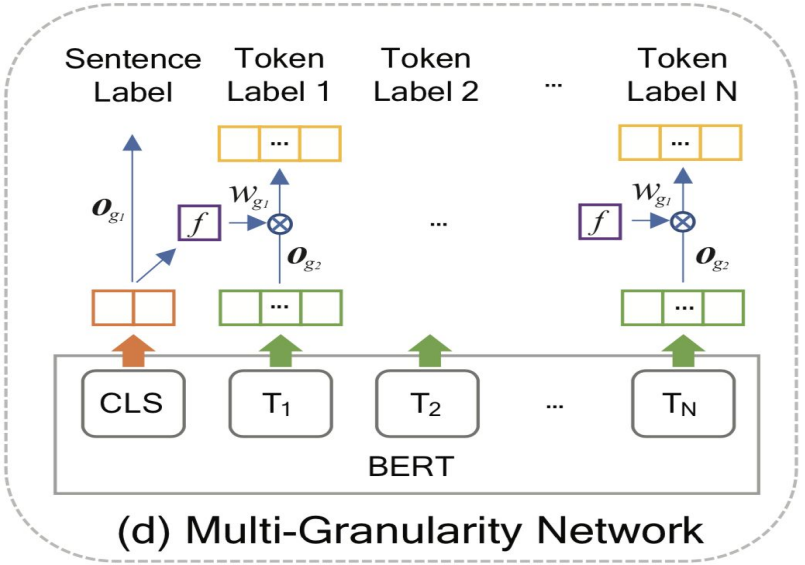
Technique	inst	avg. length
loaded language	2,547	23.70 ± 25.30
name calling, labeling	1,294	26.10 ± 19.88
repetition	767	16.90 ± 18.92
exaggeration, minimization	571	45.36 ± 35.55
doubt	562	123.21 ± 97.65
appeal to fear/prejudice	367	93.56 ± 74.59
flag-waving	330	61.88 ± 68.61
causal oversimplification	233	121.03 ± 71.66
slogans	172	25.30 ± 13.49
appeal to authority	169	131.23 ± 123.2
black-and-white fallacy	134	98.42 ± 73.66
thought-terminating cliches	95	34.85 ± 29.28
whataboutism	76	120.93 ± 69.62
reductio ad hitlerum	66	94.58 ± 64.16
red herring	48	63.79 ± 61.63
bandwagon	17	100.29 ± 97.05
obfusc., int. vagueness, confusion	17	107.88 ± 86.74
straw man	15	79.13 ± 50.72
all	7,485	46.99 ± 61.45

1 *Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino and Preslav Nakov Multilingual Multifaceted Understanding of Online News in Terms of Genre, Framing, and Persuasion Techniques*

Neural Approaches to Persuasive Spans Detection

- Output at token (word) level: beginning, middle, end of a persuasive span or not part of any span





Results on English Data

Span Level

Model	Spans			Full Task		
	P	R	F_1	P	R	F_1
BERT	39.57	36.42	37.90	21.48	**21.39**	21.39
Joint	39.26	35.48	37.25	20.11	19.74	19.92
Granu	43.08	33.98	37.93	23.85	20.14	21.80
Multi-Granularity						
ReLU	43.29	34.74	38.28	23.98	20.33	21.82
Sigmoid	**44.12**	**35.01**	**38.98**	**24.42**	21.05	**22.58**

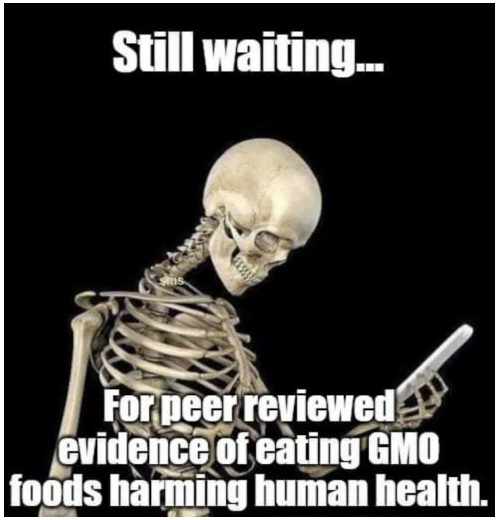
Sentence Level

Model	Precision	Recall	F1
All-Propaganda	23.92	1.00	38.61
BERT	**63.20**	53.16	57.74
BERT-Granu	62.80	55.24	58.76
BERT-Joint	62.84	55.46	58.91
MGN Sigmoid	62.27	59.56	60.71
MGN ReLU	60.41	**61.58**	**60.98**

A Demo of the System is available at https://www.tanbih.org/prta

Persuasion Techniques in Memes

- Most communication in social media is multimodal, mixing textual with visual content
- SemEval 2021 task 6: 950 memes annotated with 22 techniques[1]
- New Data (9K memes!) and a new shared task are coming soon![2]



1 Dimitar Dimitrov et al.: Detecting Propaganda Techniques in Memes. ACL/IJCNLP (1) 2021: 6603-6617
2 SemEval 2024 Task 4: https://semeval.github.io

Disseminating messages On Social Media (using bots)

- Detecting Fake Accounts: Botometer

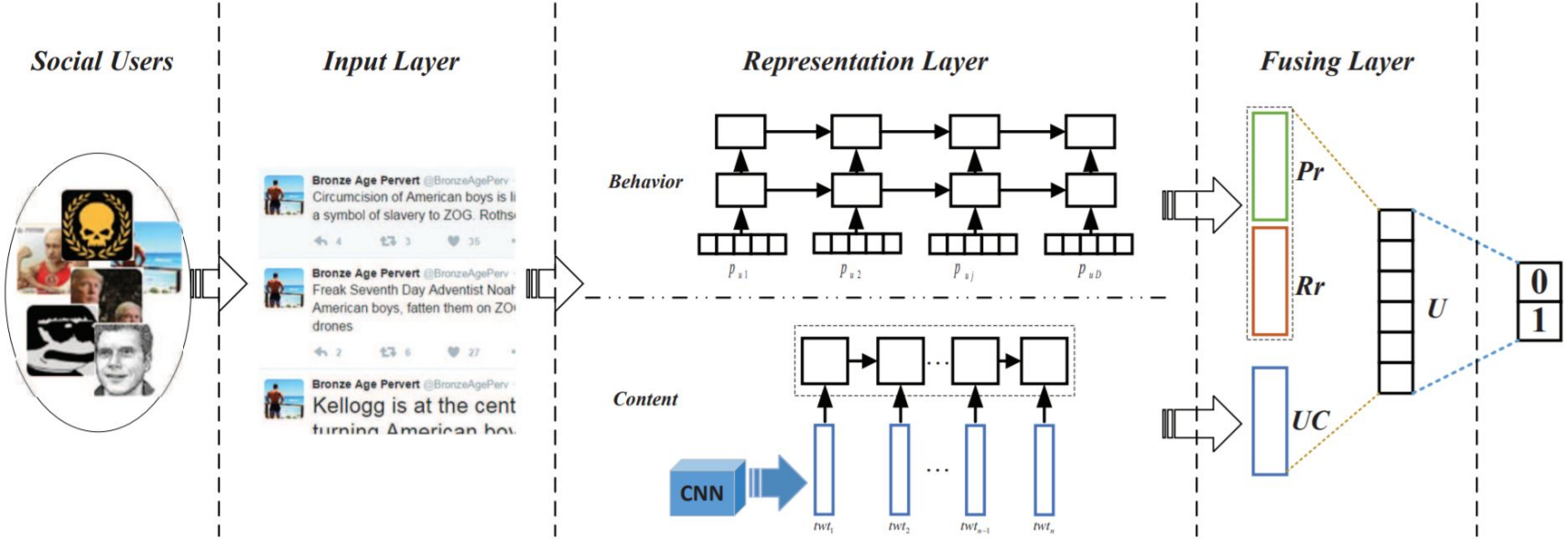
- Given a Twitter account, Botometer extracts
  - over 1,000 features relative to the account
- Yields a classification score called bot score: the higher the score, the greater the likelihood that the account is controlled by software

- Drawbacks:
  - lack of reliable ground truth
  - malicious actors evolve to avoid detection

Yang et al. Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies, 1(1):48–61, 2019

Deep Bot Detection

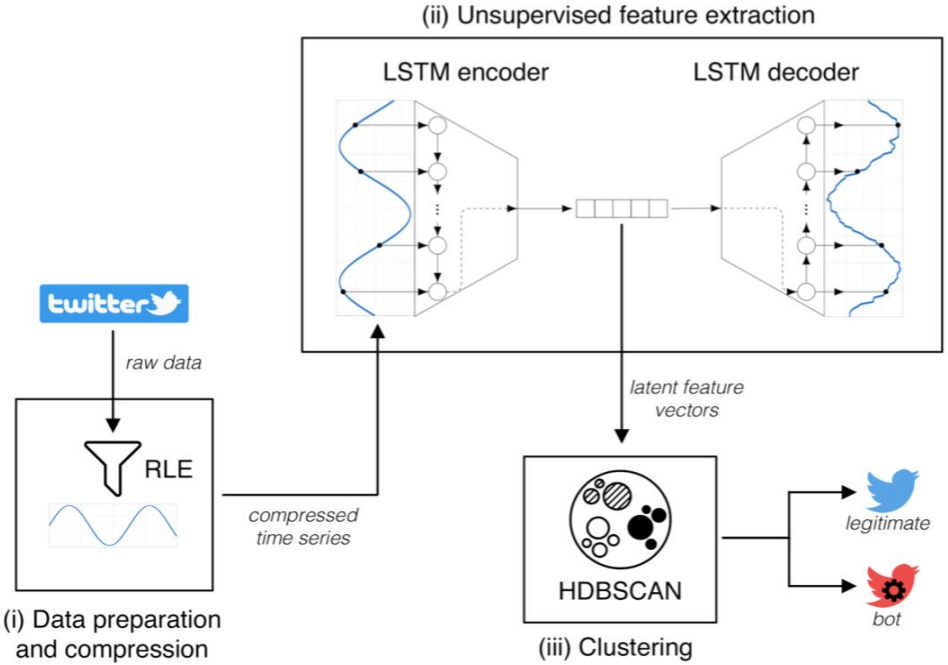
- Devise a user representation based on behaviour (posting, retweeting) and sequence of tweets' content



Cai et al. Detecting social bots by jointly modeling deep behavior and content information. In CIKM, pages 1995–1998, 2017

Detecting Coordinated Behaviour: RTBust

- Detecting coordinated behaviour instead of fake accounts
- Encode and cluster retweet patterns
- Discriminate between normal and inauthentic behaviour
- Rationale: humans exhibit more behavioural heterogeneity than bots



Mazza et al. RTbust: Exploiting temporal patterns for botnet detection on Twitter. In WebSci, pages 183–192, 2019

Conclusions

- Bot Detection is a challenging problem: most machine learning techniques are designed for stationary and neutral environments
- Coordinated behavior is not necessarily harmful

- Propaganda: “Communications that deliberately manipulate the audience to influence it towards a specific goal”
  - Detection of persuasion techniques is a recent area of research
    - encouraging results but still lots to do
  - Detecting intent is a hard task
    - Coordinated Behaviour as a proxy for intent?[1]

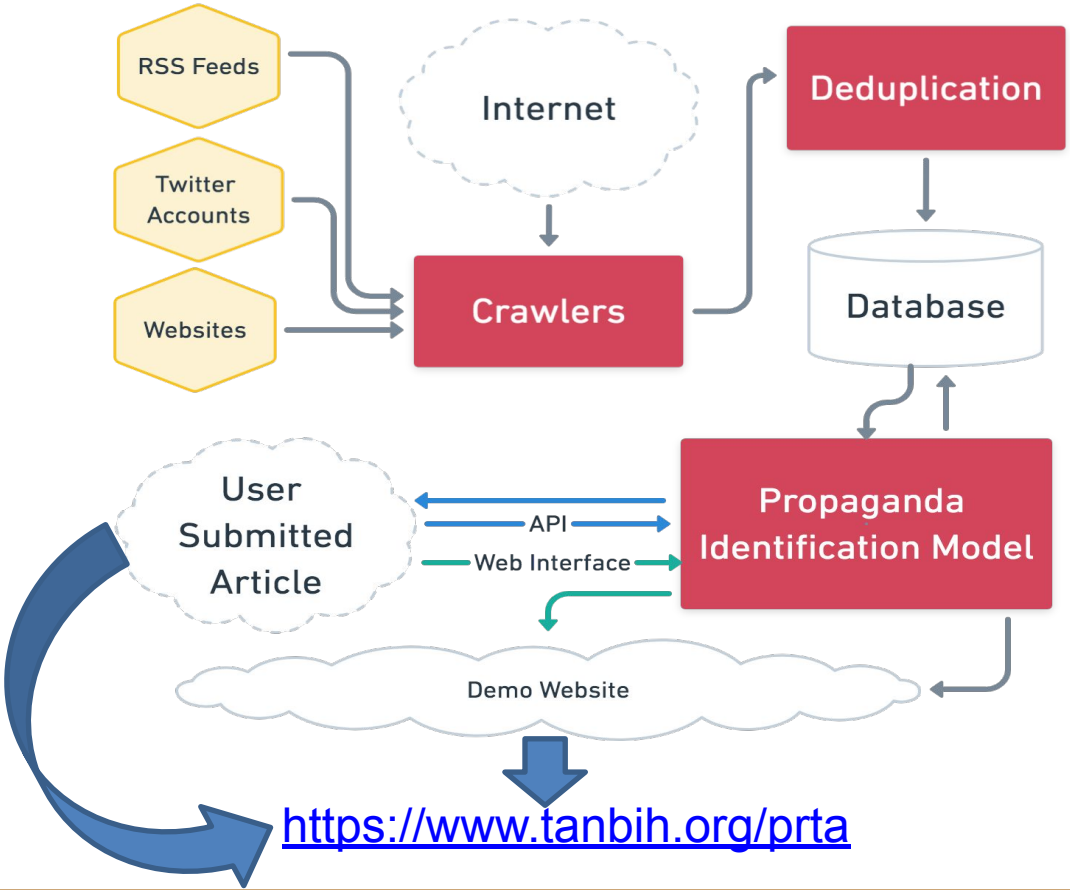
1 Giovanni Da San Martino et al. “A Survey on Computational Propaganda Detection”. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI-PRICAI ’20. Yokohama, Japan, 2020, pp. 4826–4832.

Prta: Detection of Persuasion Techniques in Texts

Prta (https://www.tanbih.org/prta)

- continuously collects news articles
    - highlights fragments with persuasion techniques
    - shows aggregated statistics
- analyses articles submitted by the user through
    - web interface
    - dedicated API (an example of usage is available here)

Prta: Detection of Persuasion Techniques in Texts



PRTA

A Tool For the Analysis of Propaganda Techniques in Texts

This page shows our tool for detecting persuasion techniques in texts.
This demo has two main functionalities:

1. Propaganda technique analysis on a topic

Collecting articles about a topic and showing aggregated statistics on the propaganda techniques our learning algorithm detects in them.
It further allows the user to customise the plots by filtering by source, date, keywords, political bias.
We apply our system to news articles, continuously collected from more than 2K sources.

Go to DEMO

Examples of topics include:

- Coronavirus Outbreak 2019-20
- Khashoggi Murder
- Gun Control and Gun Rights
- Brexit

2. Highlighting of the propaganda techniques in a text

The demo shows the list of articles related to a topic per news outlets. When clicking on an article, it shows the spans in the articles in which each technique occurs. Furthermore, we allow users to submit any text and have it analysed by our system

Submit Text for Analysis

References

The model behind the demo is also available as an API.

The demo has been published at ACL 2020:

G. Da San Martino, S. Shaar, Y. Zhang, S. Yu, A. Barrón-Cedeño, P. Nakov, Prta: A System to Support the Analysis of Propaganda Techniques in the News. In Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics (ACL 2020), Seattle, USA, July 5-10, 2020.

The learning algorithm that is used to make predictions for this demo is described in the following paper:

G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-Grained Analysis of Propaganda in News Articles. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China, November 3-7, 2019.

CORONAVIRUS OUTBREAK 2019-20

Change topic:

e.g. khashoggi

dd/mm/yyyy - dd/mm/yyyy

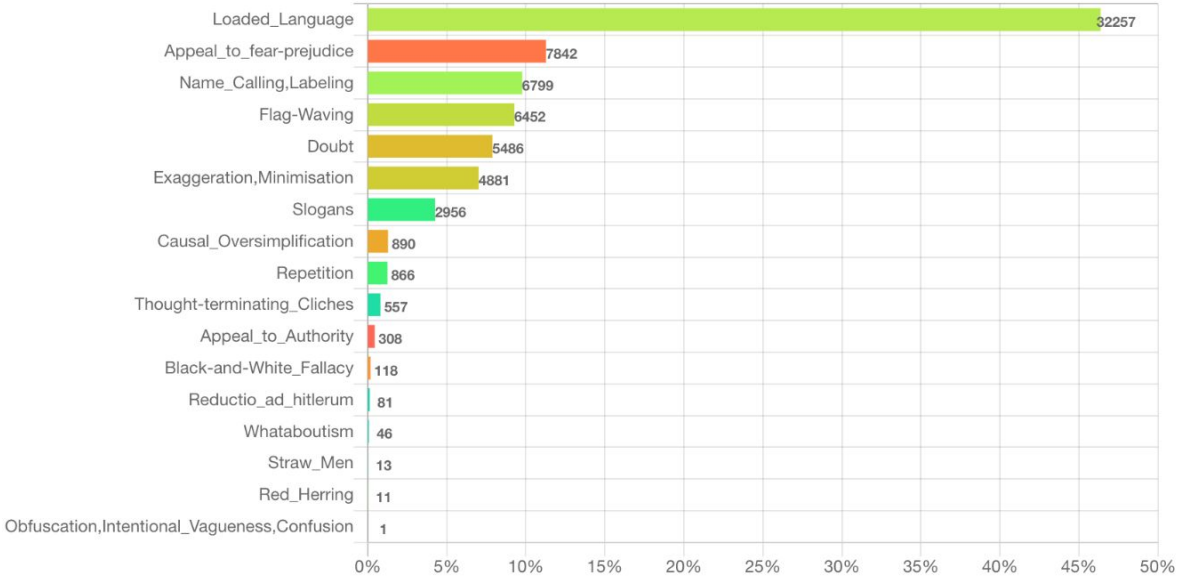
keyword, phrase Filter

Political Bias: Right Center Left N/A

- **Others** (1602)
- **Reuters** (1372)
- **International Business Times** (1063)
- **Mail Online** (635)
- **Fox News** (620)
- **CNN** (467)
- **New York Post** (387)
- **India Today** (385)
- **Guardian** (345)
- **Voice of America** (332)
- **ABC Online** (331)
- **Breitbart News** (302)
- **BBC Online** (294)
- **The Japan Times** (287)
- **The New York Times** (284)
- **Daily Mirror** (279)
- **Sputnik** (271)

Distribution of Propaganda Techniques in the Articles

Shown are the distribution and the relative proportion of use of the different propaganda techniques.
Example: Across the articles related to the topic, 46.37% of the instances of propaganda techniques are of type *Loaded_Language*. In absolute terms, this amounts to 32257 instances.



Number of Articles vs. Use of Propaganda Techniques Over Time

Change topic:

e.g. khashoggi

dd/mm/yyyy - dd/mm/yyyy

keyword, phrase Filter

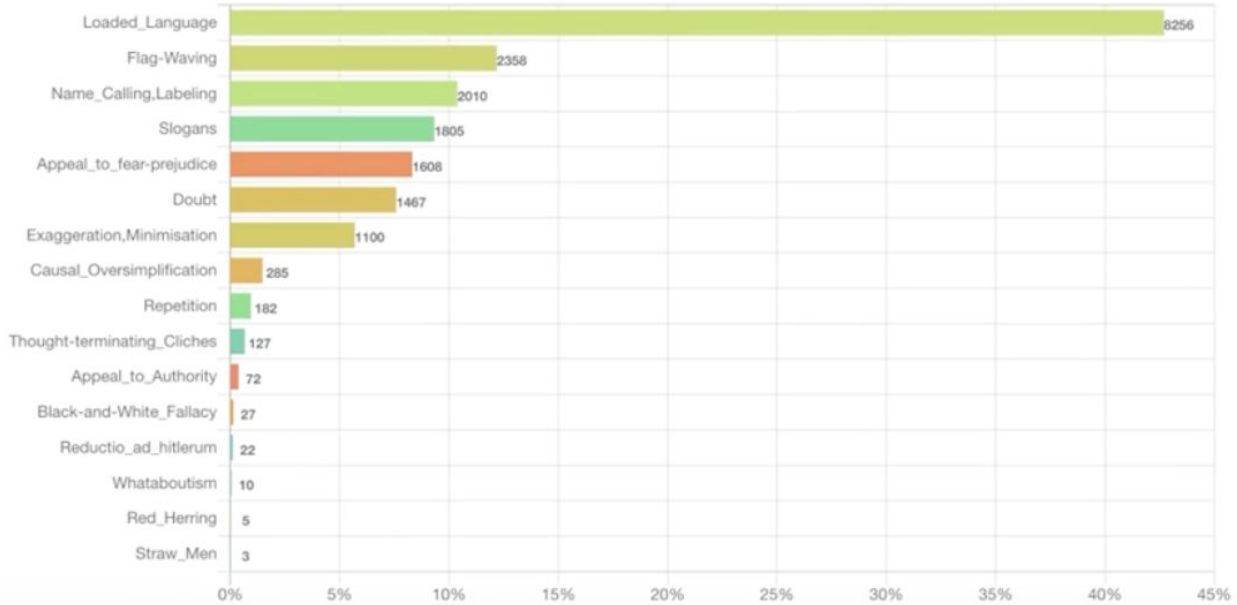
Political Bias: Right Center Left N/A

- Mail Online (635)
- Fox News (620)
- Breitbart News (302)
- Sputnik (271)
- The Hill (232)
- Epoch Times (210)
- Sightline Media Group (1)

Distribution of Propaganda Techniques in the Articles

Shown are the distribution and the relative proportion of use of the different propaganda techniques.

Example: Across the articles related to the topic, 42.7% of the instances of propaganda techniques are of type *Loaded_Language*. In absolute terms, this amounts to 8256 instances.



Change topic:

e.g. khashoggi

dd/mm/yyyy - dd/mm/yyyy

keyword, phrase Filter

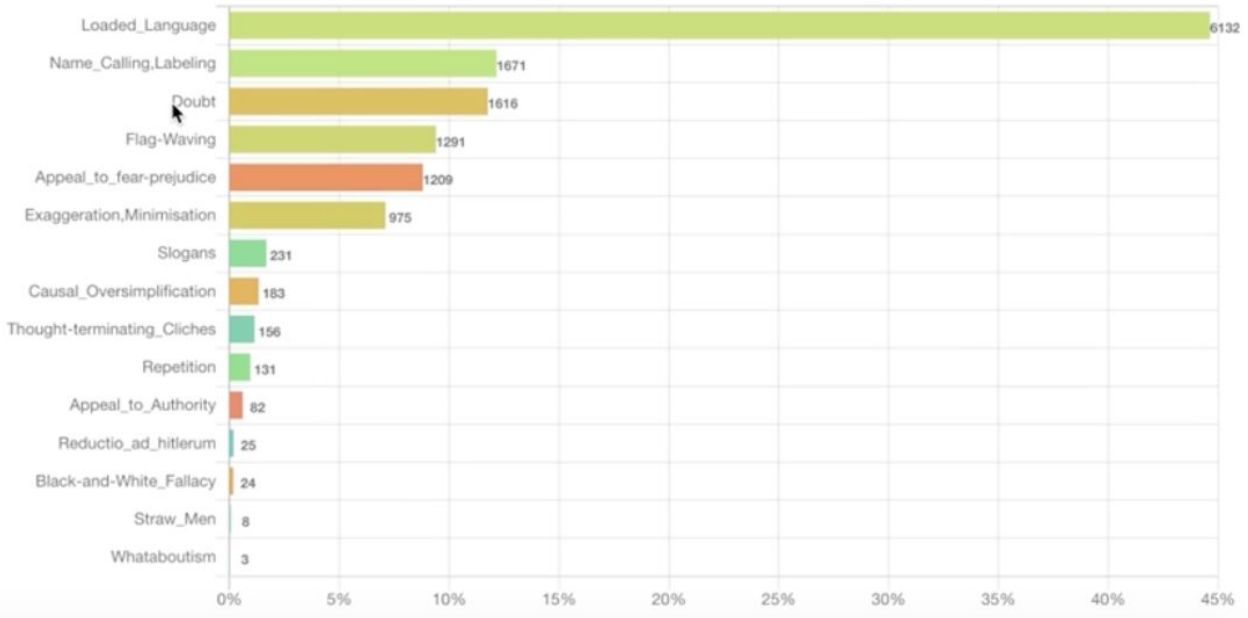
Political Bias: ☐ Right ☐ Center ☑ Left ☐ N/A

- CNN (467)
- Guardian (345)
- The New York Times (284)
- Daily Mirror (279)
- PBS (112)
- CBC.ca (98)
- The Daily Beast (67)
- Slate (magazine) (47)
- Politico (4)

Distribution of Propaganda Techniques in the Articles

Shown are the distribution and the relative proportion of use of the different propaganda techniques.

Example: Across the articles related to the topic, 44.64% of the instances of propaganda techniques are of type *Loaded_Language*. In absolute terms, this amounts to 6132 instances.



SUBMIT YOUR TEXT

What General Weygand called the Battle of France is over. I expect that the Battle of Britain is about to begin. Upon this battle depends the survival of Christian civilization. Upon it depends our own British life, and the long continuity of our institutions and our Empire. The whole fury and might of the enemy must very soon be turned on us.

Hitler knows that he will have to break us in this Island or lose the war. If we can stand up to him, all Europe may be free and the life of the world may move forward into broad, sunlit uplands. But if we fail, then the whole world, including the United States, including all that we have known and cared for, will sink into the abyss of a new Dark Age made more sinister, and perhaps more protracted, by the lights of perverted science.

Let us therefore brace ourselves to our duties, and so bear ourselves that if the British Empire and its Commonwealth last for a thousand years, men will still say, "This was their finest hour."

Copy Text From URL | e.g. https://www.bbc.com/nev

SUBMIT

- What General Weygand called the Battle of France is over. I expect that the Battle of Britain is about to begin.
Upon this battle depends the survival of Christian civilization. Upon it depends our own British life, and the long continuity of our institutions and our Empire.
[8] The whole fury and might[9] of the enemy must very soon be[2] turned on us.

- Hitler knows that he will have to[2] break us in this[2] Island or lose the war. If[2] we can[5] stand up to him, all Europe may be free[5] and the[2] life of the world[5] may move[2] forward into broad, sunlit uplands.[2] But if we fail, then[2] the whole[8] world,[7] including the United States, including all that we have known[8] and cared for,[8] will sink into the abyss of a new Dark Age made more sinister,[9] and perhaps more protracted, by the lights of perverted science.

Show only predictions with confidence ≥ 0.05

0 — 1

Technique Types (More info)

- [x] 2 - Appeal to fear prejudice (?)
- [x] 5 - Causal/Oversimplification (?)
- [x] 7 - Exaggeration, Minimisation (?)
- [x] 8 - Flag Waving (?)
- [x] 9 - Loaded Language (?)

Dark Age made more sinister, and perhaps more protracted, by the lights of perverted science.

Let us therefore brace ourselves to our duties, and so bear ourselves that if the British Empire and its Commonwealth last for a thousand years, men will still say, "This was their finest hour."

Copy Text From URL | e.g. https://www.bbc.com/nev

SUBMIT

Flag Waving
Playing on strong national feeling (or to any group; e.g., race, gender, political preference) to justify or promote an action or idea

- What General Weygand called the Battl…he Battle of Britain is about to begin.

Upon this battle depends the survival of Christian civilization. Upon it depends our own British life, and the long continuity of our institutions and our Empire.

[8] The whole fury and might[9] of the enemy must very soon be[2] turned on us.

- Hitler knows that he will have to[2] break us in this[2] Island or lose the war. If[2] we can[5] stand up to him, all Europe may be free[5] and the[2] life of the world[5] may move[2] forward into broad, sunlit uplands.[2] But if we fail, then[2] the whole[8] world,[7] including the United States, including all that we have known[8] and cared for,[8] will sink into the abyss of a new Dark Age made more sinister,[9] and perhaps more protracted, by the lights of perverted science.

- Let us therefore brace ourselves to our duties, and so bear ourselves that if the British Empire and its Commonwealth last for a thousand years, men will still say, "This was their finest hour."

Show only predictions with confidence ≥ 0.05

0 — 1

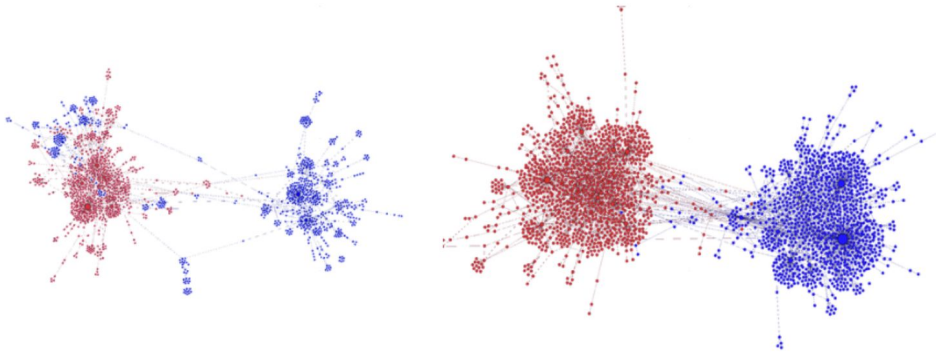
Technique Types (More info)

- [x] 2 - Appeal to fear prejudice (?)
- [x] 5 - Causal/Oversimplification (?)
- [x] 7 - Exaggeration, Minimisation (?)
- [x] 8 - Flag Waving (?)
- [x] 9 - Loaded Language (?)
- [ ] 1 - Appeal to Authority
- [ ] 3 - Bandwagon
- [ ] 4 - Black and White Fallacy
- [ ] 6 - Doubt
- [ ] 10 - Name Calling, Labeling
- [ ] 11 - Obfuscation, Intentional Vagueness, Confusion
- [ ] 12 - Red Herring

Controversy Detection

Polarization on Social Media

Tutorial || KDD 2018, WebConf 2018, ICWSM 2017



https://gvrkiran.github.io/polarization/

Kiran Garimella

Gianmarco De Francisci Morales

Michael Mathioudakis

Aristides Gionis

controversy noun

con·tro·ver·sy ˈkän-trə-ˌvər-sē *British also* kən-ˈträ-və-sē

plural **controversies**

Synonyms of *controversy* ›

1 : a discussion marked especially by the expression of opposing views : DISPUTE

> The decision aroused a *controversy* among the students.

2 : QUARREL, STRIFE

values

polarization noun

po·lar·i·za·tion ˌpō-lə-rə-ˈzā-shən

plural **polarizations**

Synonyms of *polarization* ›

: division into two sharply distinct opposites

especially : a state in which the opinions, beliefs, or interests of a group or society no longer range along a continuum but become concentrated at opposing extremes

CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE

In total, 26 out of the 52 observed episodes (or 50% of cases) saw their country's Regimes of the World score downgraded, with the vast majority of those -- 23 -- descending into some form of authoritarianism.

TABLE 4: OUTCOMES OF EPISODES OF PERNICIOUS POLARIZATION

Backsliding Within Democracy [From Liberal Democracy to Electoral Democracy]	Erosion From Democracy to Electoral Autocracy [From Liberal or Electoral Democracy to Electoral Autocracy]	Democratic Collapse [From Liberal or Electoral Democracy to Closed Autocracy]
• Mauritius, 1968–2019	• Bangladesh, 1992–2002	• Argentina, 1964–1966
• Poland, 2011–2016	• Comoros, 2010–2015	• Argentina, 1974–1976
• Slovenia, 2018–2020	• Dominican Republic, 1982–1990	• Chile, 1970–1973
	• Hungary, 2010–2018	• Fiji, 1993–2000
	• India, 2014–2019	• Fiji, 2002–2006
	• Indonesia, 1956–1958	• Malta, 1950–1957
	• Kosovo, 2002–2005	• Thailand, 2004–2006
	• Lebanon, 2010–2018	• Turkey, 1966–1980
	• Maldives, 2009–2013	• Uruguay, 1966–1973
	• Montenegro, 2004–2006	
	• Nepal, 2010–2012	
	• North Macedonia, 2008–2012	
	• Suriname, 1988–1991	
	• Turkey, 2002–2013	

https://carnegieendowment.org/2022/01/18/what-happens-when-democracies-become-perniciously-polarized-pub-86190

Causes of polarization: cognitive dissonance

"Cognitive dissonance" – psychological conflict resulting from incongruous beliefs and attitudes held simultaneously

Selective exposure

Klapper. "The effects of mass communication." 1960

Subjects choose to examine items that agree with their decision

Biased assimilation

Lord et al. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." 1979

Subjects find consonant evidence more convincing

Free-choice

Brehm. "Postdecision changes in the desirability of alternatives." 1956

Spreading-apart-of-alternatives

Induced compliance

Festinger and Carlsmith. "Cognitive consequences of forced compliance." 1959

Subjects justify their decisions a-posteriori, even if they originally disagreed

Causes of polarization: group bias

Social identity complexity

Roccas, S. and Brewer, M.B., 2002. Social identity complexity. Personality and Social Psychology Review.

Individuals associate themselves with social identities

race, religion, gender, class

Group polarization

Sunstein, C.R., 2002. The law of group polarization. Journal of political philosophy.

The tendency for a group to make decisions that are more extreme than the initial inclination of its members

Causes of polarization: algorithmic bias

Personalization

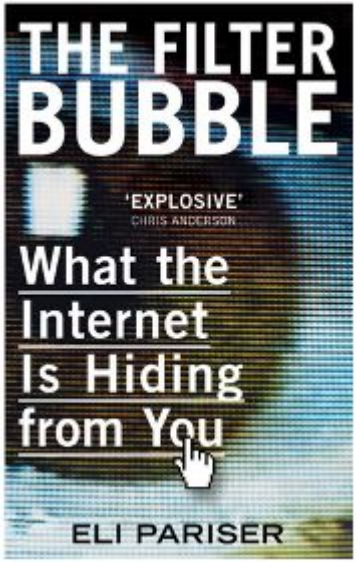
in news, search, shopping, dating, any content

Social feedback

homophily, rich gets richer, groupthink

Filter bubble - intellectual isolation that can result from personalized search/recommendation

Echo chambers - environment or ecosystem in which participants encounter beliefs that amplify or reinforce their preexisting beliefs by communication and repetition inside a closed system and insulated from rebuttal



Cinelli, Matteo, et al. "The echo chamber effect on social media." *PNAS*. 2021

Defining polarization

A society can be thought of as an amalgamation of *groups*, where two individuals drawn from the same group are "similar," and from different groups, are "different" relative to some given set of *attributes*. The *polarization* of a distribution of individual attributes must exhibit the following basic features:

1. There must be a high degree of homogeneity *within* each group
2. There must be a high degree of heterogeneity *across* groups
3. There must be a small number of significantly sized groups. In particular, groups of insignificant size (e.g., isolated individuals) carry little weight.

less polarized

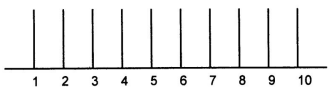


FIGURE 1A

more polarized

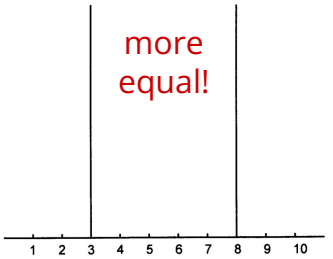


FIGURE 1B

even more

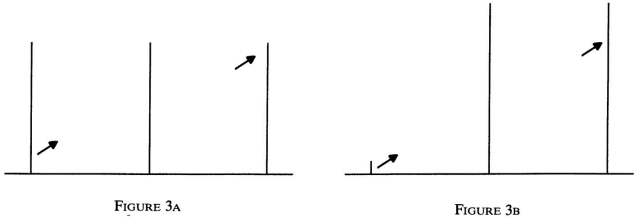


FIGURE 3A

FIGURE 3B

Esteban, J. M., & Ray, D. (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, 819-851.

Defining polarization

highly polarized

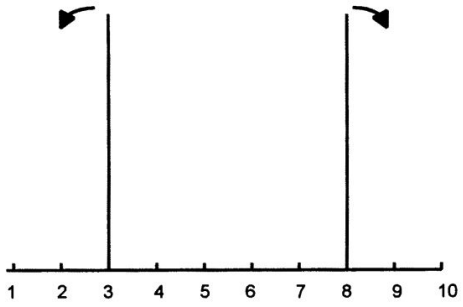


FIGURE 5A

less polarized
(less cohesive groups)

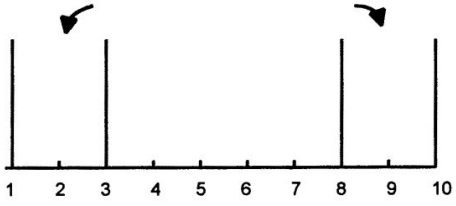


FIGURE 5B

more polarized

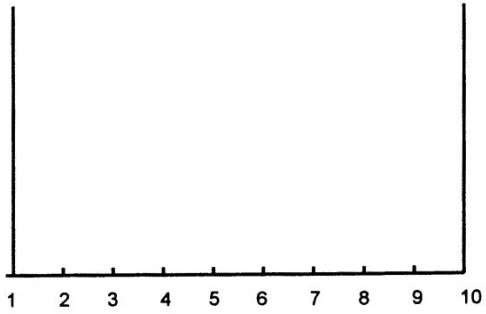


FIGURE 5C

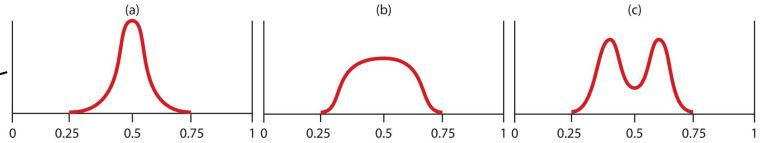
Esteban, J. M., & Ray, D. (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, 819-851.

Defining polarization

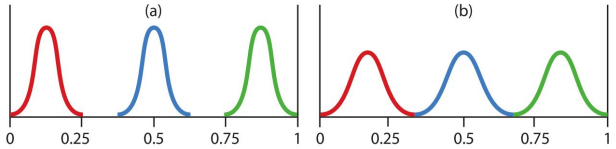
Measures of polarization:

1. spread
2. dispersion
3. coverage
4. regionalization
5. community fragmentation
6. role of groups in some senses of polarization
7. distinctness
8. group divergence
9. group consensus
10. size parity

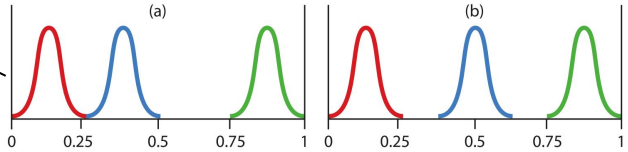
$$\text{dispersion} = \frac{2}{N}\sum_{x_i}^{N} |x_i - \bar{X}|$$



$$\text{coverage}_d = \frac{1}{|R|}\sum_{r\in R}\begin{cases} 1 & y(x_r) > 0 \\ 0 & y(x_r) = 0 \end{cases}$$



$$\sum_{r_i\in R}\begin{cases} 1 & y(r_0) = 0 \\ 1 & y(r_i) > 0 \text{ and } y(r_{i+1}) = 0 \\ 0 & \text{otherwise} \end{cases}$$

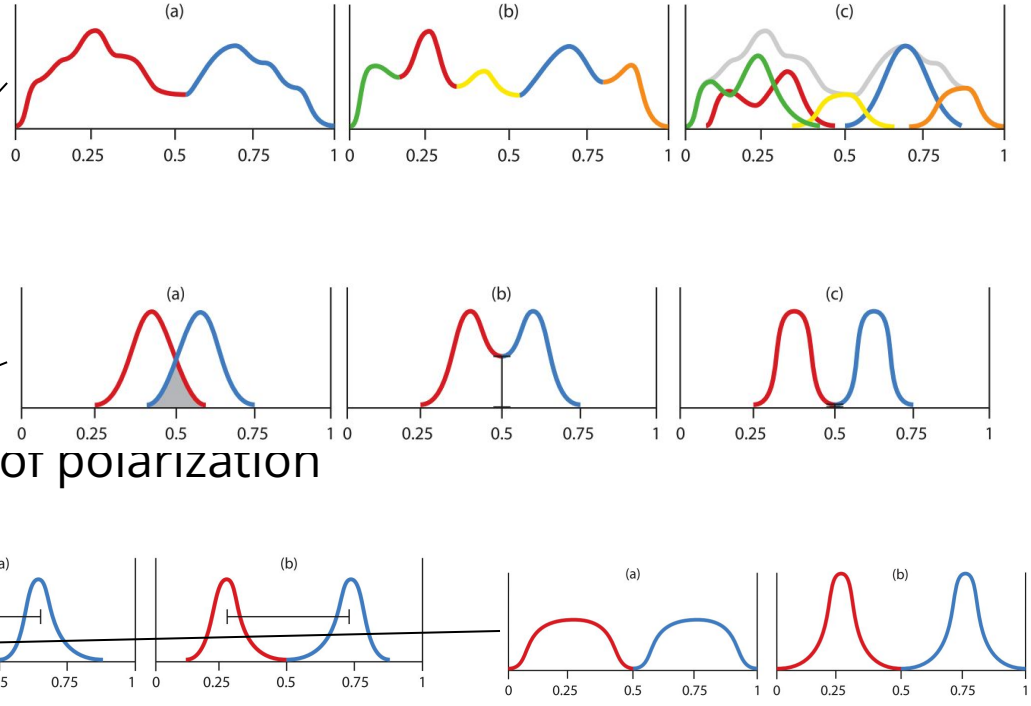


Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. The Journal of Mathematical Sociology, 40(2), 80-111.

Defining polarization

Measures of polarization:

1. spread
2. dispersion
3. coverage
4. regionalization
5. community fragmentation
6. role of groups in some senses of polarization
7. distinctness
8. group divergence
9. group consensus
10. size parity



Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. The Journal of Mathematical Sociology, 40(2), 80-111.

Detecting polarization: content

Social media: hashtags as topic indicators

$$sim(h_s, h_t) = \frac{1}{1 + \log(df(h_t))} (\alpha \cos(W_s, W_t) + (1 - \alpha) \cos(H_s, H_t))$$

Inverse document frequency (in general subset)

sets of *W*ords and *H*ashtags that co-occur with hashtag h_x

#baltimoreriots

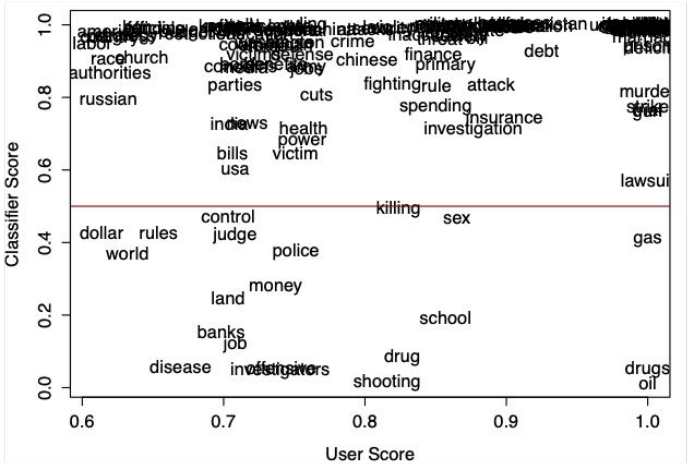
#netanyahuspeech

Garimella, De Francisci Morales, Gionis & Mathioudakis. “Quantifying Controversy in Social Media.” WSDM 2016.
Klenner, Amsler, Hollenstein & Faaß. “Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification.” KONVENS 2014.

Detecting polarization: content

- Controversy lexicons
- Controversial topics have:
  - strongly biased terms
  - more negative terms
  - fewer strongly emotional terms

"we show that we can indicate to what extent an issue is controversial, by comparing it with other issues in terms of how they are portrayed across different media."



(b) Controversial words; correctly classified words appear above the horizontal line.

Figure 2: Scores of controversial and non-controversial words including classification errors. "User score" is the confidence with which the manual labeling was done (with at least 7 annotators per element), while "classifier score" is the output of the classifier on the training data.

Mejova, Zhang, Diakopoulos, and Castillo. "Controversy and Sentiment in Online News." C+J Symposium 2014.

Detecting polarization: content

Controversial topic - a concept that invokes conflicting sentiments

Subtopic - factor that gives a particular sentiment (+ve or -ve) - noun phrases

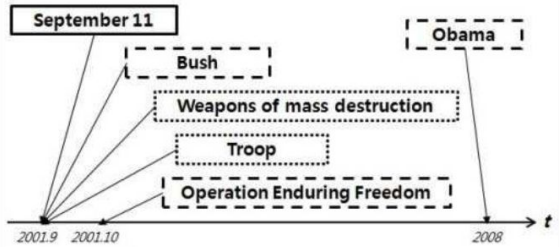
Assumption - a controversial topic receives contrasting sentiment (of different kind)

Controversiality - sum of magnitudes of sentiments around subtopics, and their difference

Issue: ***Afghanistan War***

September 11	positive	The Afghanistan war launched after the September 11 …
	negative	The Afghanistan war was of revenge by the Americans for September 11
Troops	negative	Most Americans oppose sending more troops to Afghanistan war …
Weapons of mass destruction	negative	The Afghanistan war is perilous because of weapons of mass destruction …
Obama	positive	Obama supports the Afghanistan war …

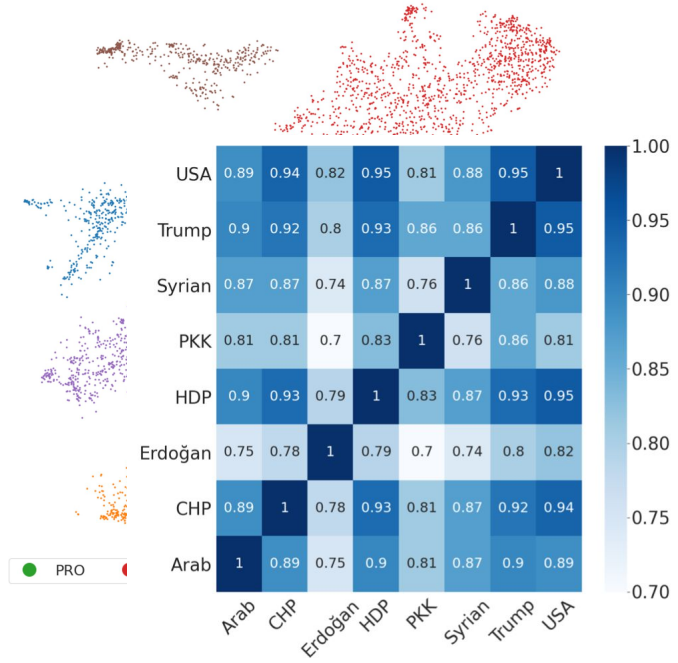
Issue: ***Afghanistan War***



Choi, Jung & Myaeng. “Identifying controversial issues and their sub-topics in news articles.” PAW-ISI 2010.

Detecting polarization: content

- Use Google's Multilingual Universal Sentence Encoder (MUSE) with pre-trained CNN embeddings to represent posts
- Project each user vector onto a two-dimensional plane using Uniform Manifold Approximation and Projection (UMAP) algorithm
- Cluster the projected user vectors using hierarchical density based clustering (HDB-SCAN)
- Compare clusterings of embeddings for different topics



Rashed, A., Kutlu, M., Darwish, K., Elsayed, T., & Bayrak, C. (2021, May). Embeddings-based clustering for target specific stances: The case of a polarized turkey. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 15, pp. 537-548).

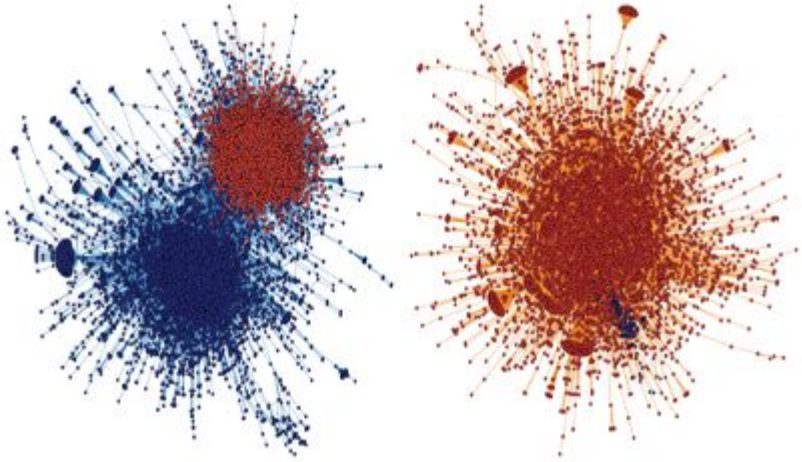
Detecting polarization: content

- Find out if a Web page discusses a (known) controversial topic
- Map topics (named entities) in a Web page to Wikipedia articles
  - A Web page is controversial if it is similar to a controversial Wikipedia article
  - E.g., If a news article mentions Abortion it is controversial
- Related:
  - There is a lot of work on identifying controversial topics on Wikipedia
  - Edit wars, hyperlink structure, etc.

Dori-Hacohen and Allan. “Detecting Controversy on the Web.” CIKM 2013.

Detecting polarization: network

- Retweet network for political hashtags has a bi-clustered structure
  - Retweet network exhibits a highly modular structure, segregating users into two homogenous communities corresponding to the political left and right
- Users mention/reply to others from their opposing viewpoint

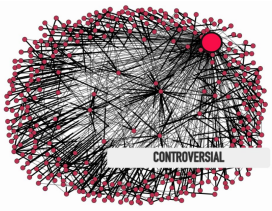


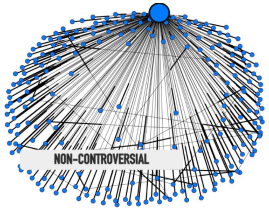
Retweet Mention

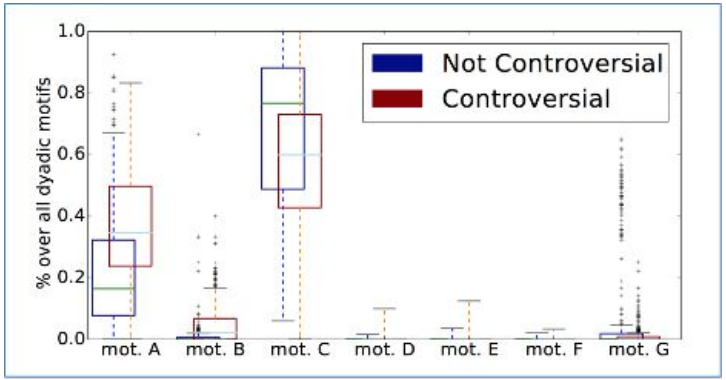
Conover, Ratkiewicz, Francisco, Gonçalves, Menczer, and Flammini. “Political Polarization on Twitter.” ICWSM 2011.

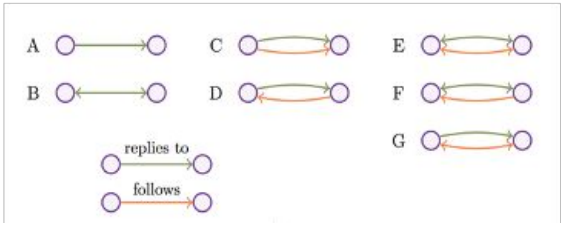
Detecting polarization: network

- Define reply trees
- Identify frequency of motifs in these trees
- Take into account also social graph (follower information)





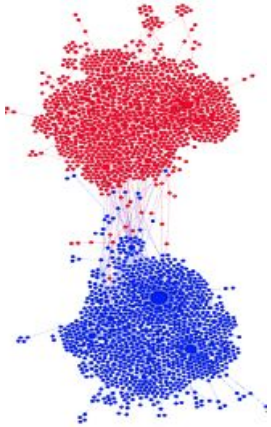




Coletto, Garimella, Luchesse, and Gionis. “A Motif-based Approach for Identifying Controversy.” OSNEM. 2017.

Detecting polarization: network

- Community boundary - possible expressions of antagonism
- Boundary node:
  - have at least one edge that connecting to the other community
  - have at least one edge connecting to a member of its community which does not link to the other community
- “polarized networks tend to exhibit low concentration of popular nodes along the boundary”
- $P(v) = d_{internal}(v)/(d_{external}(v) + d_{internal}(v)) – 0.5$
- $P(v) > 0 \rightarrow$ v prefers internal connections (antagonism?)
- $P(v) < 0 \rightarrow$ v prefers connections with members of the other group (increased homophily!)



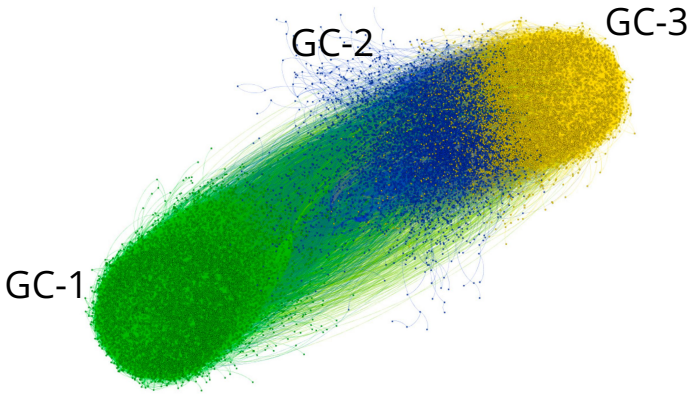
Guerra, Meira, Cardie, and Kleinberg. “A Measure of Polarization on Social Media Networks Based on Community Boundaries.” ICWSM 2013.

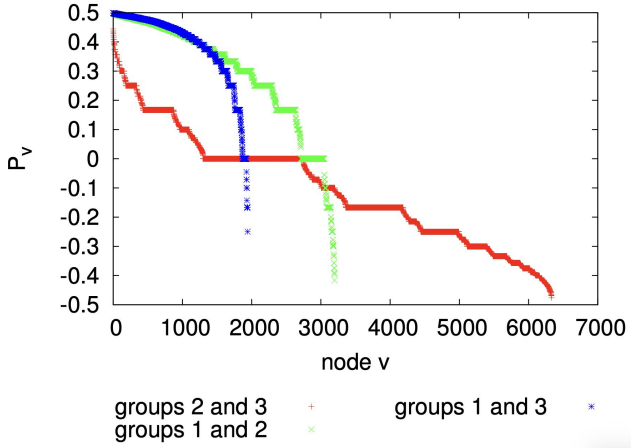
Detecting polarization: network

Gun debate network

Boundary polarization shows communities 2 & 3 agree

communities	modularity Q	polarization P
GC-1 and GC-2	0.31	+0.23
GC-1 and GC-3	0.47	+0.32
GC-2 and GC-3	0.26	-0.14

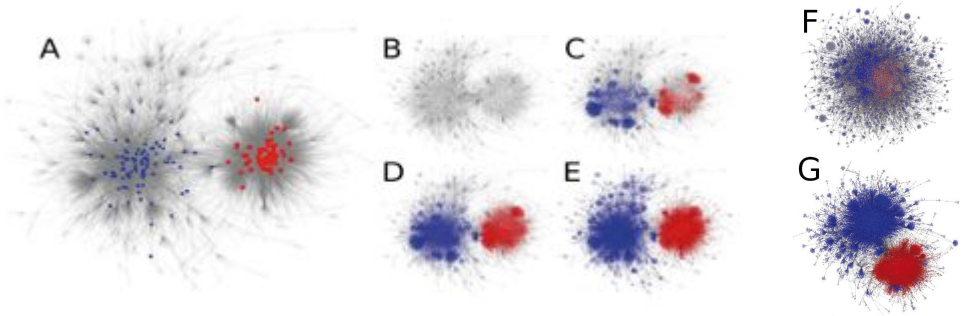


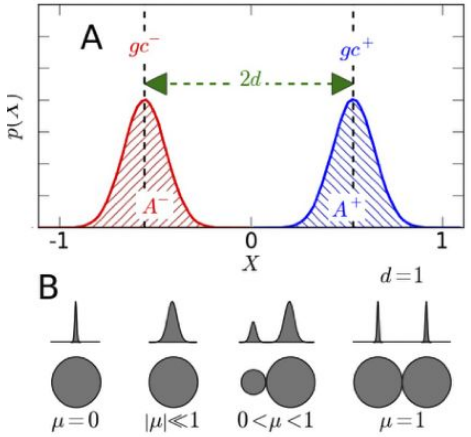


Guerra, Meira, Cardie, and Kleinberg. “A Measure of Polarization on Social Media Networks Based on Community Boundaries.” ICWSM 2013.

Detecting polarization: network

- Opinion formation:
  - Propagation of opinions from “elite” users to “listeners”
- Measure: distance between distributions
  - Distance between two gravity centers of opinions
  - Accounts for the mass of the population





Morales, Borondo, Losada, and Benito. “Measuring political polarization: Twitter shows the two sides of Venezuela.” Chaos. 2015.

Detecting polarization: network

- Bi-partition retweet (endorsement) graph using METIS
- Random Walk Controversy (RWC) Score: *"Consider two random walks, one ending in partition X and one ending in partition Y , RWC is the difference of the probabilities of two events: (i) both random walks started from the partition they ended in and (ii) both random walks started in a partition other than the one they ended in"*

$$RWC = P_{XX}P_{YY} - P_{YX}P_{XY}$$

$$P_{AB} = Pr[\text{start in partition } A \mid \text{end in partition } B]$$

(i) These probabilities are not skewed by the size of each partition, as the random walk starts with equal probability from each partition, and
(ii) they are not skewed by the total degree of vertices in each partition, as the probabilities are conditional on ending in either partition (i.e., the fraction of random walks ending in each partition is irrelevant).

Garimella, De Francisci Morales, Gionis, and Mathioudakis. "Quantifying Controversy on Social Media." Transactions on Social Computing. 2018.

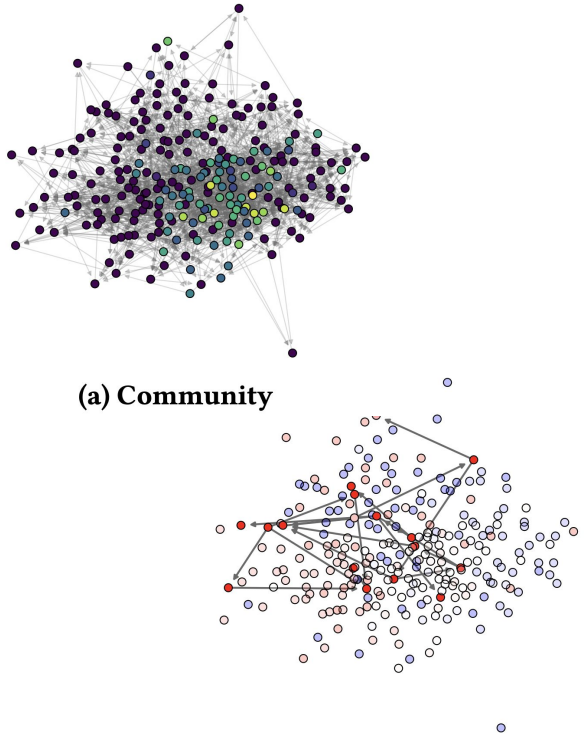
Detecting polarization: network

Input: information cascade (retweet) and social (follow) networks

Probabilistic generative model with latent variables:

- θc ,u ∈ [0, 1]: the level of polarized engagement of user u in a echo chamber c
- φc ,u ∈ [0, 1]: the level of social engagement of user u in a social community c

Inferred using Generalized Expectation Maximization algorithm

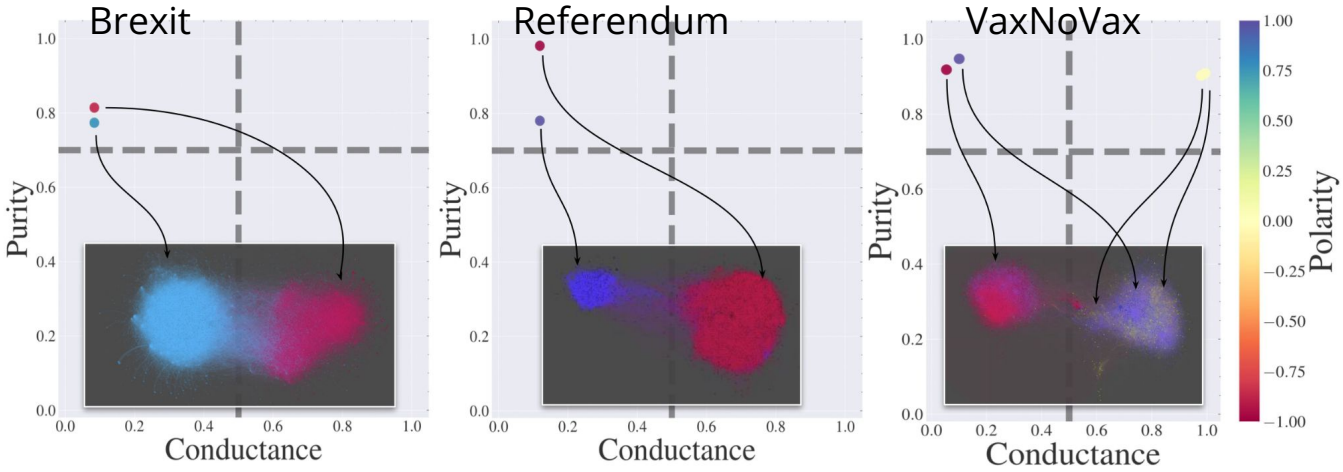


(a) Community

(b) Propagation

Minici, M., Cinus, F., Monti, C., Bonchi, F., & Manco, G. (2022, October). Cascade-based echo chamber detection. I *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1511-1520).

Detecting polarization: network



conductance —how closely-knitted is the community with the rest of the graph
purity—the ratio of users with the same ideological alignment, measured as the average polarity of the tweets they reshare

Minici, M., Cinus, F., Monti, C., Bonchi, F., & Manco, G. (2022, October). Cascade-based echo chamber detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1511-1520).

Audience suggestion

Suggested reading during discussion:

> Salloum, A., Chen, T. H. Y., & Kivelä, M. (2022). Separating polarization from noise: comparison and normalization of structural polarization measures. Proceedings of the ACM on human-computer interaction, 6(CSCW1), 1-33.
>
> https://dl.acm.org/doi/abs/10.1145/3512962

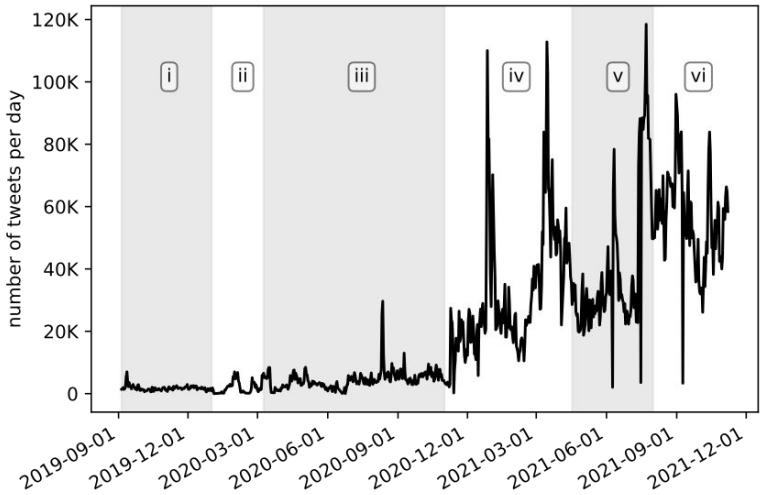
Case Study: measuring anti-vax echo chambers

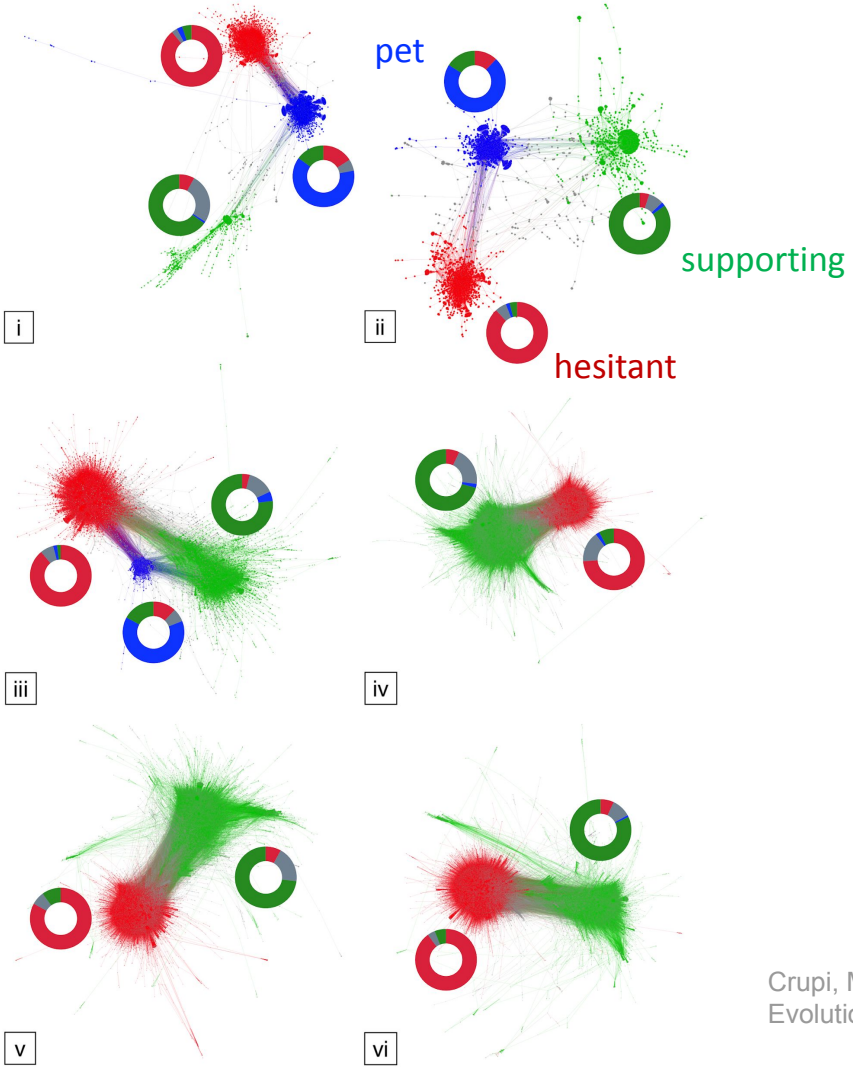
Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate

Giuseppe Crupi, Yelena Mejova, Michele Tizzani, Daniela Paolotti, Andre Panisson

@ International AAAI Conference on Web and Social Media (ICWSM) 2022

- Twitter Streaming API
- Italian language filter
- Sep 5, 2019 – Nov 7, 2021
- >16M users, 665K tweets
- 6 time periods
- 6 retweet “endorsement” retworks (weight > 1)





Opinion communities

- hierarchical clustering + selection using modularity
- manual annotation of users

- strong separation
- pet users bridge hesitant and supporting camps

Crupi, Mejova, Tizzani, Paolotti, Panisson. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. ICWSM 2022

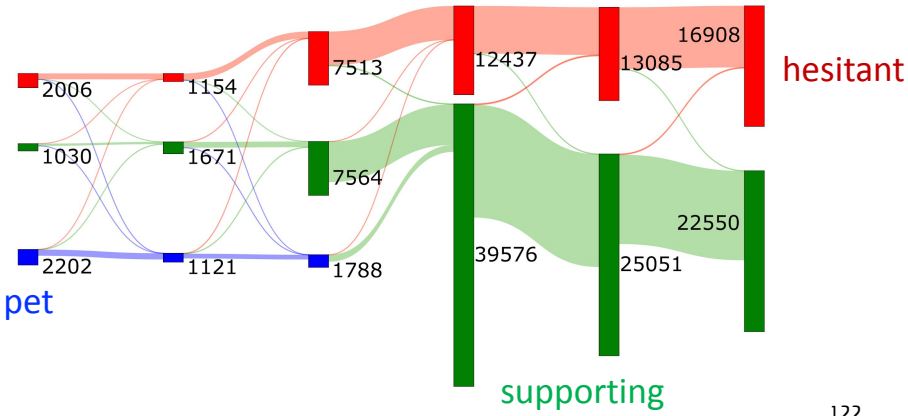
Measuring echo chambers: membership

- almost nobody changes sides
- most vaccine supporters: early vaccine period (iv)
- most vaccine hesitant: late vaccine period (vi)

tweets per user per day

		$\lvert H_{usr} \rvert$	$\lvert S_{usr} \rvert$	$\lvert H_{day} \rvert$	$\lvert S_{day} \rvert$
i.	pre-Covid	41.9	8.7	0.35	0.07
ii.	early-Covid	24.0	7.0	0.40	0.12
iii.	pre-vaccine	61.8	18.8	0.27	0.08
iv.	early-vaccine	137.4	48.7	0.83	0.30
v.	vaccine-drive	153.2	37.4	1.45	0.35
vi.	late-vaccine	191.3	43.2	1.95	0.44

i

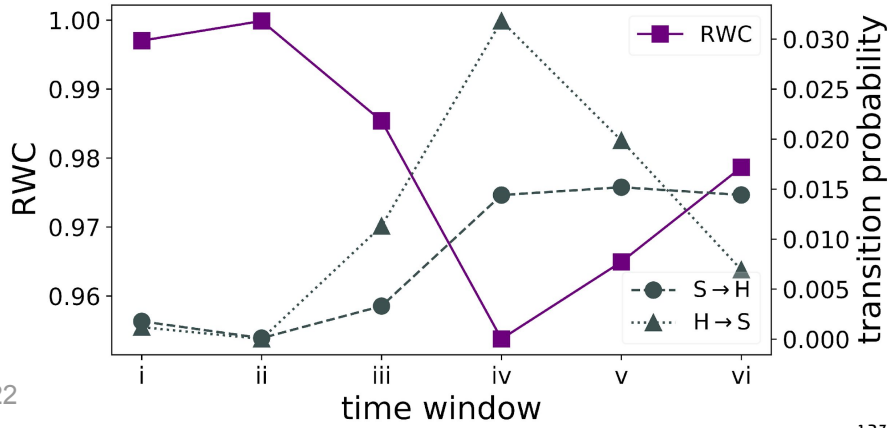


Crupi, Mejova, Tizzani, Paolotti, Panisson. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. ICWSM 2022

Measuring echo chambers: endorsement

- Random Walk Controversy score:
  - “how likely a random user on either side is to be exposed to authoritative content from the opposing side”

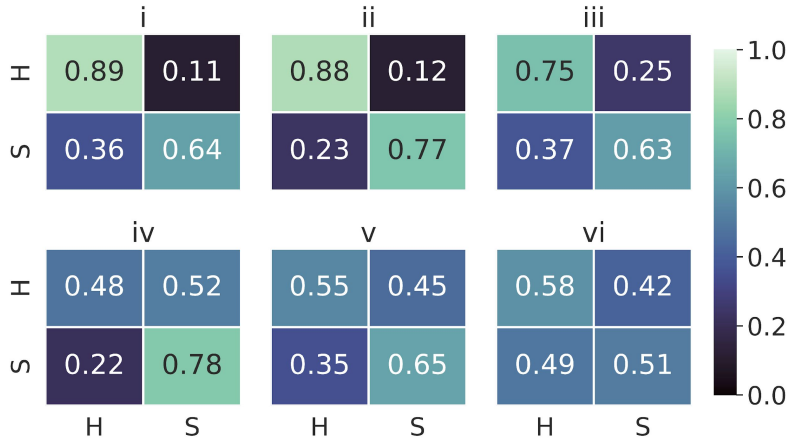
  [Garimella et al. TSC’18]



Crupi, Mejova, Tizzani, Paolotti, Panisson. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. ICWSM 2022

Measuring echo chambers: mentions

- pre-COVID two sides almost do not mention each other
- during vaccine rollout, both sides mentioned each other almost as much as themselves

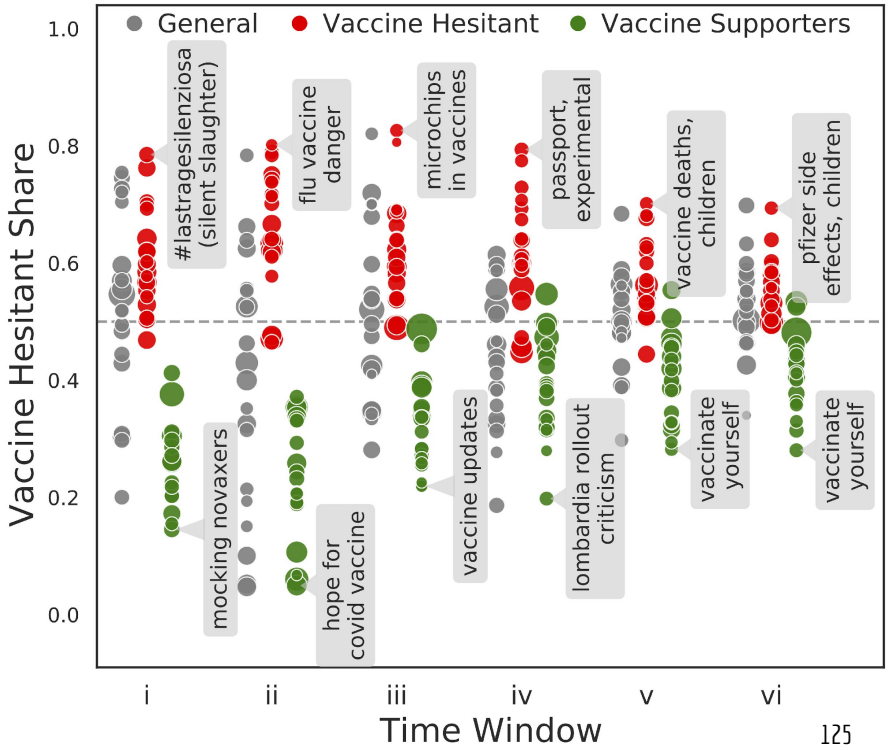


out of all mentions by row x, how many are from column y

Crupi, Mejova, Tizzani, Paolotti, Panisson. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. ICWSM 2022

Measuring echo chambers: topics

- extract 20 topics using NMF
- compute how much of the topics is attributable to each side
- topical convergence over time



Crupi, Mejova, Tizzani, Paolotti, Panisson. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. ICWSM 2022

Critique

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, *45*(3), 188-206.

Definition of polarization is vague and inconsistent

Most studies are platform-specific (Twitter), possibly not generalizable

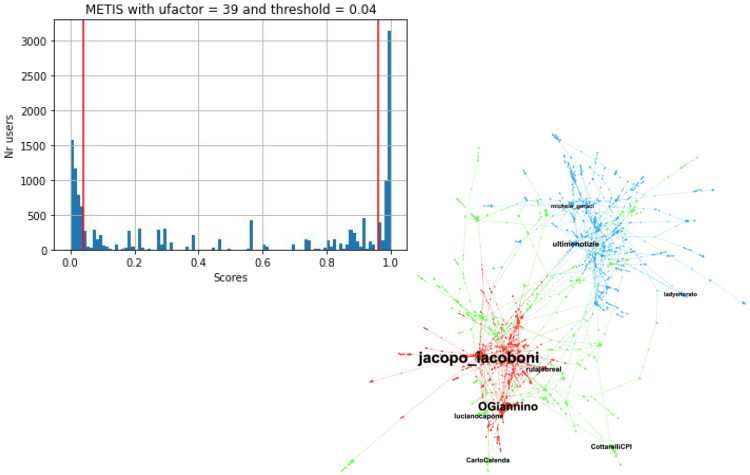
Causality is unclear (studies suggest both ways, and even none)

No standard measures (related to lack of definition)

Demo

1. Stance Detection using Network Partitioning

   Topic: Ukraine-Russia conflict



2. Stance Propagation using Network Clustering

   Topic: Vaccination debate

Discussion

Pay attention to the **limitations and strengths of each approach!**
Lexicons have higher interpretability but have troubles detecting sarcasm.
In large scale analysis the results are stable.

Are there causal effects in the interplay of morality and political views?
We don't know! But **our values do change over time** a
according to our exposure to events.

The moral content of lemmas may vary according to the context!

'brotherhood' has a low impact on betrayal moral the ALM domain
but a considerably higher impact in BLM!

Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C., Kalimeri, K. and Murukannaiah, P.K., 2023, July. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14113-14132).

Discussion

Is detecting the **presence** of a moral foundation enough?
Should Morals foundations be used as a **presence (virtue yes/virtue no), or polar (virtue/vice)** scale?

E.g. Fairness/Cheating
Support for fairness and equality/ Refrain from cheating or exploiting others

C: All citizens should have free access to healthcare
F: Only taxpayer should have access to healthcare

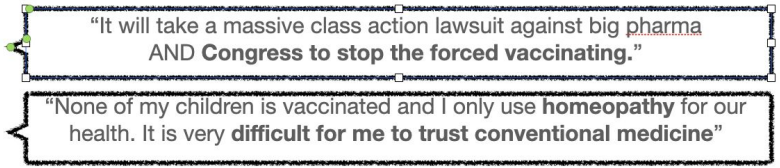
Discussion

Ok, so we assessed morals in text. Now what?

Important for successful communication campaigns!

Health Organisations create campaigns **addressing their own core values (care)!**

However people are concerned about their **Freedom of choice (liberty) & holistic therapies (purity)!**



Kyriaki Kalimeri
ISI Foundation
Ph.D. Brain and Cognitive Sciences
@KyriakiKalimeri

Giovanni Da San Martino
University of Padova
Ph.D. Computer Science
@giodsm

Yelena Mejova
ISI Foundation
Ph.D. Computer Science
@YelenaMejova

Oscar Araque
Universidad Politécnica de Madrid
Ph.D. Telecommunication Engineering
@oaraquei

Webpage to the tutorial