IDC HERZLIYA

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

HYBRID CONTENT ANALYSIS

A COMPUTER-ASSISTED STRATEGY FOR SCALING UP THE THEORY-DRIVEN CLASSIFICATION OF MULTI-PLATFORM SOCIAL MEDIA DATA

Christian Baden – Neta Kligler-Vilenchik – Moran Yarchi

RATIONALE

NEED FOR SCALE

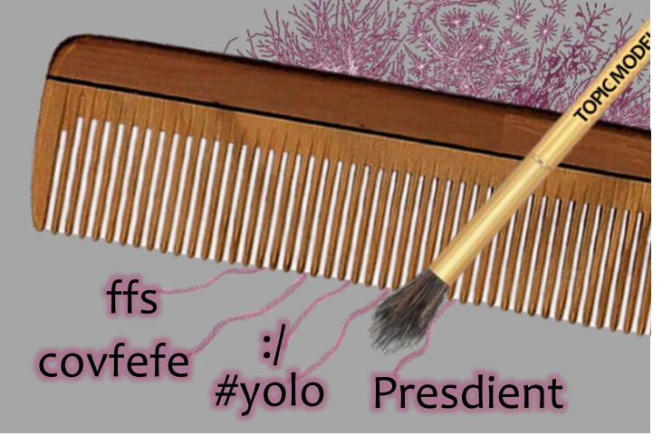
The interactive, unbounded structure of social media conversations bars the use of conventional sampling strategies capable of reducing large scale corpora for use in manual content or discourse analysis:

- intertextual links are broken by random sampling
- fluid topic evolution, ongoing change in participation and multi-platform uses disable a delineation of cases

Analyses of interactive talk often require a scale that is unattainable without computational methods.

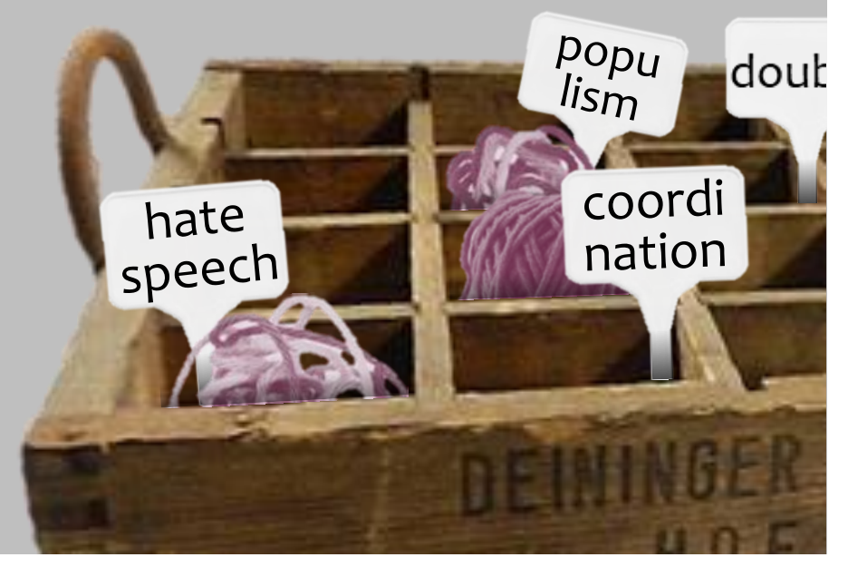
NEED FOR INDUCTIVE EXTRACTION

The unconventional, fast-evolving and imprecise use of language that is characteristic for social media requires analytic techniques capable of treating unanticipated contents. Dictionary-based approaches suffer from low recall, and also supervised machine learning (SVM) fails to capture unseen language uses. Only inductive, unsupervised techniques (e.g., topic models) can adapt to the varied texture of social media discourse.

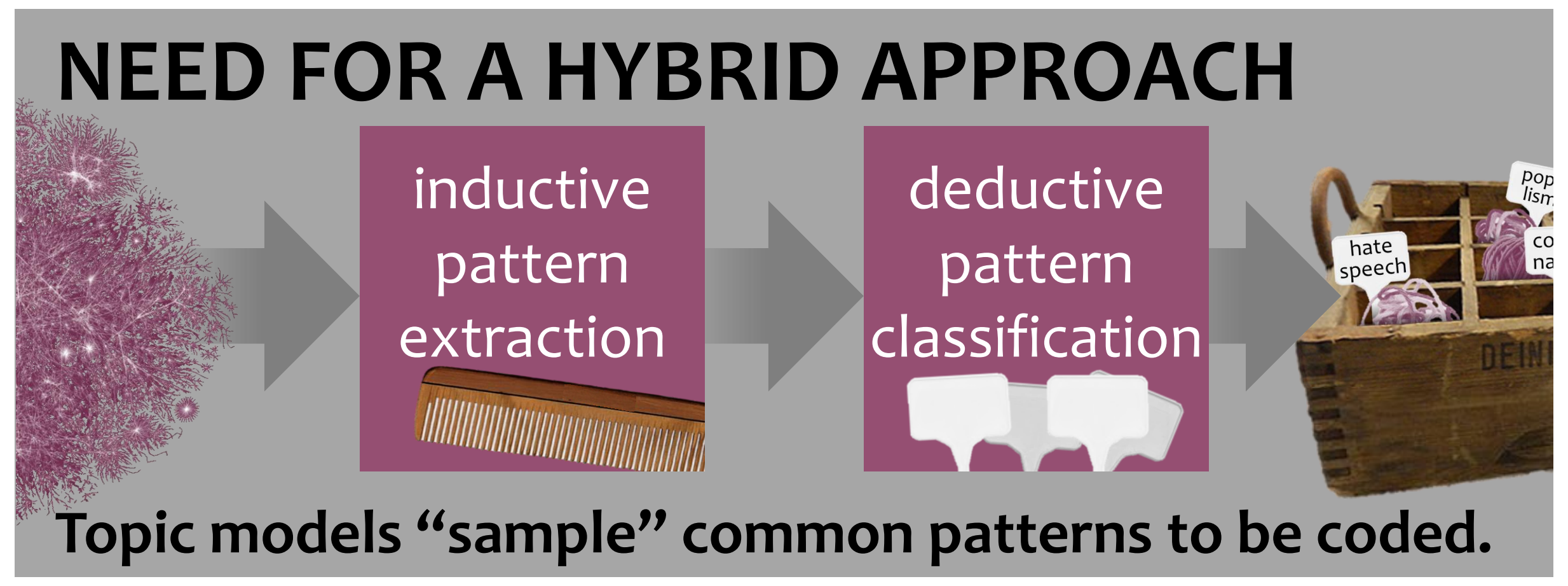


NEED FOR DEDUCTIVE ANALYSIS

Existing **unsupervised tools** used to capture ill-defined patterns in complex textual data cannot be trained to operationalize specific theoretically relevant constructs. **Supervised machine learning** is capable of deductive classification, but incurs considerable costs for nuance and researcher control. **Dictionaries** permit adequate control, but miss any unforeseen instances. To advance theoretical knowledge about social media, there is a need for new techniques that can operationalize theoretically relevant constructs in large-scale social media discourse.



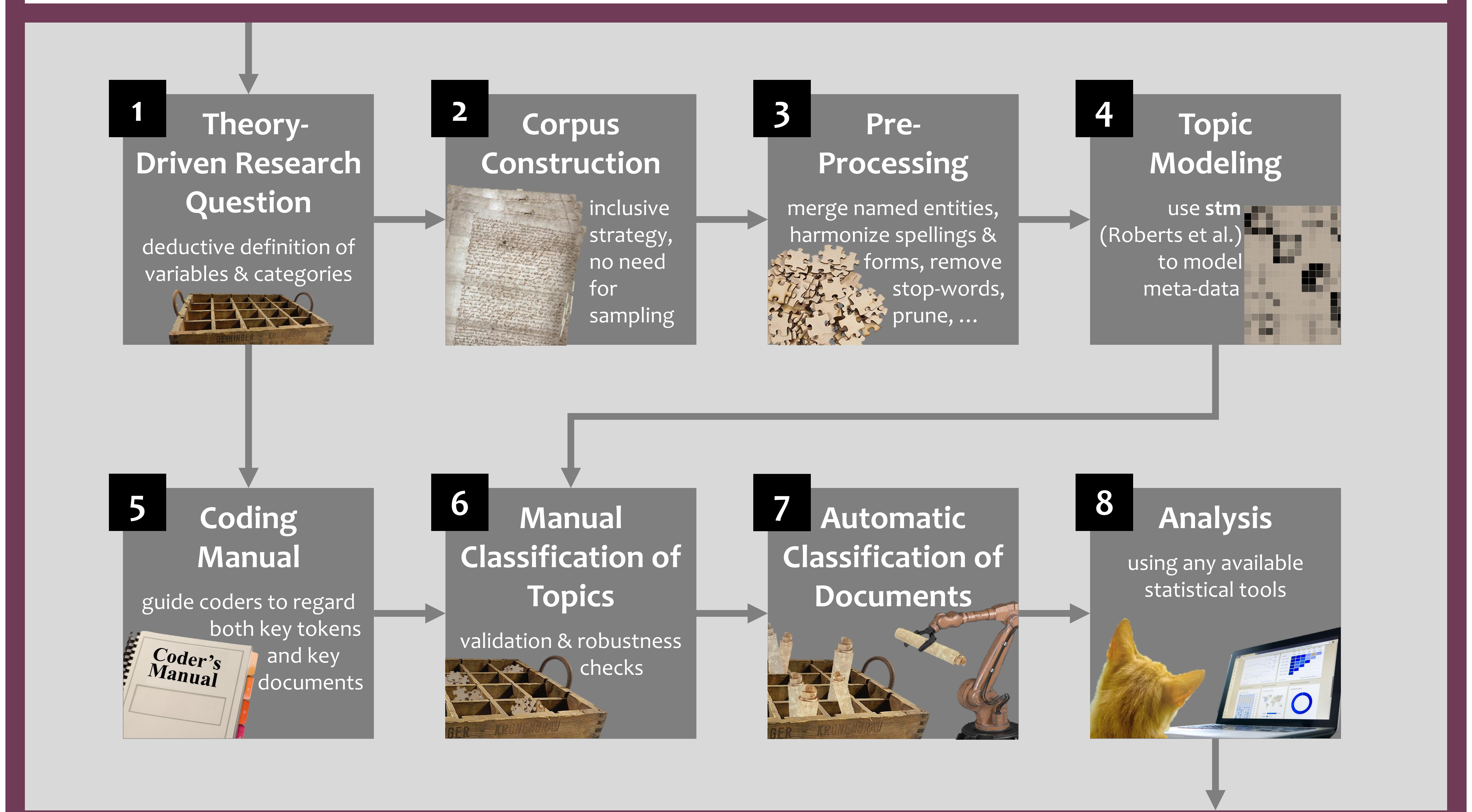
NEED FOR A HYBRID APPROACH


inductive pattern extraction
deductive pattern classification

Topic models “sample” common patterns to be coded.

IDEA IN A NUTSHELL

Hybrid Content Analysis (HCA) uses Topic Models to extract regular patterns from large textual corpora, and subsequently subjects these patterns to manual classification based on a theory-driven codebook. The classification is then transferred back to the original documents based on their use of extracted topics. Thereby, HCA **enables a nuanced, deductive and fully researcher-controlled classification of inductively extracted patterns in huge text corpora.**



1 Theory-Driven Research Question
deductive definition of variables & categories

2 Corpus Construction
inclusive strategy, no need for sampling

3 Pre-Processing
merge named entities, harmonize spellings & forms, remove stop-words, prune, …

4 Topic Modeling
use stm (Roberts et al.) to model meta-data

5 Coding Manual
guide coders to regard both key tokens and key documents

6 Manual Classification of Topics
validation & robustness checks

7 Automatic Classification of Documents

8 Analysis
using any available statistical tools

VALIDATION N = 200 documents,

HCA classification vs. manual coding
Precision: 0.89 (SD = 0.10)
Recall: 0.89 (SD = 0.07)

ROBUSTNESS all documents,

HCA based on joint topic model vs. separate topic models per platform
Holsti: 0.80 (SD = 0.18)

Hybrid Content Analysis (HCA) is specifically suited to the analysis of interactive social media conversations:

- preserves intertextual links, as no sampling is required
- inductively organizes innovative language uses
- classifies short snippets based on common embeddings
- models distinct conventions in multi-platform discourse

APPLICATIONS & WAY FORWARD

HCA opens up new avenues for the theory-guided study of interactive discourse at scale.

STRENGTHS/LIMITATIONS

- ⊕ full deductive researcher control
- ⊕ nuanced, manual classification
- ⊕ picks up on unforeseen patterns
- ⊕ hardly affected by robustness issues
- ⊕ probabilistic/multiple classification
- ⊕ capacity to treat very short texts
- ⊖ more difficult coding of topics
- ⊖ some residual robustness issues

DEMONSTRATION

CASE STUDY: The Hebron Shooting

On 24 March 2016, a Sergeant of the Israeli army shot dead a Palestinian assailant after he had already been disarmed and neutralized. In the ensuing trial, the Israeli public polarized between those condemning the extrajudicial killing, and those defending the soldier’s acts as legitimate defense against a terrorist attack.

RQ: Interpretative Polarization

To what extent do social media users focus on different issues depending on their stance toward the incident?

DATA: Facebook, Twitter, WhatsApp

We obtained all relevant conversations on Twitter, all Israeli public Facebook pages, and in two political discussion groups on WhatsApp (24.03.16 – 02.10.17).

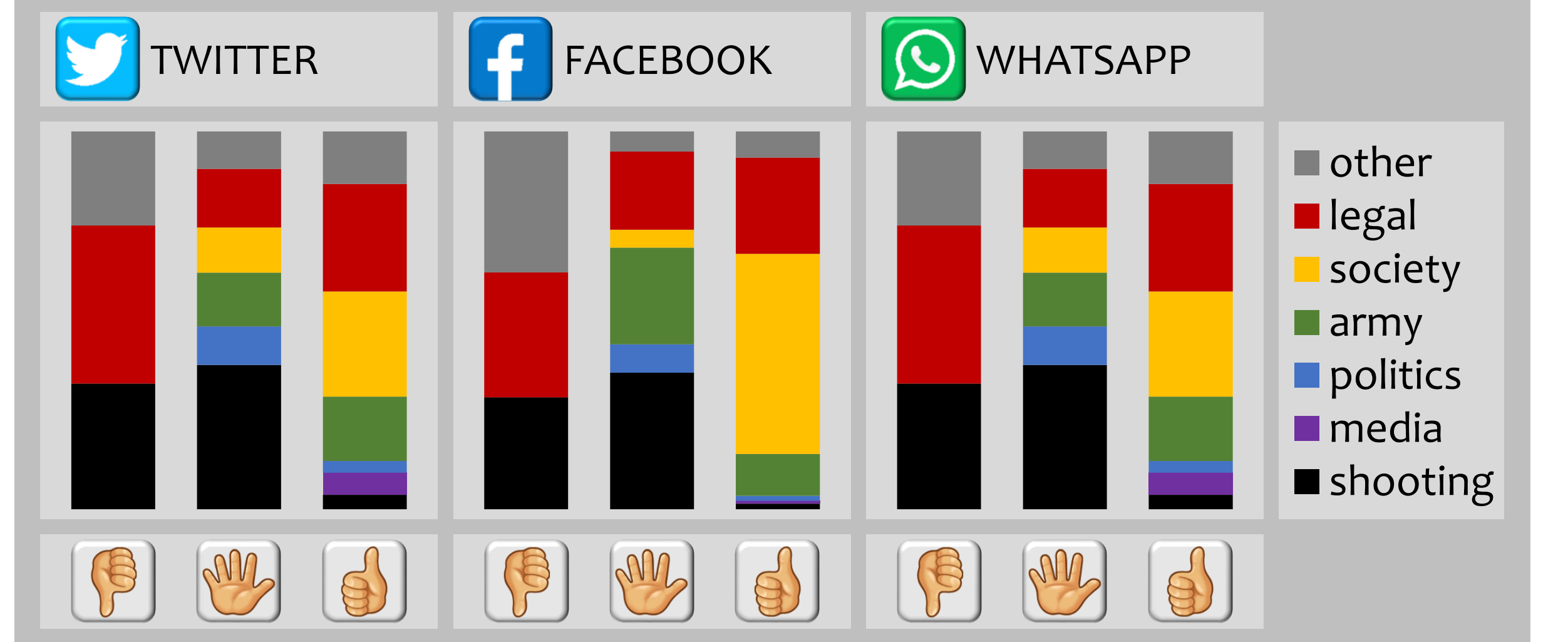
	Twitter	Facebook	WhatsApp
Posts:	29,250	6,508	6,245
Comments:	61,772	145,542	

HYBRID CONTENT ANALYSIS

We removed retweets/reposts. Data were tokenized, acronyms/spellings/emojis harmonized, named entities concatenated and stop words removed. One topic model (stm, k = 100) was selected for the combined data. We also ran separate models with k = 80, 80, 70 for each model. Topics were coded and documents classified based on their use of topics if the weight of one combined category exceeded 0.5.

FINDINGS

The data shows consistent interpretative polarization:



Supporters of the defendant (👍) relied on very different issues than opponents (👎) and ambivalent voices (✋).

This poster has received the
ICA Political Communication Top Poster Award 2019

Sunday 26 May 2019 | 17:00-18:15 | International Terrace | 5735 Political Communication Interactive Poster Session | No. 62 | 69th Annual Conference of the International Communication Association | “Communication Beyond Boundaries” | Washington, DC, USA