Linking Surveys and Social Media Data: Shaping the Future of Public Opinion Research

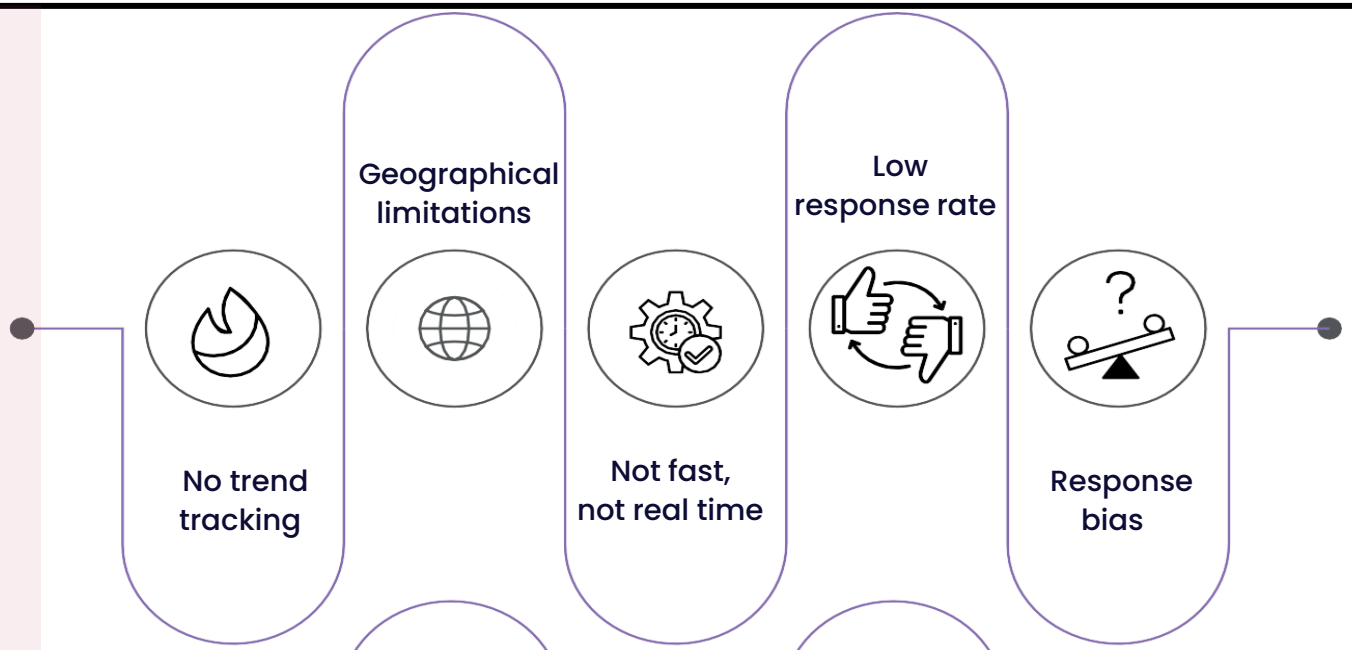
Politus Analytics

Şükrü Atsızelti & M. Fuat Kına

Koç University

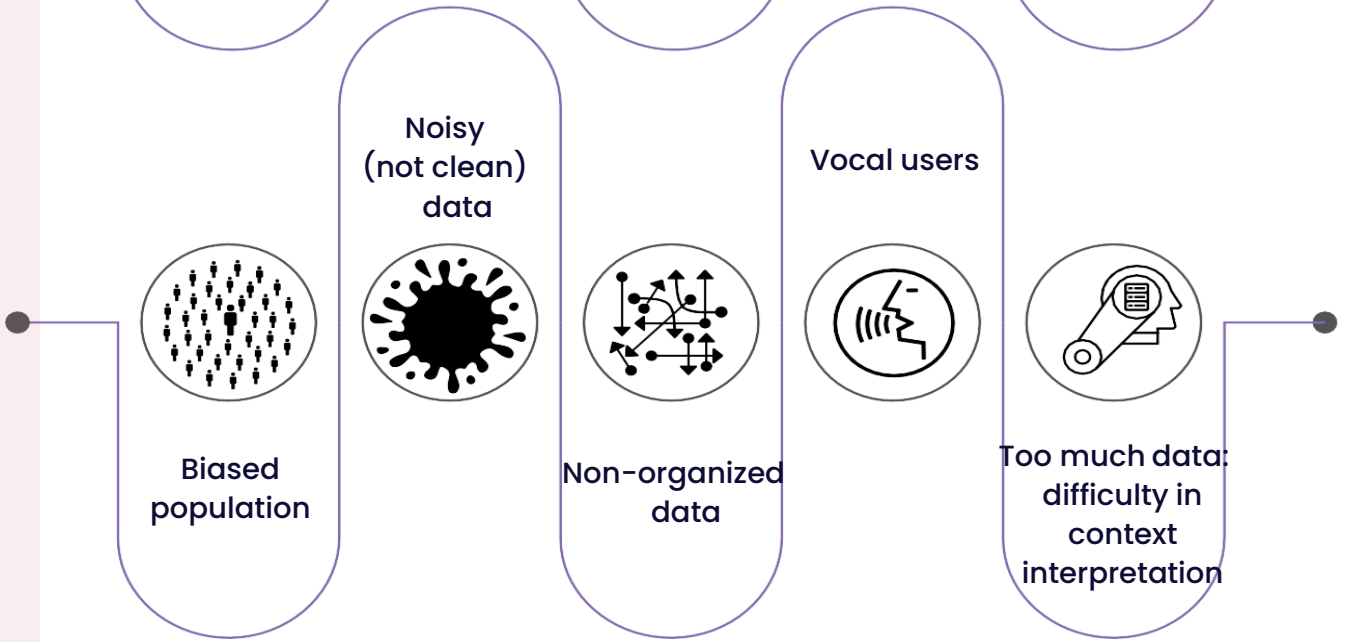
CLASSICAL SURVEYS

Classical surveys have problems and limitations in understanding **public opinion** in a fast-moving world



SOCIAL MEDIA MONITORING

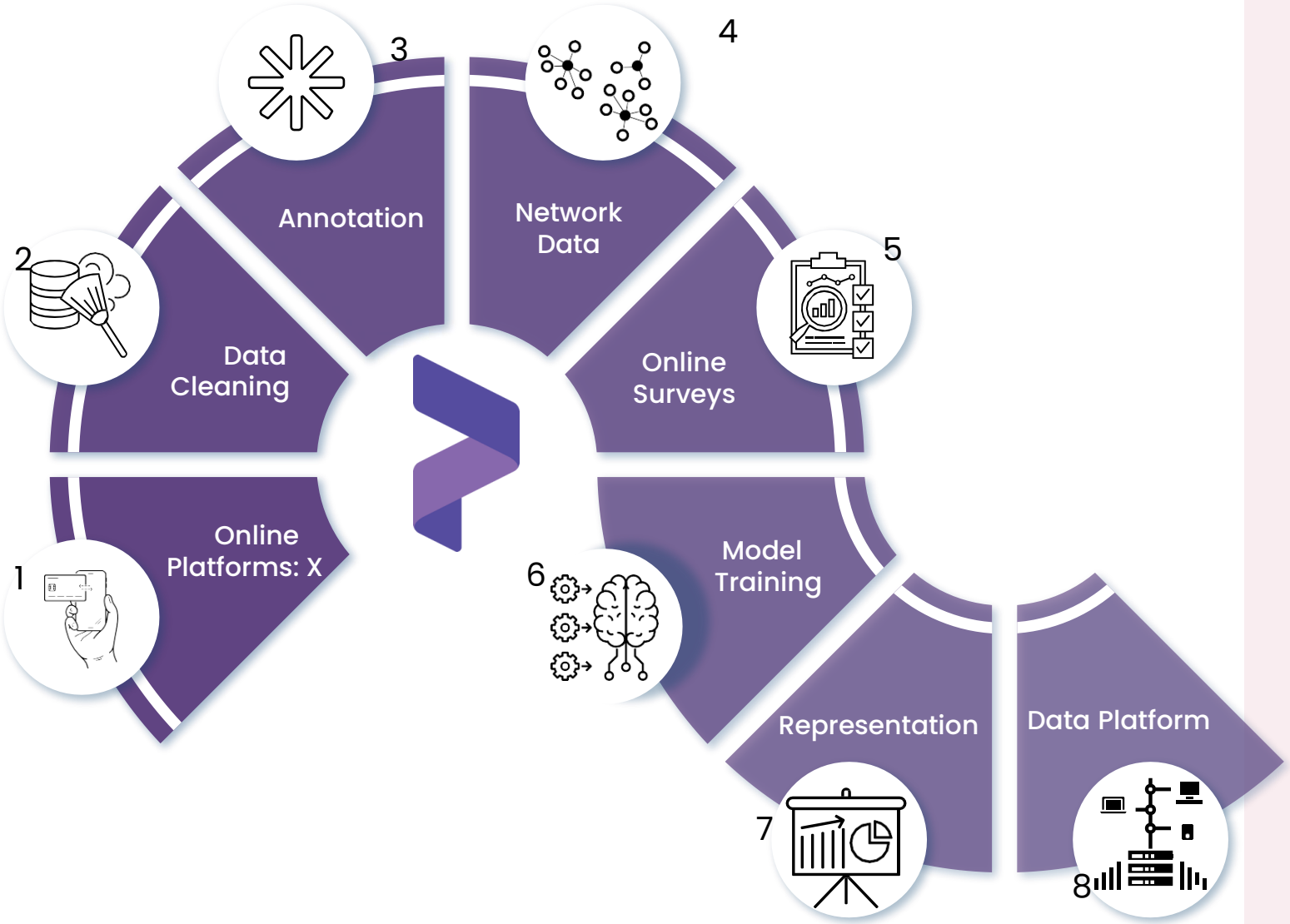
Social media monitoring / listening enables catching up with the high pace yet bringing data-based problems in itself.



Politus Project

ERC proof of concept grant that processes digital trace data to offer timely insights into social and political trends

Aim: To extract GDPR-compliant public opinion from online platforms using AI



Methodology

NLP, Machine Learning, Network Analysis

Data

Social Media Data
Text Data

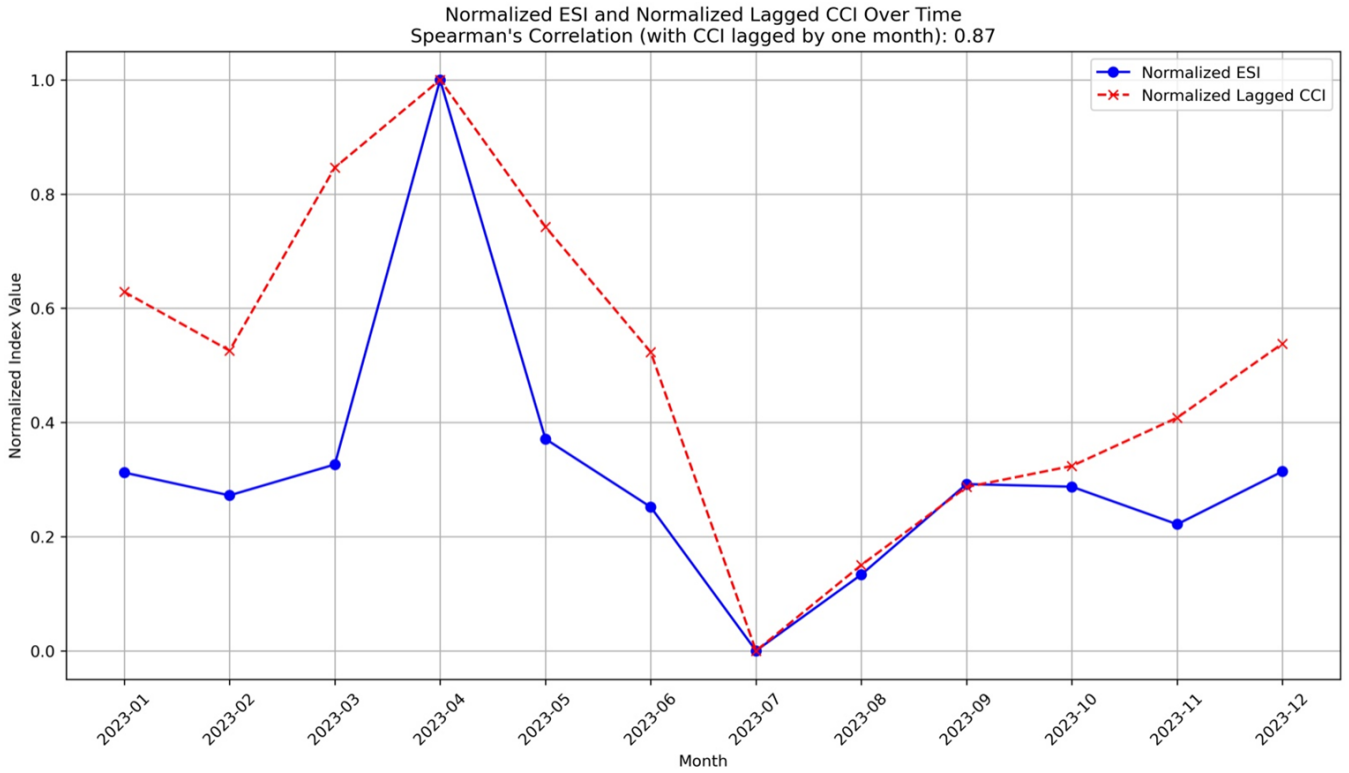
Data Platform

Demography,
Geographical Breakdown
Emotions,
Values,
Topics,
Approval Rates,
Time Trends,
Sectors

An example: Predicting Consumer Confidence Index

CCI and public emotions about the topic of “economy”: anger, hope, despair

Spearman's rank correlation (with CCI lagged by one month): 0.87



Methodology of Linking

Social Media Data Collection

Online Survey

Annotation and Model Building

Tweet-to-User Transformation

Linking Survey and Social Media Data

Data at Hand

Data at Hand: Twitter

Adding users in time

Currently:

6,681,771 account

718,867,236 tweets

Hand-Picked 100 prominent accounts

55 million follower Ids

Filtration (language and location)

3.5 million users

Data at Hand: Survey

TWITTER ADS

GENERAL TURKISH POPULATION (NO TARGETING)

Data at Hand: Survey

- We tried Facebook and Twitter, we were successful on Twitter.
- Advertised via ccss_ku account
- Qualtrics
- Questions ranging from political preferences to education, tweeted topics to stances toward politicians

Data at Hand: Survey

- The advertisement received over 100 thousand views.
- 8500 people clicked on the link.
- Approximately 2000 completed the survey.
- 1000 of which are successfully linked.
- 10,000 TL was given.

Some validation scores

- Demographics:
  - **Gender** f1-weighted: 0.79
  - **Age** f1-weighted: 0.56 (four categories)
  - **Location** f1-weighted: 0.79
  - This led us to concern about age detection method.

- Ideology:
  - At least one tweet, at least 5 tweets, or dynamic threshold.
  - We tried different thresholds for ideologies based on frequency of ideology;
  - And for users depending on users' activity.
  - The scores on Ideology were quite low, around 50-60% on average.

Challenges

Challenges related to social media data

- Missing participants
  - Many survey participants have out-of-reach locked accounts.
- It became limited and costly
  - Like many other social media platforms, Twitter limited the reach to data and make it APIs costly.
- Building models for a left-out platform
  - Twitter-based models is hardly generalizable to other platforms and due to new costs of Twitter data, building models on it may not be a rational investment.

Challenges related to online survey

- Changing platform behavior
  - Ads methods and platform behavior have been changed a few times during the last year (e.g. giving ad to an ID list option)

- Company vs. Local Legal System
  - Due to incoordination between Turkish authorities and Twitter, giving ad is impossible for Turkish citizens anymore.

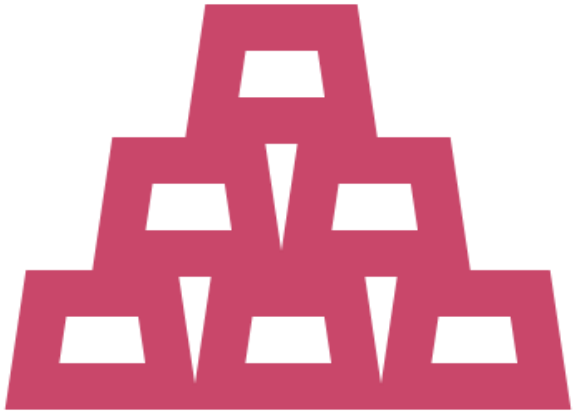
- Anonymization Problems
  - Anonymization techniques are not still ready for fully anonymized, repeated surveys linked with social media posts without compromising the data quality.

Challenges Related to Linkage

- It is hard to link all parts of the data on survey and social media.
- They may be showing different phenomena.

Example: User ideologies determined with thresholds and users' remarks are incompatible.

A question: Which one is ground-truth?



Challenges Related to Model Building

- Survey data is highly imbalanced in terms of politics and education.
- Result: poor machine learning model outputs
- Solution a) Data imputation
  - Known imputation techniques didn’t yield hopeful results for education prediction for now.
- Solution b) Simulating Data with LLMs
  - As a rising method, simulating data with LLMs and using them for missing data and balancing the data, may open new doors in social sciences.

Conclusion

- Demographics prediction models are planned to be tuned by survey data.
- Self-reported attitude (survey) vs observed behavior (tweets)
- Validating or complementing
  - We couldn’t validate the outputs of ideology models via survey, but we validated them with nationwide election results. Survey and Tweets may indicate different phenomena.

- Our future tasks:
  - Adding expert view to self-reported attitude and observed behavior as a third source of data
  - Using LLM for filling missing data and data imputation via simulation

Thanks for listening