Political Tweet Sentiment Analysis for Public Opinion Estimation

A. Kaimakamidis, Prof. I. Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 2.0

Artificial Intelligence & Information Analysis Lab

Public Opinion Estimation

- **Problem statement**
- Heuristic popularity score estimation
- Regression popularity score estimation

Problem statement

Social media vs polls

Social Media Analytics

- Advantages:
  - Live, every day feedback.
  - Cover a big part of society.
  - Low cost data acquisition.
- Disadvantages:
  - Imprecise political views.
  - Can be easily biased.
  - Offensive political speech.

Public Opinion Polls

- Advantages:
  - Carefull population sampling.
  - Rather low estimation errors.
- Disadvantages:
  - Expensive.
  - Cannot provide every day results.
  - Occasional failures.

Problem statement

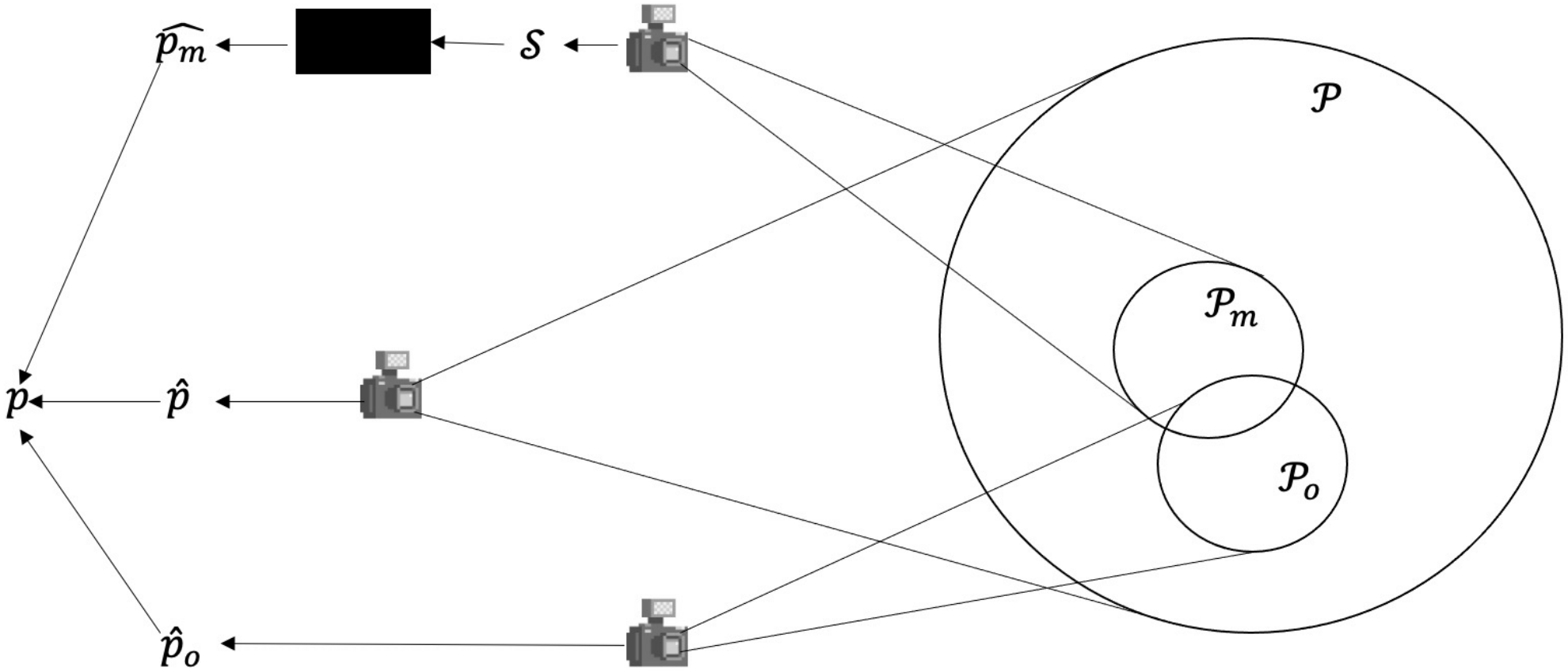
- Let $\mathcal{P}$ be the total population set.
- $\mathcal{P}_m$ be the people that are politically active in social media.
- $\mathcal{P}_o$ be the people participating in a public opinion poll.

- Let set $\mathcal{S}$ consist of the total numbers of positive, neutral and negative political tweets about a political entity.
- Aim: Estimation the ***popularity score distribution***

$$\mathbf{p}^T = [p_1, \cdots, p_n]$$

for n political entities.

Problem statement



Problem statement

- Set $\mathcal{P}$ can be sampled only in special occasions
  - ***election***/***referendum*** procedures.
- Opinion polling for set $\mathcal{P}_o$ are accurately but rather costly and facing difficult problems:
  - correct ***population sampling***.
- ***Popularity score distribution*** estimation for set $\mathcal{P}_m$ would constitute a cheaper and daily solution for estimating the popularity score distribution.

Problem statement

Twitter Data Gathering

- Twitter API has been used for tweet gathering from 14 June 2022 until 31 Dec 2022.
- More than 300,000 tweets have been gathered about six Greek political parties, currently in Greek parliament.

Parties	neutral	positive	negative
ND	48,238	5,449	35,014
SYRIZA	83,329	9,677	86,473
KINAL	20,774	4,645	8,090
KKE	10,142	2,937	4,016
ELLINIKI LISI	6,819	1,249	1,390
MERA25	1,585	539	524
Total	170,887	24,496	135,507

Public Opinion Estimation

- Problem statement
- **Heuristic popularity score estimation**
- Regression popularity score estimation

Heuristic popularity score estimation

Two candidate elections (candidates i, j):

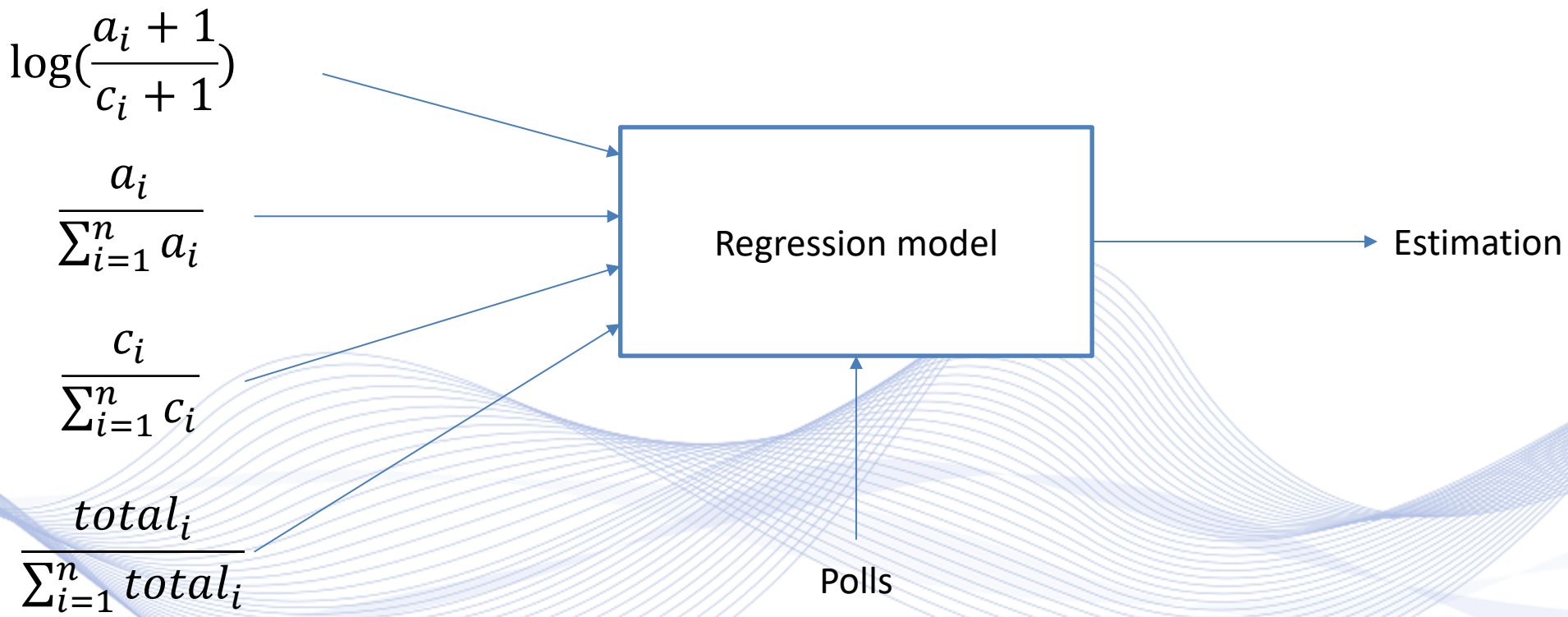
- $\mathbf{a}, \mathbf{c}$: n - dimension vector where a_i, c_i represent the total number of positive, negative tweets for party i respectively.
- $total = a_i + a_j + c_i + c_j$.
- Estimators [BOV2018], [CON2010], [WAN2017]:
  - $p_i(\mathbf{a}, \mathbf{c}) = \frac{a_i + c_j}{total}$.
  - $p_i(\mathbf{a}, \mathbf{c}) = \frac{a_i}{c_i}$.
  - $p_i(\mathbf{a}, \mathbf{c}) = \frac{a_i}{a_i + c_i} \frac{total_i}{total}$.

Heuristic popularity score estimation

Multi candidate elecations (n candidates):

- $popularity_i(\boldsymbol{a}, \boldsymbol{c}) = \frac{a_i}{c_i}$ [CON2010].
- $popularity_i(\boldsymbol{a}, \boldsymbol{c}) = \frac{a_i}{a_i + c_i} \frac{total_i}{total}$ [WAN2017].
- $popularity_i(\boldsymbol{a}, \boldsymbol{c}) = \log(\frac{a_i + 1}{c_i + 1})$ [BER2017].
- $popularity_i(\boldsymbol{a}) = \frac{a_i}{\sum_{i=1}^{n} a_i}$ [BAN2018].
- $popularity_i(\boldsymbol{a}, \boldsymbol{c}) = \frac{total_i}{\sum_{i=1}^{n} total_i}$ [TUM2010].

Heuristic popularity score estimation



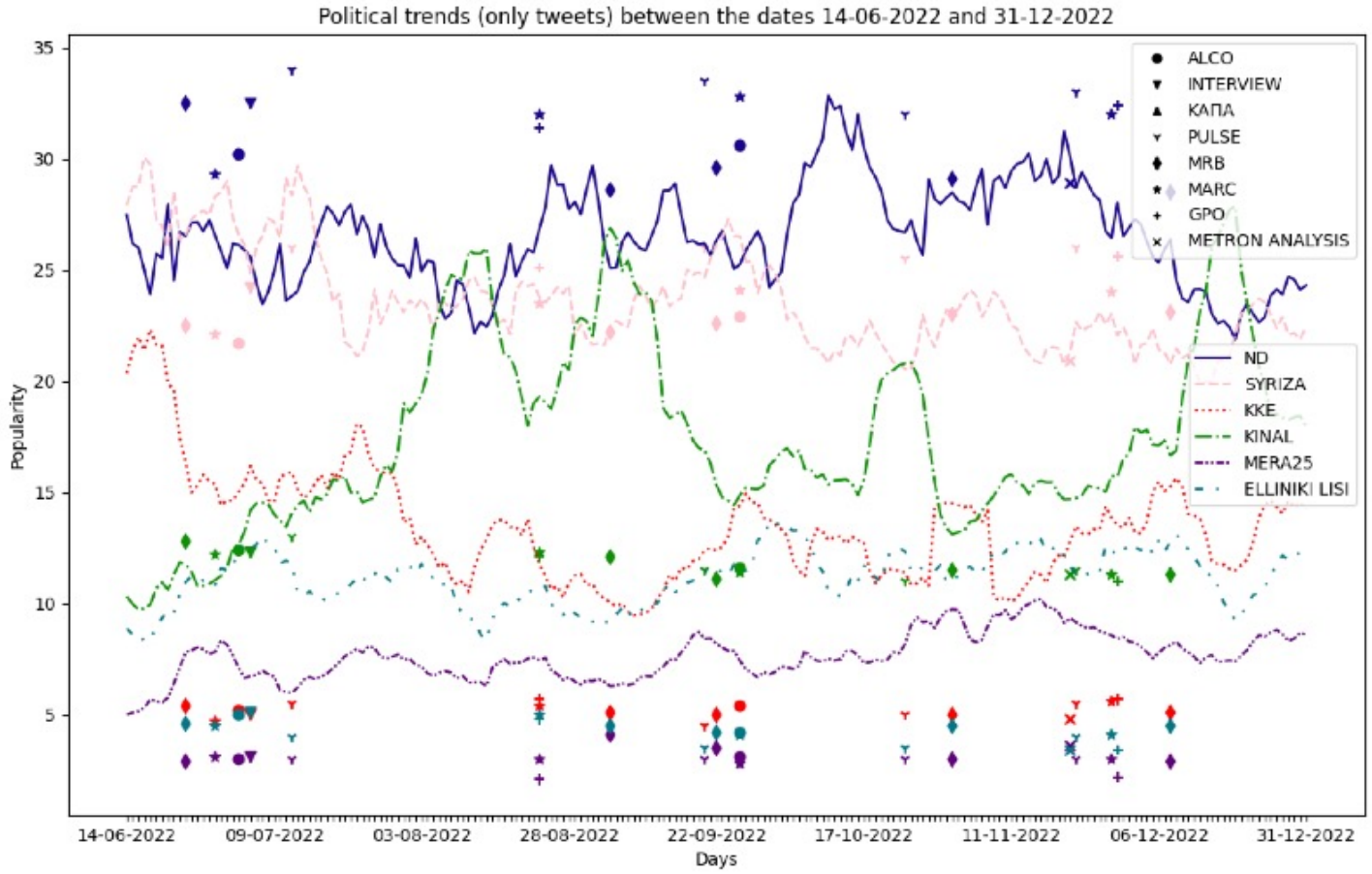
Regression scheme [BER2017].

Heuristic popularity score estimation

Proposed method for estimating the popularity score distribution $\mathbf{p}_m$.

- Let $\mathbf{a}, \mathbf{b}, \mathbf{c}$ be n - dimension vectors, where a_i, b_i, c_i represent the total number of positive, neutral, negative tweets for party i respectively.
- The number of negative tweets c_i for party i is distributed to other parties j in proportion of their positive and neutral tweet numbers a_j, b_j.

Heuristic popularity score estimation



Heuristic popularity score estimation

Estimators	200 days	180 days	150 days	100 days
Proposed	5.07%	5.23%	5.3%	5.3%
$\frac{a_i}{c_i}$	20.73%	20.51%	20.38%	20.5%
$\log(\frac{a_i+1}{c_i+1})$	18.7%	18.87%	18.96%	19.14%
$\frac{a_i}{\sum_{i=1}^{n} a_i}$	7.01%	6.91%	7%	7.29%
$\frac{a_i}{a_i + c_i} \frac{total_i}{total}$	9.39%	9.5%	9.47%	8.95%
$\frac{total_i}{\sum_{i=1}^{n} total_i}$	6,44%	6.91%	7.22%	7.9%

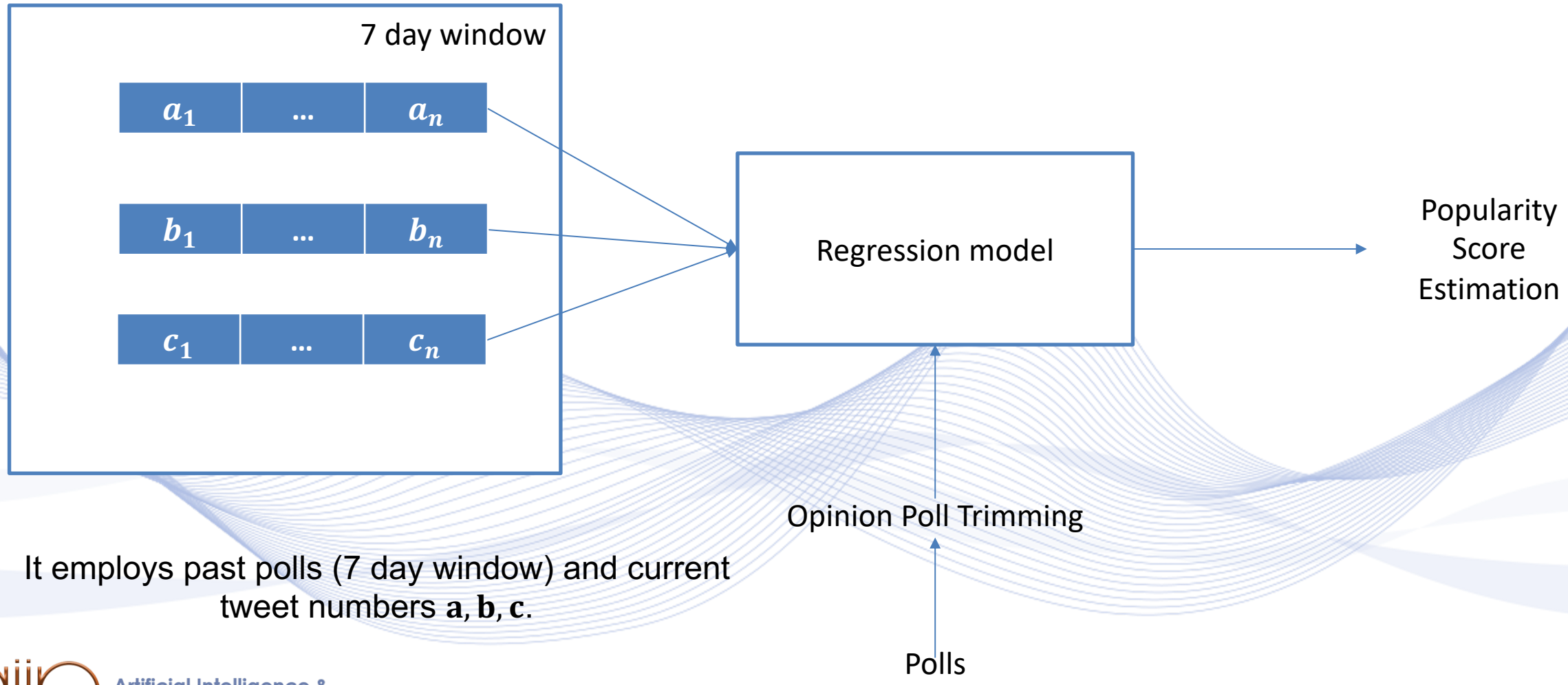
Parties	Proposed	$\frac{a_i}{\sum_{i=1}^{n} a_i}$	$\frac{total_i}{\sum_{i=1}^{n} total_i}$	Polls
ND	0.265	0.214	0.272	0.314
SYRIZA	0.237	0.385	0.524	0.238
KINAL	0.173	0.173	0.1	0.119
KKE	0.138	0.138	0.062	0.053
ELLINIKI LISI	0.112	0.068	0.033	0.043
MERA25	0.076	0.021	0.009	0.03

Comparison of heuristic popularity score estimators.

Public Opinion Estimation

- Problem statement
- Heuristic popularity score estimation
- **Regression popularity score estimation**

Regression popularity score estimation



It employs past polls (7 day window) and current tweet numbers $\mathbf{a}, \mathbf{b}, \mathbf{c}$.

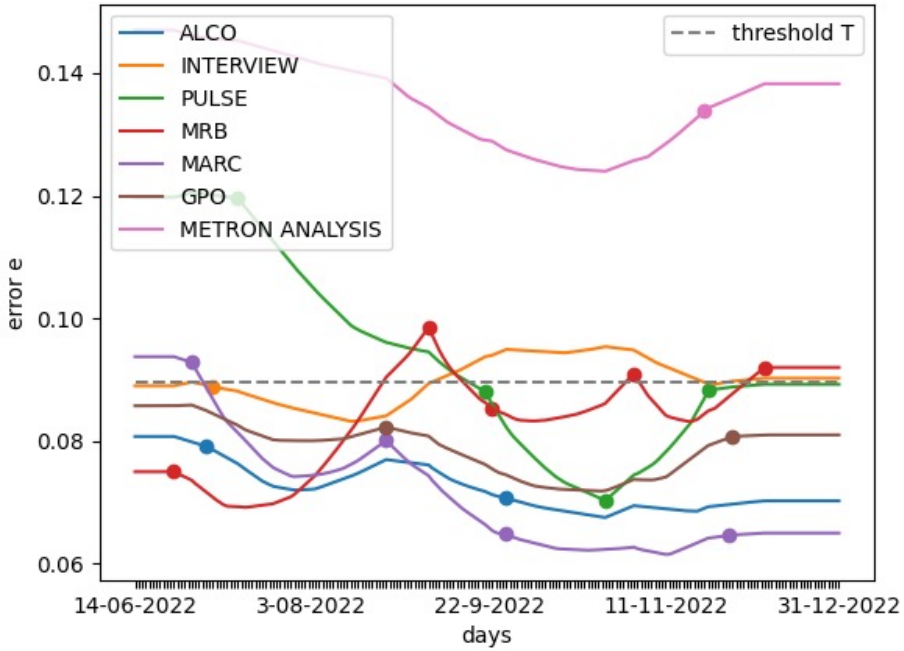
Regression popularity score estimation

Opinion Poll Trimming

- Certain opinion polls of some polling companies may be ***outliers***.
- In order to have 'ground truth' opinion polls for regression, we trim out poll outliers.
- As opinion polls are conducted at various dates by various companies:
  - ***linear temporal interpolation*** is applied to their poll results and in order to compare them.

Regression popularity score estimation

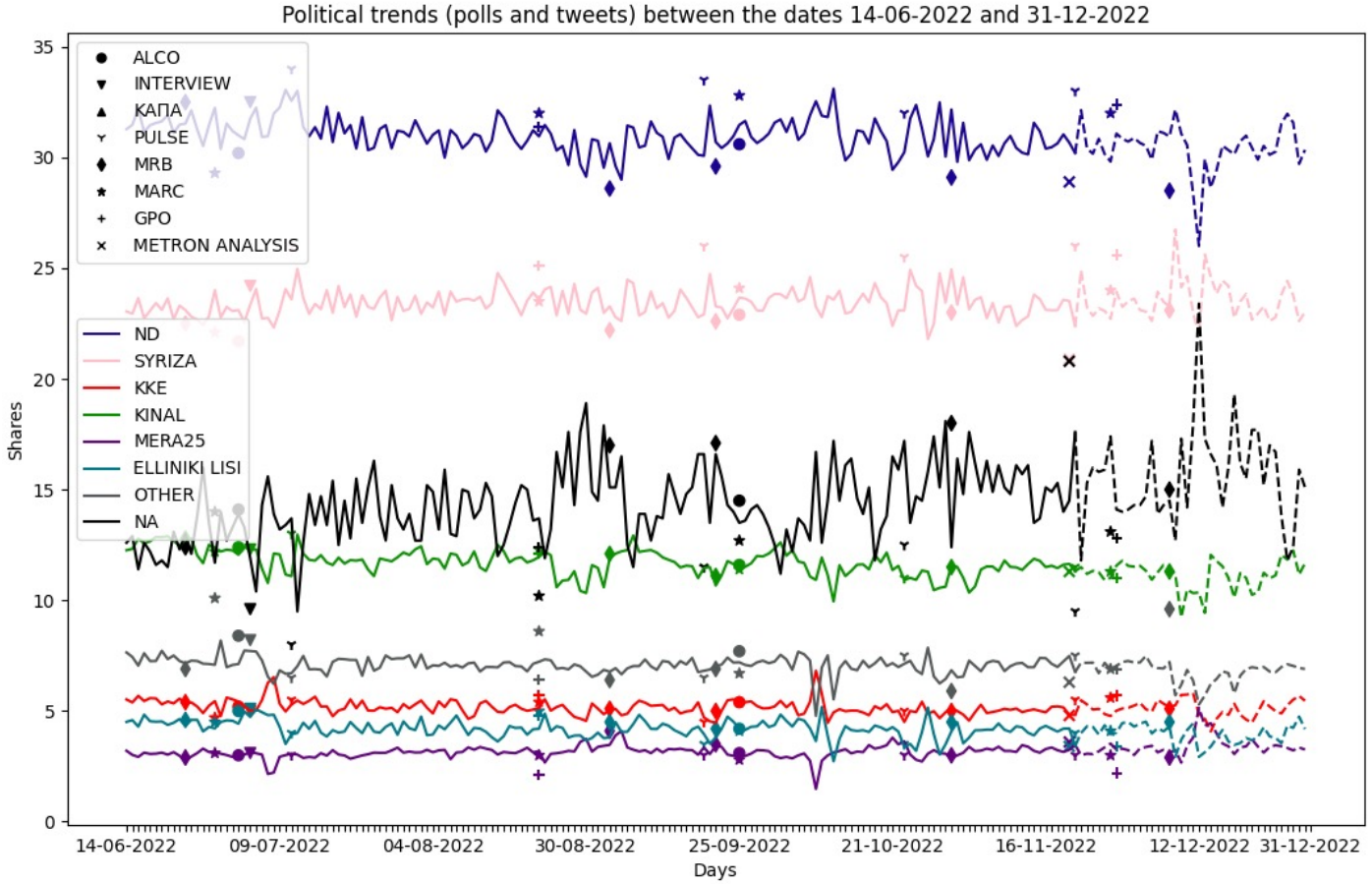
- The **MAE** error e_i is measured between the polls of company i and the rest of the companies for each date.
- If e_i is above a threshold T, it is excluded from regression.



Regression popularity score estimation

- During the examined period we managed to access 19 polls from various companies.
- Four polls were kept for testing.
- We trained two regression models in two ways:
  1. Opinion poll regressor (OPR), trained with 15 polls.
  2. Opinion poll regressor (OPR) with opinion poll selection, trained with 10 polls.

Regression popularity score estimation



Popularity score regression results.

Regression popularity score estimation

MAE of popularity score regression results.

Testing polls	[BER2017] proposition	OPR	OPR outlier trimming
23/11/2022	1.67%	1.43%	0.91%
29/11/2022	0.84%	0.61%	0.22%
30/11/2022	1.33%	1.3%	0.72%
9/12/2022	1.14%	1.31%	1.85%
Average MAE	1.245%	1.163%	0.925%

Conclusions

- Twitter data analysis can predict political trends rather accurately.
- Its results are biased when trying to estimate public opinion based only on Twitter data.
- The proposed popularity score regression uses twitter and past opinion poll data.
- It outperforms classical popularity score estimators.

Bibliography

[1] I. Pitas, “Artificial Intelligence Science and Society Part A: Introduction to AI Science and Information Technology“, Amazon/Kindle Direct Publishing, 2022,
https://www.amazon.com/dp/9609156460?ref_=pe_3052080_397514860

[2] I. Pitas, “Artificial Intelligence Science and Society Part B: AI Science, Mind and Humans“, Amazon/Kindle Direct Publishing, 2022,
https://www.amazon.com/dp/9609156479?ref_=pe_3052080_397514860

[3] I. Pitas, “Artificial Intelligence Science and Society Part C: AI Science and Society”, Amazon/Kindle Direct Publishing, 2022,
https://www.amazon.com/dp/9609156487?ref_=pe_3052080_397514860

[4] I. Pitas, “Artificial Intelligence Science and Society Part D: AI Science and the Environment“, Amazon/Kindle Direct Publishing, 2022,
https://www.amazon.com/dp/9609156495?ref_=pe_3052080_397514860

Bibliography

[BOV2018] A. Bovet, F. Morone, H. A. Makse, “Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump,” Scientific Reports, vol. 8, no. 1, pp. 8673, June 2018.

[CON2010] Brendan O’Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith, “From tweets to polls: Linking text sentiment to public opinion time series,” 01 2010, vol. 11.

[WAN2017] Lei Wang and John Q. Gan, “Prediction of the 2017 french election based on twitter data analysis,” in *2017 9th Computer Science and Electronic Engineering (CEEC)*, 2017, pp. 89–93.

Bibliography

[BAN2018] Barkha Bansal and Sangeet Srivastava, “On predicting elections with hybrid topic based sentiment analysis of tweets,” Procedia Computer Science, vol. 135, pp. 346–353, 2018.

[TUM2010] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment,” Proceedings of the International AAAI Conference on Web and Social Media, 2010.

[BER2017] Adam Bermingham and Alan Smeaton, “On using Twitter to monitor political sentiment and predict election results,” in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, Chiang Mai, Thailand, Nov. 2011, pp. 2–10.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr