Predicting Survey Response with Quotation-based Modeling

A Case Study on Favorability towards the United States

Speaker: Alireza Amirshahi

Co-authors:

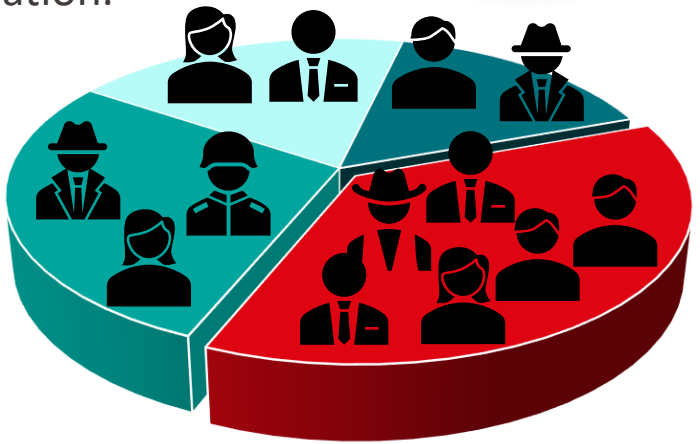
Nicolas Kirsch, Jonathan Reymond, Saleh Baghersalimi

École Polytechnique Fédérale de Lausanne (EPFL)

10th IEEE Swiss Conference on Data Science (SDS2023)

Importance of Surveys

- Survey: A method for collecting information from a group of individuals by asking them questions.
- Surveys provide valuable insights for strategic decision-making and risk reduction.
- Surveys offer high representativeness, allowing for a better understanding of the general population.



Challenges in Conducting Surveys

- Cost constraints
- Time-consuming process
- Low response bias
- Privacy concerns
- Survey fatigue
- Technological accessibility
- Language and cultural barriers
- …

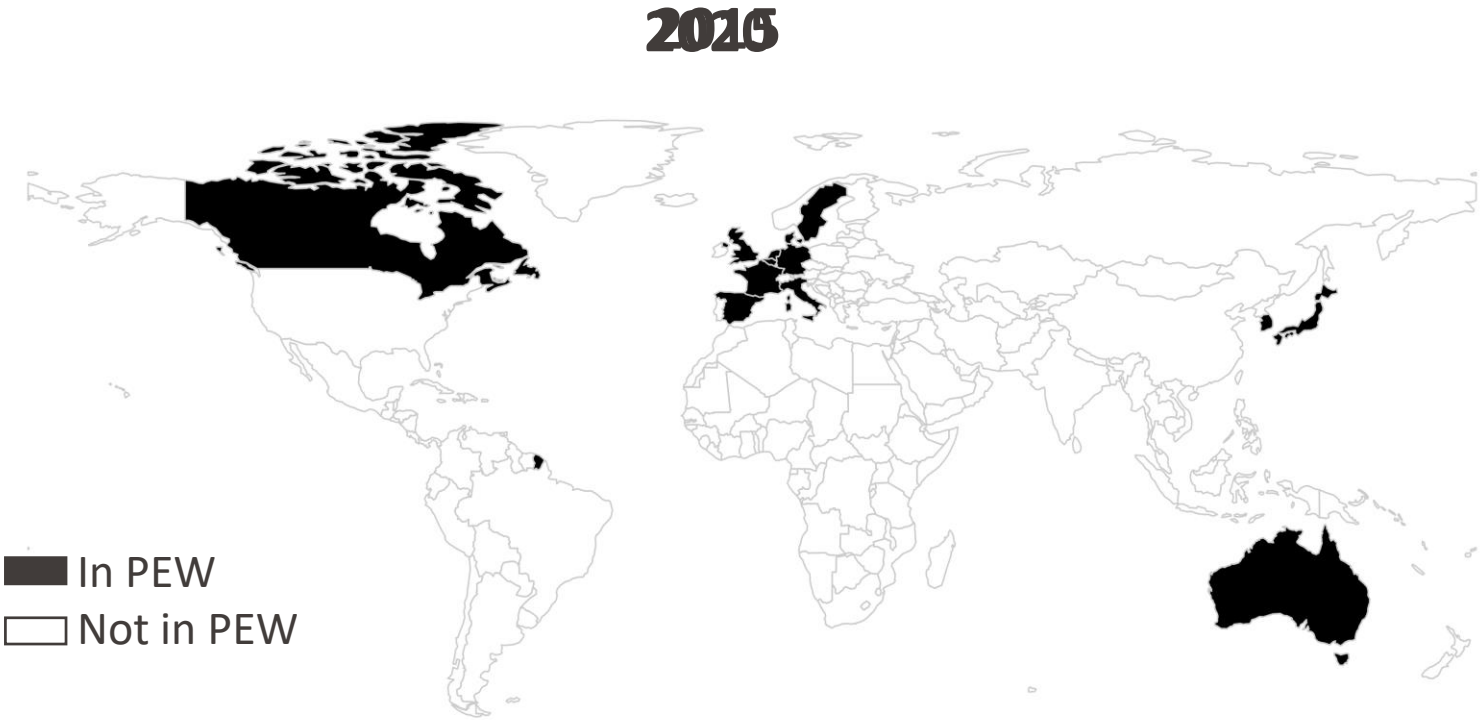
Overview of Our Proposed Solution

PEW Dataset

- PEW Research Center: A nonpartisan institution that informs the public about the global issues, trends, and public attitudes
- PEW global attitudes across 70 countries since 2001
- Survey questions:
  - Public life
  - Religion
  - Internet and technology
  - Economic situation
  - Favorability of the countries and international organizations
  - …
- In this work we focus on U.S. favorability as a case study without loss of generality

https://www.pewresearch.org/

The Problem: Missing Data



Quotebank Dataset

- A corpus of quotations from a decade of news
- 178 million unique attributed quotations
- Over 900,000 distinct speakers
- 377,000 unique web domains

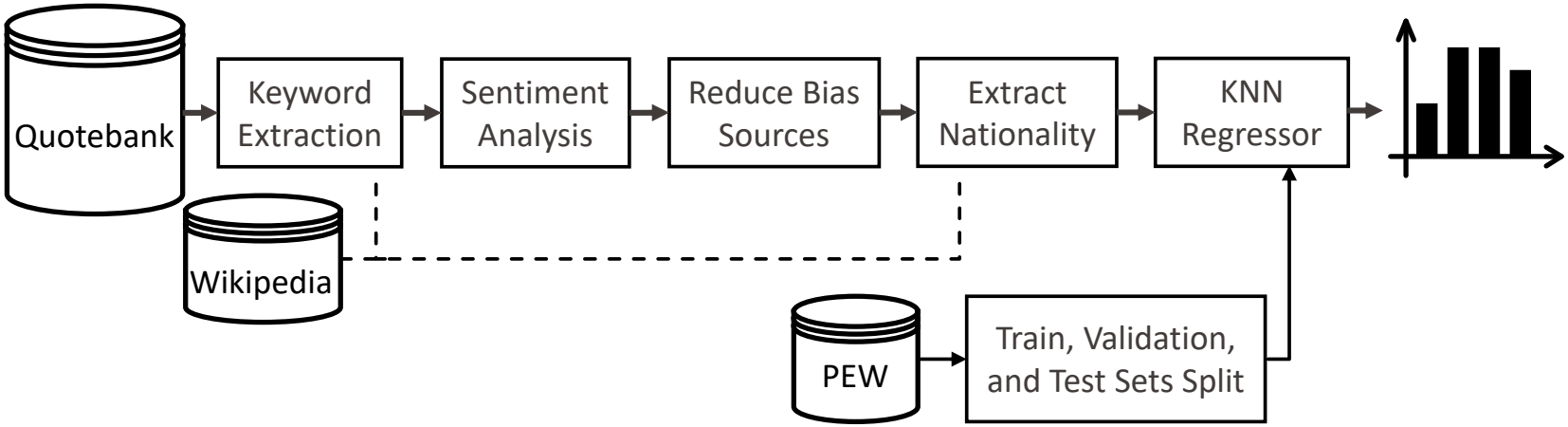
Example

Interior Minister Alain Berset, who holds the rotating Swiss presidency this year, told parliamentarians that the clock was ticking and "the government was forced to act, in the interest of the country, the institutions and the national economy"

www.swissinfo.ch

Vaucher, Timoté, et al. "Quotebank: a corpus of quotations from a decade of news.", WSDM2021

Our Framework



Keyword Extraction

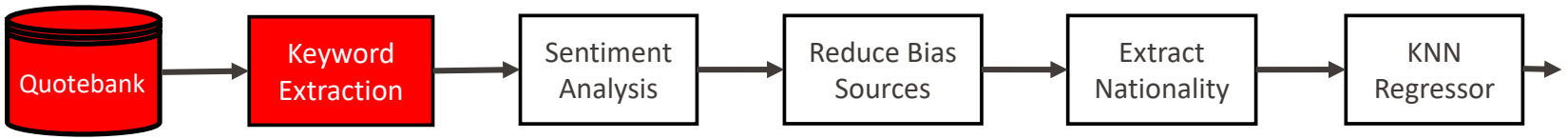
- Manual keywords: ‘US’, ‘U.S.’, ‘USA’, and ‘United States’

- Enriched keywords
  - Most-frequent American speakers in the dataset
  - If their names are in the quotations

Example

“Protects the United States from future intrusions on the United States' sovereignty”
Donald Trump,
2 June 2017

“I respect the decision of President Trump,”
Emmanuel Macron,
14 July 2017

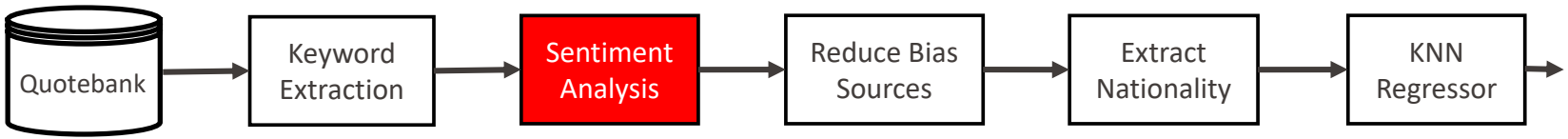


Sentiment Analysis

- Determining text polarity to identify positive to negative attitudes in text
- Using a pre-trained model for sentiment analysis of individual quotes, assigning a sentiment score from -1 to 1

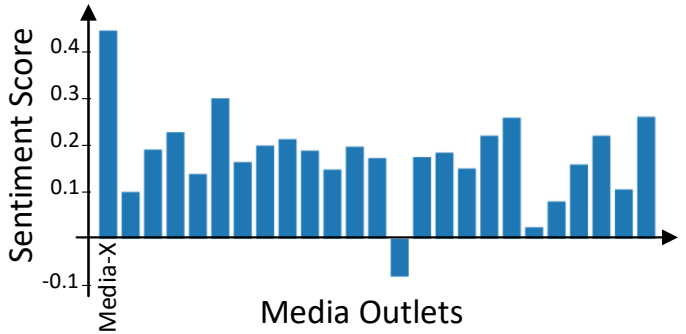
Example

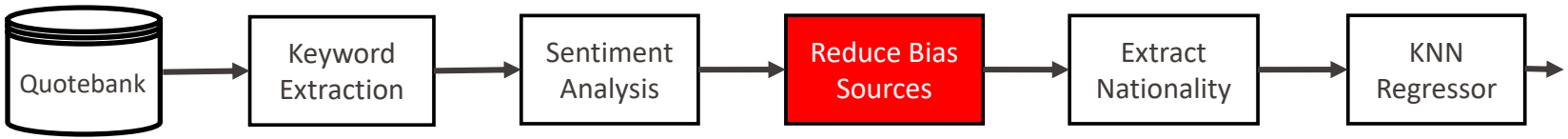
"We love the United States and you love the French, although you're sometimes too shy to say so"
François Hollande,
12 February 2014
Sentiment: 0.99



Reduce Bias Sources

- Ensuring consistency and reliability
- Enhancing the quality and comparability of data by eliminating significantly divergent media sources (p-value < 0.05)
- Based on the sentiment score distribution



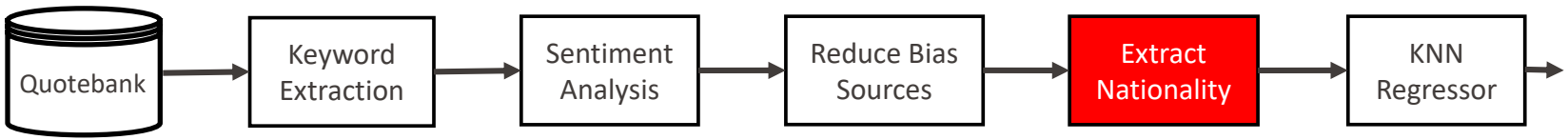


Extract Nationality

- Utilizing a Wikipedia database to establish the nationalities of quoted speakers.
- Discarding quotes with unidentifiable speakers and nationalities

Example

"For people who are not from the United States, the situation in the United States is not easy"
Unknown,
22 October 2018
Sentiment: -0.05

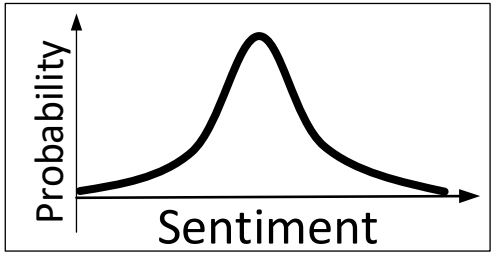


Sentiment Distribution

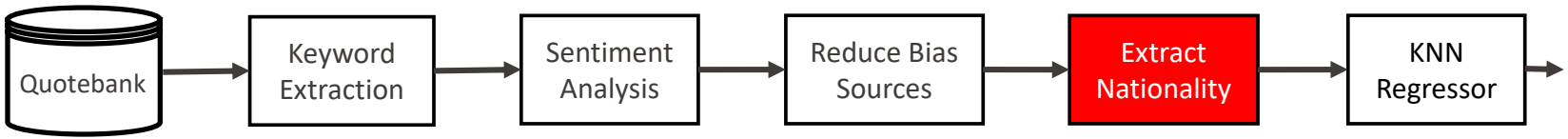
Quote	Country	Year	Sentiment
“ … ”	Brazil	2016	+0.1
“ … ”	Brazil	2016	+0.4
“ … ”	Germany	2017	-0.25
“ … ”	France	2017	-0.2
“ … ”	France	2018	+0.1

Group by Country and Year

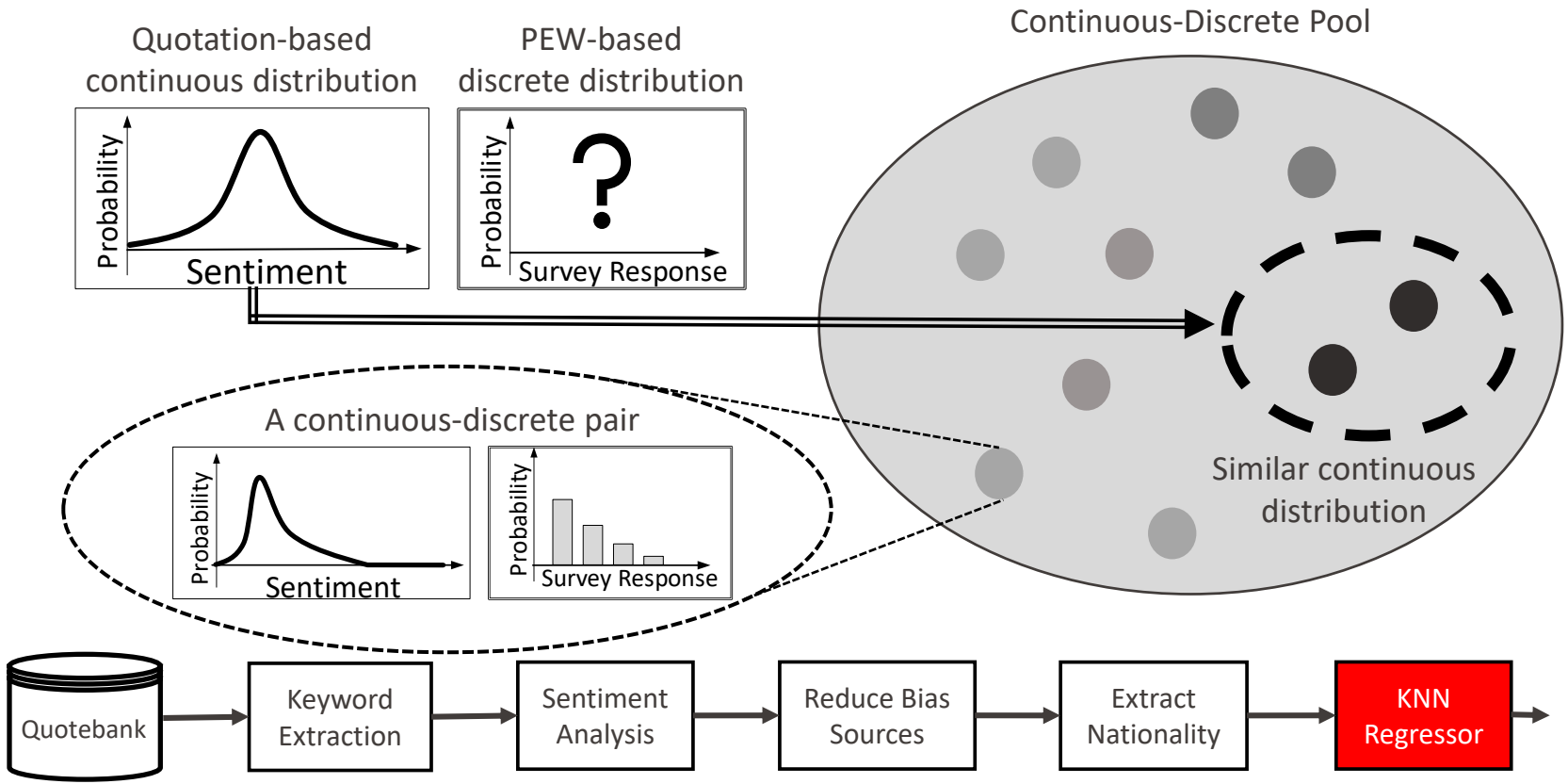
Distribution for Brazil 2016



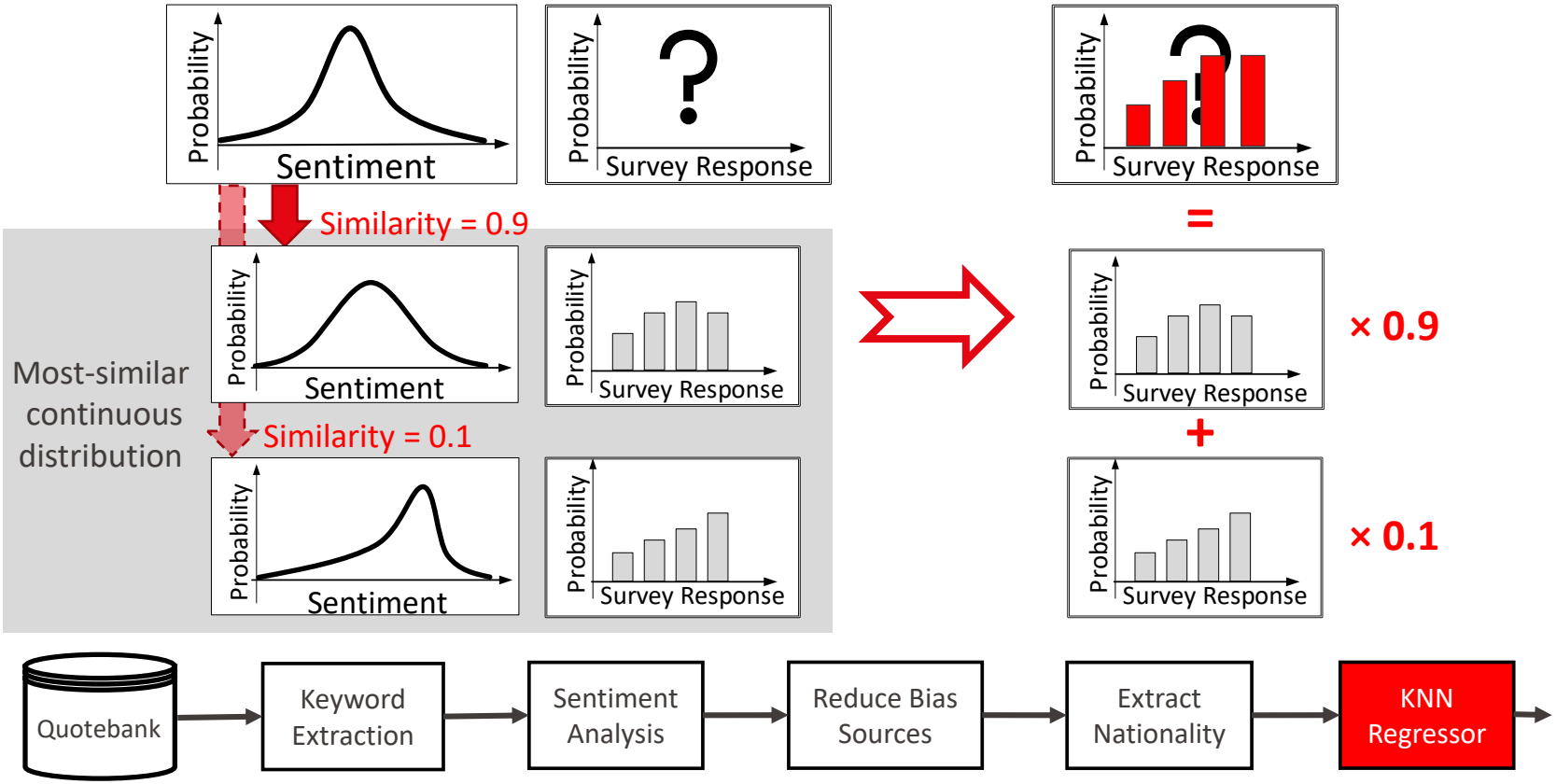
Target: Discrete distribution, similar to PEW data



Continuous-Discrete Transfer



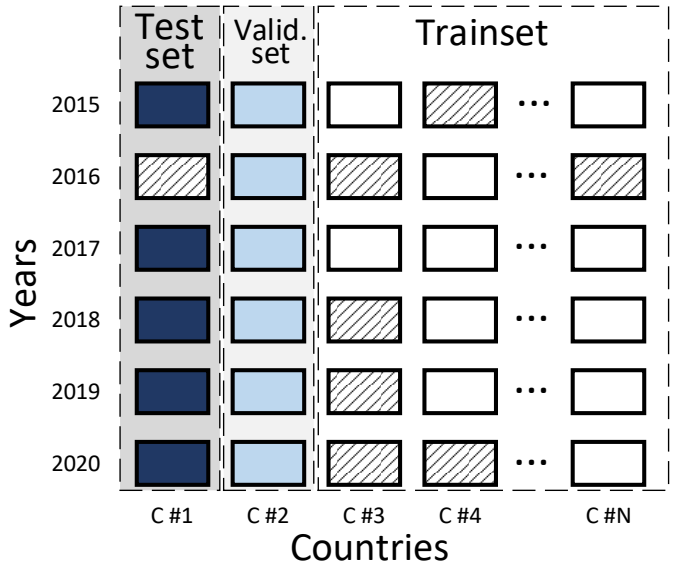
Continuous-Discrete Transfer, KNN Regressor



Real-world Scenarios

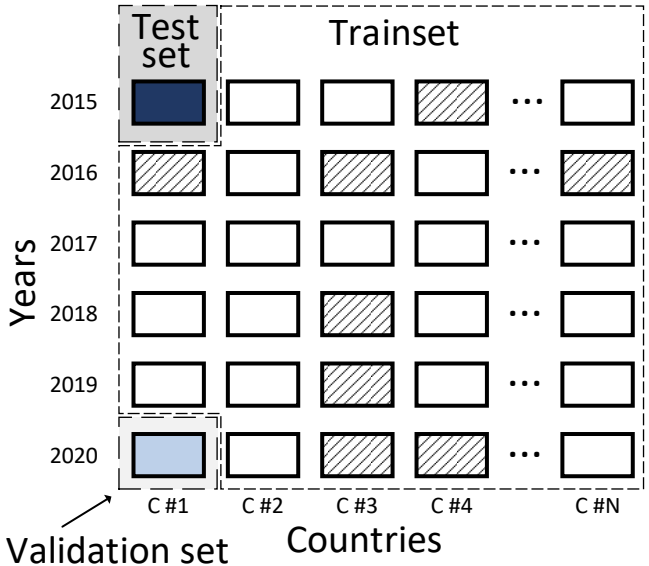
1. Leave-One-Country-Out (LOCO)

There is NO data for the target country in PEW



2. Same-Country-Validation (SCV)

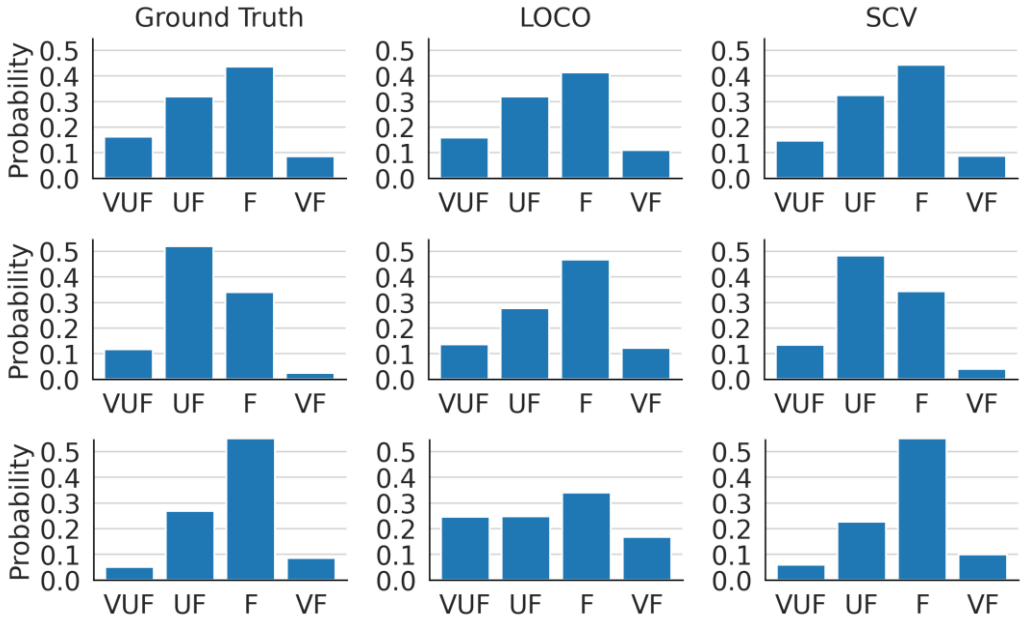
There is data for the target country in PEW BUT in specific years



Samples of Survey Prediction Results

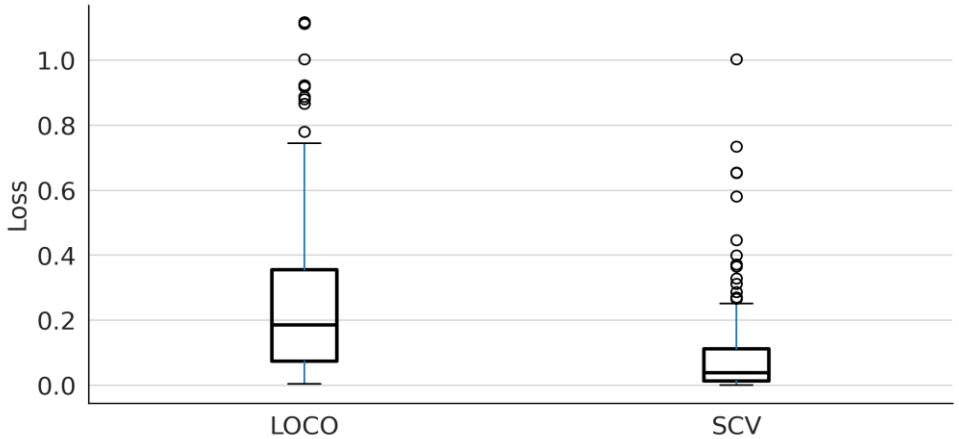
1. Australia, 2019
2. Germany, 2017
3. Hungary, 2018

VUF: Very unfavorable
UF: Somewhat unfavorable
F: Somewhat favorable
VF: Very favorable



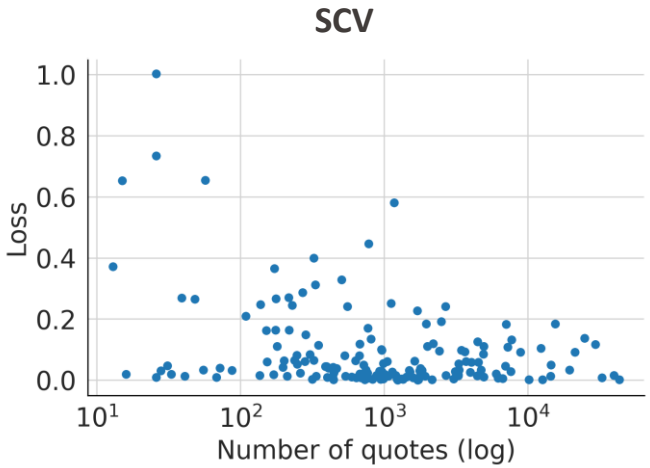
Overall Survey Prediction Results

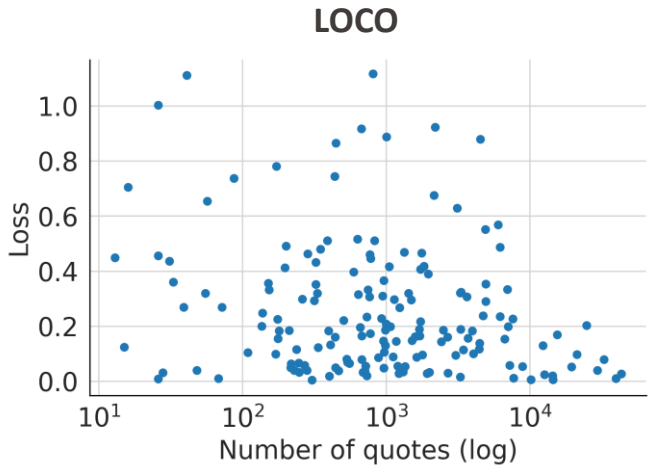
- $\mathcal{L} = \mathrm{KL}(\mathrm{PEW}(c, y) \parallel Predict(c, y); K = k)$
- (c, y) represents a country and year in the test set
- k is the number of neighbors in KNN, obtained by the validation set
- $0 \leq \mathcal{L} < \infty$
- The lower the loss, the better the performance
- SCV outperforms LOCO



Correlation with Number of Quotations

- Pearson correlation coefficient with p-value < 0.05
- There is a significant correlation between the number of quotes and the loss
- More quotations in a particular country and a specific year can lead to lower loss and improved prediction accuracy

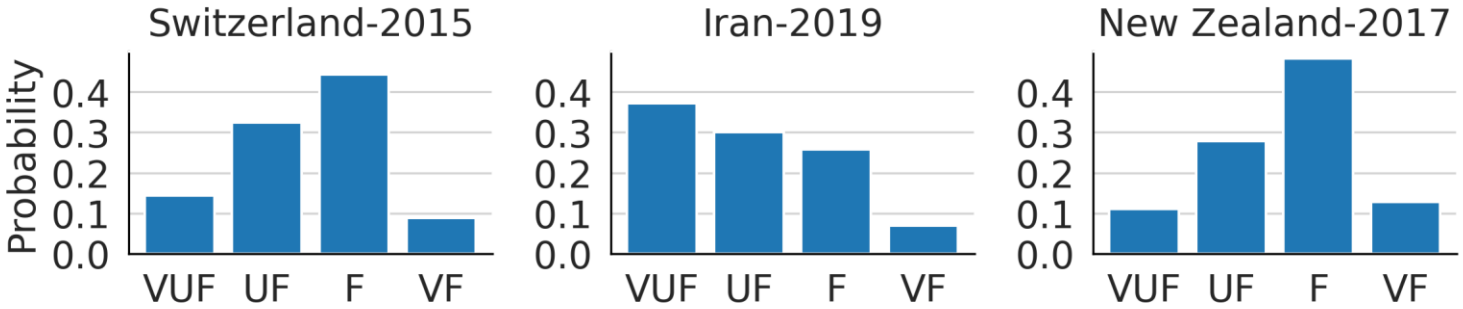




The Time Tunnel

- Countries not included in the PEW research
- => LOCO Scenario
- The potential of our quotation-based model to be used for predicting survey response in countries that are not covered in existing surveys

VUF: Very unfavorable
UF: Somewhat unfavorable
F: Somewhat favorable
VF: Very favorable



Conclusion

- A novel approach for predicting survey response
- By analyzing quotations using machine learning techniques
- U.S. favorability as a case study without loss of generality

- Leave-one-country-out (LOCO) for the countries NOT included in the existing survey data
- Same-country-validation (SCV) for the countries included in the existing survey data BUT NOT for all the years

GitHub Link

Thank you for your attention!