Datasheet for 'Perceptions of Democracy in South Korea'*

Hailey Jang

2024-04-18

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

- The dataset was created to explore the impact of globalization on democratic values in South Korea. Specifically, it aimed to understand how economic and political dissatisfaction influence the democratic perceptions of South Korean citizens. The research utilized a comprehensive questionnaire distributed in 2016, focusing on citizens' perceptions of economic and political aspects. The purpose was to address the contemporary democratic ethos in South Korea using a linear regression model to analyze the relationships between citizens' economic conditions and political contexts.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

- The dataset was compiled under the auspices of The Comparative Study of Electoral Systems (CSES), a collaborative research project among national election studies worldwide. It utilized responses from the "Voter Political Consciousness Survey" conducted following the 20th presidential election in 2016 in South Korea.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

- The document does not specifically mention who funded the creation of the dataset. However, since the dataset was used under the Comparative Study of Electoral Systems (CSES), it is likely that it was supported through the framework of this international research project.

*Code and data are available at: https://github.com/Hailey-Jang/Democracy_Perception_in_South_Korea.git.

4. *Any other comments?*

- The study emphasizes the significance of addressing economic and political dissatisfaction to foster a more inclusive and participative democratic environment in South Korea. It highlights the importance of such research in contributing to a broader understanding of how globalization shapes political systems worldwide.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The dataset consists of survey responses from individual participants. The instances represent individual entries in the survey, primarily capturing political and socio-economic perceptions related to elections in South Korea.

2. *How many instances are there in total (of each type, if appropriate)?*

- The specific total number of instances (individual survey responses) in the dataset is 119.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The dataset is a sample of responses from the broader population of voters in South Korea, derived from the "Voter Political Consciousness Survey." It aims to be representative of the population's political consciousness among 119 citiznes.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

- Each instance contains a large number of features (605 columns), which include both coded survey responses and some metadata (e.g., survey version, DOI). Data types include integers, floating-point numbers, and strings, suggesting a mixture of quantitative answers, categorical data, and identifiers.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

- The dataset includes labels associated with each instance, primarily the levels of satisfaction or dissatisfaction with democracy and the economy, perceptions of corruption, and political performance. These are derived from survey responses and coded into numerical scales

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- While the initial preview doesn't explicitly show missing values, some entries like '9999999' or '9999996' in certain columns may represent coded missing data or specific survey coding for non-responses or non-applicable answers.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - There is no specific relationships between individual instances

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - There is no information provided about recommended splits for training, validation, or testing within this dataset. Usage would depend on the specific research design.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Given the complex nature of survey data, errors or redundancies might exist, but cannot be definitively identified without further detailed analysis or access to the survey design documentation.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset contains metadata linking to external standards or documentation (e.g., DOI links), suggesting reliance on external definitions or protocols for data collection and coding. No information about the permanence or changes over time for these resources is available from the dataset alone.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The dataset appears to handle personal perceptions and opinions which are generally anonymized; however, direct information about confidentiality protections isn't available in the dataset preview.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - There is no indication from the preview that the dataset contains offensive or insulting data. It focuses on political and economic perceptions.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset likely categorizes respondents by demographic factors such as age or gender, which are common in survey datasets. Exact distributions would require further analysis.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Without more detailed metadata, it is unclear if individuals could be indirectly identified; the dataset appears to be anonymized for general research use.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset contains politically sensitive information, such as voting behavior and political opinions, which could be considered sensitive. Other sensitive demographic information might also be included.

16. *Any other comments?*

    - To fully leverage this dataset, detailed documentation on the survey methodology, question coding, and data cleaning procedures would be necessary for accurate analysis and interpretation.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was reported by subjects through a survey. Such data is directly reported by the participants and typically involves questions related to political, economic, and social perceptions. Validation or verification of this kind of self-reported data is often limited to consistency checks and logical validation (e.g., ensuring responses fall within the allowed range).

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Data collection for surveys is typically done using paper forms where respondents enter their answers. The validation of these mechanisms usually involves pre-testing the survey (pilot testing) to ensure questions are understood as intended and the data collection system works smoothly.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset likely employed a probabilistic sampling method to ensure representativeness of the South Korean electorate, but specifics such as the exact method (stratified, random, etc.) require access to the survey methodology documentation.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Such surveys are generally conducted by research teams or through survey firms. Compensation details for these personnel would depend on the employment or contract terms but are typically salaried positions or contracted services.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected following the 20th presidential election in 2016 in South Korea. The creation timeframe matches the survey's focus, aiming to capture contemporary political consciousness around that period.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - For studies involving human subjects, especially on sensitive topics like political opinions, ethical review approval (e.g., from an Institutional Review Board) is commonly required. This process would ensure that the study complies with ethical standards, though specific details would be documented in the associated study protocols.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected directly from individuals through the survey process, not obtained via third parties or secondary sources.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Participants in such surveys are typically notified about the study's purpose at the time of survey administration. This notification would include information about the survey's scope, the organization conducting it, and how the data will be used.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Consent to participate in the survey would have been obtained at the time of data collection, usually through a consent form or verbal agreement after explaining the survey's purpose and use of the data. The exact language of consent would typically emphasize voluntary participation and the confidentiality of responses.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - If provided, mechanisms to revoke consent would be detailed in the consent form, allowing participants to withdraw from the study at any time. This might be more relevant in longitudinal studies than in one-time surveys.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- An analysis like a data protection impact analysis is crucial for sensitive data but is not often detailed in the dataset itself. Such analysis would assess risks related to data handling and privacy implications.

12. *Any other comments?*

    - For a comprehensive understanding and responsible use of the dataset, access to the full survey documentation, ethical approvals, and data handling protocols would be necessary. These documents would provide deeper insights into the dataset's integrity and appropriateness for specific types of analysis.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - While the specific document doesn't detail the usage of this dataset, datasets similar to this one are typically used in political science and sociology research to analyze voter behavior, perceptions of democracy, and the influence of political events on public opinion. They might also be utilized in studies assessing the impact of economic conditions on political attitudes or the effectiveness of governmental policies.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - https://github.com/Hailey-Jang/Democracy_Perception_in_South_Korea.git

3. *What (other) tasks could the dataset be used for?*

    - Beyond political analysis, this dataset could serve educational purposes in teaching statistical analysis, political science, or data science methodologies. It could also be used in predictive modeling to forecast election results or public opinion trends based on demographic and historical data. Additionally, NGOs or policymakers might analyze the data to better understand public needs or responses to governance.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - Users of the dataset should be aware of potential biases introduced during the data collection or preprocessing stages. For instance, non-response bias or sampling errors can skew results, potentially leading to misrepresentations of certain populations. To mitigate these risks, users should perform rigorous statistical tests

to check for representativeness and bias and apply corrective measures like weighting or stratification where applicable. They should also be transparent about any limitations of the data when publishing findings to avoid misleading interpretations.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - The dataset should not be used for any purposes that could potentially harm individuals, such as targeting vulnerable groups based on political opinions or creating discriminatory policies. It's also inappropriate to use this data for commercial purposes that could infringe on the privacy or rights of the individuals who participated in the survey, especially without explicit consent covering those uses.

6. *Any other comments?*

   - Future researchers and users of the dataset should ensure compliance with ethical guidelines and legal standards, particularly regarding data privacy and protection. Regular reviews and updates to the dataset documentation can help maintain its relevance and usefulness, addressing any emerging ethical or legal issues. Collaboration with domain experts in political science can enhance the reliability and applicability of research findings derived from the dataset.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - The supporting entity would usually be the academic institution or research group that conducted the survey. Their ongoing commitment to maintaining the dataset would depend on funding and research priorities.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - Typically provided through the dataset's main documentation or via the research group's webpage.You can contact via `https://cses.org/`

3. *Is there an erratum? If so, please provide a link or other access point.* -Any updates would likely be communicated through academic publications or updates to the repository where the dataset is hosted. The process would involve the original research team or a designated data curator.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - Limits on data retention would be especially important for sensitive data, and would typically be governed by ethical guidelines or legal requirements related to data protection (e.g., GDPR). Documentation should detail how these aspects are managed.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - Mechanisms for external contributions would be rare for such datasets, but if allowed, would likely require thorough validation to maintain data integrity and reliability.

1 References