Datasheet(enhancement)

George Zhiqi Chen

2022/4/29

1. Motivation:

- For what purpose was the dataset created:

These datasets are constructed to understand the demographic structure of the Canadian Armed Forces. The six datasets describe the number of active-duty Canadians who serve in the Canadian military each year, and the number of troops attrition each year. Numbers of English-speaking and French-speaking groups in the military, racial minorities in the military, gender, rates of disability, and levels of education in the military and other important facts. Each data set has its own purpose, such as the CAF member data to understand the trend of the military active duty in the last ten years. The purpose of the attrition data is to understand the trend of active-duty attrition in the past decade. Data on the number of English and French groups in the armed forces explore the comparison of different language groups in the Canadian Armed Forces. The proportion of social groups in the CAF data discusses the contribution of minority groups in the armed forces, while the military education level data explores trends in educational and technical requirements within the development of armed forces.

- Who created the dataset and on behalf of which entity:

These data set is provided by the Canadian Armed Forces - Chief of Military Personnel (CMP) annual report. Military Personnel Command (MPC) supports the requirement to release accurate and timely information to Canadians, in line with the principles of Open Government.

- Who funded the creation of the dataset:

The Government of Canada and the Ministry of National Defence funded the report and creation of the dataset.

2. Composition:

- What do the instances that comprise the dataset represent? Are there multiple types of instances?
  The instances that comprise the dataset represent the regular service members of the Canadian Armed Forces, including officers and non-commissioned member. The only type of instance is people.

- How many instances are there in total: Given that the numbers vary from year to year .
- 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?

These data sets contain all possible instances and all registered regular personnel are in the data sets.

- What data does each instance consist of ("Raw" data or features):

Each instance is just a number, but each data contains military rank.

- Is there a label or target associated with each instance:

Not applicable.

- Is any information missing from individual instances:

Not applicable.

- Are relationships between individual instances made explicit:

Not applicable

- Are there recommended data splits (e.g., training, development/validation, testing):

There is no recommended data split.

- Are there any errors, sources of noise, or redundancies in the dataset:

There are some uncertainties in the military education level data. For example, Officer and NCM's data from 1998,2008, and 2018 are mostly outlier data. Maybe there was a statistical error.

- Is the dataset self-contained, or does it link to or otherwise rely on external resources:

The dataset can be downloaded from the Open Government website: https://search.open.canada.ca/en/od/?od-search-subjects=Military&_ga=2.80338195.501191627.1651213633-1505714419.1648749825

- Does the dataset contain data that might be considered confidential:

The dataset does not contain the confidential information. All data were collected from public sources.

- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

- Does the dataset identify any subpopulations? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Most of the data sets did not identify any subgroups, but in the CAF social group dataset they did primarily identify gender, ethnic minorities, aboriginal people, and people with disabilities.

- Is it possible to identify individuals, either directly or indirectly (i.e., in combination with other data) from the dataset:

It is impossible to identify individuals.

- Does the dataset contain data that might be considered sensitive in any way:

Thanks to government safeguards, the dataset does not contain sensitive information.

3. **Collection process**:

- How was the data associated with each instance acquired? Was the data directly observable, reported by subjects, or indirectly inferred/derived from other data?

The data associated with each instance were collected from the Chief of Military Personnel (CMP) Annual Report, which is based on Personnel registration information.

- What mechanisms or procedures were used to collect the data? How were these mechanisms or procedures validated?

Since the dataset contains the personnel registration information. Just Use Excel to record personnel registration information. There may exist measurement error, but how this data collection procedure was validated was unknown.

- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Data sets are not samples.

- Who was involved in the data collection process and how were they compensated?

All officially enrolled regular active-duty officers and non-commissioned members, as well as Chief of Military Personnel.

- Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances?

Each dataset may vary. Most of the dataset was collected during the 1997-2020 year. But CAF education level was collected during the 1997-2018 year. And because of its specialty the CAF social group dataset was updated recently so it was gathered during 2021.

- Were any ethical review processes conducted:

Whether there were ethical review processes conducting was unknown.

- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources:

Obtained from the official website, by downloading it from the website Open Government where government share the collected data. The Website link is: https://search.open.canada.ca/en/od/?od-search-subjects=Military&_ga=2.80338195.501191627.1651213633-1505714419.1648749825

- Were the individuals in question notified about the data collection: Unknown.

- Did the individuals in question consent to the collection and use of their data: Unknown.

- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses:

Unknown.

- Has an analysis of the potential impact of the dataset and its use on data subjects been conducted: Unknown.

4. **Preprocessing/cleaning/labeling:**

• Was any preprocessing/cleaning/labeling of the data done:

Since most data in the dataset is categorical data and a few missing values exist in the dataset, there was no preprocessing/cleaning/labeling of the data done. But I cleaned the data by removing the missing values and for better explanation I kept outlier.

5. **Uses:**

- Has the dataset been used for any tasks already:

These data sets are widely referenced and used, often appearing in papers and military policy studies

- 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.

The paper that uses the dataset can be accessed here in pdf:
http://www.forces.gc.ca/assets/FORCES_Internet/docs/en/about/code-eng.pdf e on SAGE:

- What (other) tasks could the dataset be used for?

The dataset and the report were also used by student cadets and military media.

- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Is there anything a future user could do to mitigate these undesirable harms?

No.

- Are there tasks for which the dataset should not be used? Unknown.

6. **Distribution**:

- Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?

The dataset has been shared on the open government website that is a public access research data sharing repository.

- How will the dataset will be distributed? Does the dataset have a digital object identifier (DOI)? The dataset has been uploaded in the csv and xlsx file directly on the open government website. The dataset has a DOI: OD-2019-00001(Regular force Total number), OD-2019-00004(Attrition), OD-2019-00006(English-French speaker), OD-2018-00012(Education level), OD-2020-00005(social group).

- When will the dataset be distributed?

The Social group dataset was distributed on 2020-03-18. The Regular force Total number dataset was distributed on 2019-09-02. The Attrition dataset was distributed on 2019-12-01. The English-French speaker dataset was distributed on 2019-12-19. The Education level) dataset was distributed on 2018-10-21.

- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is distributed under Open Government Licence - Canada, the link is available here: https://open.canada.ca/en/open-government-licence-canada .

- Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Unknown

- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown.

7. **Maintenance**:

- Who is supporting/hosting/maintaining the dataset?

The original organization and stuff will keep supporting/hosting/maintaining and updating the dataset.

- How can the owner/curator/manager of the dataset be contacted:

Contacted by open government official email: open-ouvert@tbs-sct.gc.ca.

- Is there an erratum? If so, please provide a link or other access point. Unknown.

- Will the dataset be updated? If so, please describe how often, by whom, and how updates will be communicated to users?

Maintenance and Update Frequency: As Needed.

- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances? If so, please describe these limits and explain how they will be enforced. Unknown.

- Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Most of the old data will become part of the new data, but it may also be erased.

- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Unknown.