Introduction to NLP
Chapter 1:
Dataset Collection and Building

Muhammad Okky Ibrohim, S.Mat., M.Kom.

Credit

This slide was built using **Blue Connections Cordelia Presentation Template** theme (https://www.slidescarnival.com/cordelia-free-presentation-template/216).

Several images and other media from Google Images that maybe I forgot to add specific credit (if there is belong yours, please contact me if you want me to add specific credit to you or remove it).

Video course

For those who can not attend the synchronous session or want to relearn this material, you can watch this video for this slide presentation: https://youtu.be/baK17aj8kvc

Outline

1.1 Why we need to learn dataset collection and building?

1.2 How to collect dataset

1.3 Data annotation and how to validate it

1.1 Why we need to learn dataset collection and building? Definition

- Data collection

  Data collection is the process to collect a dataset that is relevant to the research/experiment focus, whether collecting a raw dataset (e.g. scraping) or collecting a ready-to-use dataset (e.g. already annotated) for the experiment (e.g. downloading open dataset from Kaggle).

- Data building

  Data building is a process to build (e.g. annotate) the dataset and validate it dataset by ourself.

- Other definition

  Some researchers define data collection and building as an inseparable unit and simply called it as Data Collection.

1.1 Why we need to learn dataset collection and building? Why it is important

- Data not yet available

  Not all data that we need is already available in particular open dataset website so that we need to collect and build it by ourself.

- Garbage in, garbage out

  Not all open datasets are validated, having a good dataset will make the building model process easier.

- Data need to be improved

  Sometimes we find a dataset that appropriate for our research but the dataset is too small for our experiment so we need to add it.

1.2 How to collect dataset Source

- Get from an open dataset website

  Some website like UCI (https://archive.ics.uci.edu/ml/datasets.php) and Kaggle (https://www.kaggle.com/datasets) provides a thousand dataset that may fit to our research/study.

- Get from a research paper

  Several researchers open their dataset to the public by giving a particular link (GitHub, Google Drive, or a specific website).

- Ask directly to the author

  Several researchers open their dataset to the public with a particular restriction (e.g. restricted for research purposes only) so that we need to send dataset requests to the author.

- Scraping

  If the dataset we need is not available in an open dataset, just scrape it (whether through API or direct scraping). Python with Selenium or other libraries can help us with this.

1.2 How to collect dataset
Attention in Collecting Dataset

- Read the license carefully

  Some open datasets are restricted for research purposes only which makes us can not to use them for commercial purposes.

- Ask the dataset restriction to the author

  If we want to use an open dataset for commercial purposes but the author does not put a license or explain the dataset restrictions, better for us to ask directly to the author first.

- Read the data description carefully

  A good open dataset has a clear data description to explain how they build and validate their dataset (including providing the evaluation metric they used). Read it carefully to make sure the open dataset is fit for our research/experiment.

1.3 Data annotation and how to validate it

Data Annotation Step by Step

1. Data collection
2. Annotation guideline and ground truth examples
3. Annotation guide readability test
4. Annotator recruitment and pilot study
5. Data annotation process
6. Final label and data annotation validation

1.3 Data annotation and how to validate it

Data Annotation Step by Step:

1. Data collection

- Collecting data from scratch

  If we do not find any open dataset that is similar to our task, we need to scrap it from scratch.

- Reannotate an open dataset

  If we find an open dataset that is similar to our task, we can use and reannotate it to fit our task.

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
2. Annotation guideline ground truth examples (1)

- Arrange an annotation guideline

  We need to write our annotation guidelines as clear as possible to make the annotator easy to understand the task so that we can get a valid annotation result. Do a literature review to get a valid definition and example for our annotation task.

- Consult with the expert

  Sometimes we must consult with an expert to get a valid fundamental definition and example for our annotation guideline (e.g. annotation guideline for hate speech detection task).

1.3 Data annotation and how to validate it

Data Annotation Step by Step:
2. Annotation guideline ground truth examples (2)

- What we need to put in our annotation guideline
  - Task definition
  - Motivation
  - Annotation example with the reason explanation
  - How to use the annotation tool
  - Contact person

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
2. Annotation guideline ground truth examples (3)

- Attention in arranging an annotation guideline
  - Do not ask annotator to annotate too many label at the same time (use multi-stage annotation instead)
  - If we do not have to, as much as possible use "**click**" for annotation instead of "**manual typing**“.

1.3 Data annotation and how to validate it

Data Annotation Step by Step:

2. Annotation guideline ground truth examples (4)

- Building ground truth examples
  - We need a set of ground truth examples for our pilot study (will be explained more later).
  - It ranges between 100-200 data or as needed.

1.3 Data annotation and how to validate it

Data Annotation Step by Step:
3. Annotation guide readability test

- Definition

  A process to check whether our annotation guideline is easy to understand or not and do not contain a typo before we send it to our real annotator.

- With whom we did it
  - Ask research partners that do not write the guideline
  - Consult with an expert

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
4. Annotator recruitment and pilot study (1)

- Annotation by expert vs paid crowdsourcing vs voluntary crowdsourcing
  - By expert:

    We can get the most valid annotation result, but this scheme is sometimes very expensive, and difficult to find an expert that has spare time to annotate a data.
  - By paid crowdsourcing:

    Cheaper than using an expert, but sometimes we are unlucky when getting inconsistent annotator.
  - By voluntary crowdsourcing

    We can annotate our dataset for free by asking for help from our friends or asking a community/organization that has the same consent/problem as us, but we may face difficulties to find those kinds of people.

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
4. Annotator recruitment and pilot study (2)

- Attention in choosing/recruiting annotators
  - Annotating non-subjective task

    If we want to annotate a non-subjective task (e.g. spam classification), 1 or 2 annotators per data is enough (no need for an expert).
  - Annotating subjective task

    If we want to annotate a subjective task, especially for a high subjective task (e.g. hate speech detection, political sentiment/stance classification), we may need 1 or 2 expert annotators per data or a minimum of 3 crowdsourcing annotators from various backgrounds (last education level, religion, ethnic, etc.) to reduce the annotation bias.

1.3 Data annotation and how to validate it

Data Annotation Step by Step:
4. Annotator recruitment and pilot study (3)

- Do a pilot study (annotator testing)
  - A pilot study (annotator testing) is done by asking the annotators candidate to annotate the ground truth examples.
  - This is done in order to test whether the annotators read and understand the annotator guideline or not.
  - We can validate the annotators' understanding by counting the Cohen's Kappa (will be explained more later) between annotation results with the ground truth example for each annotator separately.
  - If an annotator has Cohen's Kappa below the threshold, do an investigation to check whether the annotator does not read carefully/understand the guideline or the guide difficult to understand for the majority of annotators by comparing the Cohen's Kappa between annotators.

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
5. Data annotation process

- Split the dataset

  We can split the dataset that will be annotated by annotators to manage bad things (e.g. annotator suddenly declared unable to complete, annotator disappeared without news, etc.) easier.

- Monitoring the annotation progress

  If possible, do the annotation process using an online tool (using Google Sheets or specific online annotation tools) to monitor the progress in real time. If it is offline, keep in touch with the annotators periodically. We must manage the annotators to do the task gradually (not done at once before the deadline).

1.3 Data annotation and how to validate it

Data Annotation Step by Step:

6. Final label and data annotation validation (1)

- The most used techniques in deciding the final label:
  - *Majority Voting*

    Decide the final label by most votes. If there are not most votes in a data, we can add annotators or simply remove the data from the dataset.
  - *100% Agreement*

    Decide the final label if only if all annotators vote the same label. If not, the data is removed from the dataset.

- Do not remove the annotation details:

  In several advanced NLP techniques, annotator disagreement can be used as a feature.

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
6. Final label and data annotation validation (2)

- Annotation validation metric: Cohen's Kappa

  We can use this metric only if we use two same annotators for the whole dataset we annotated.

Data	Annotator 1	Annotator 2
Data 1	0	1
Data 2	1	0
Data 3	0	0
Data 4	1	1
Data 5	1	1

- How to do it:

  Just give it to a Python library to solve it :)

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
6. Final label and data annotation validation (3)

- Annotation validation metric: Fleiss' Kappa

  If we use more than two annotators for annotating the whole dataset, we can use Fleiss' Kappa instead of Cohen's Kappa.

Data	Annotator 1	Annotator 2	Annotator 3
Data 1	0	1	0
Data 2	1	0	1
Data 3	0	0	0
Data 4	1	1	0
Data 5	1	1	1

- How to do it:

  Just give it to a Python library to solve it :)

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
6. Final label and data annotation validation (4)

- Annotation validation metric: Krippendorff's Alpha

  Different from the two previous metrics that only can be used for nominal labels and can not handle missing values, this metric can handle ordinal and ratio labels and missing values.

Data	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5
Data 1	0	1	0		
Data 2		0		0	1
Data 3	0	0	0	0	0
Data 4	1	1	0		0
Data 5			1	1	1

Data	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5
Data 1	0	1		1	
Data 2		6	5	7	1
Data 3	0	0	0		
Data 4	1	2	0	3	0
Data 5		3	2	4	3

- How to do it:

  Just give it to a Python library to solve it :)

1.3 Data annotation and how to validate it
Data Annotation Step by Step:
6. Final label and data annotation validation (5)

- Annotation validation metric: Kappa and Krippendorff's Alpha Interpretation

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

- Source: https://fammedarchives.blob.core.windows.net/imagesandpdfs/fmhub/fm2005/May/Anthony360.pdf

(Note: some references may have different ranges for the interpretation)

Thank You!

Bisa.ai
PT. Bisa Artifisial Indonesia

Kampus Merdeka
INDONESIA JAYA