Here we consider dataset(s) produced by the data preparation phase, used for modeling or for the major analysis work of the project.

1. SELECT DATA

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

1.1. Rationale for inclusion/exclusion

List the data to be used/excluded and the reasons for these decisions.

• Collect appropriate additional data (from different sources—in-house as well as externally) • Perform significance and correlation tests to decide if fields should be included • Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data quality and data exploration (i.e., may wish include/exclude other sets of data)
• Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experience of modeling (i.e., model assessment may show that other datasets are needed) • Select different data subsets (e.g., different attributes, only data which meet certain conditions) • Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.)
• Document the rationale for inclusion/exclusion
• Check available techniques for sampling data

We selected 3000, randomly sampled context, question, and answer triplets from the TORQUE dataset. The selected dataset was divided equally among 3 different categories from the overall dataset, noted as categories 0, 1, 2. Each category corresponds to “What events have already finished?”, "What events have begun but not finished?” and “What happened after/before an event?” type questions respectfully.

The decision to sample only 3,000 questions from the larger dataset of 25,000 was driven by considerations of our computational resources. We determined that 3,000 questions offer a sufficiently diverse range of data for model training, while also ensuring that the dataset size remains manageable for processing within our computational constraints. This balance allows for efficient model training without overwhelming our resources.

We decided to include some questions without answers so we could supply our models with negative examples during training, as in the real world it is possible to encounter questions that cannot be answered.

Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle the weighting.

2. CLEAN DATA

Raise the data quality to the level required by the selected analysis techniques. This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.
Describe the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task. If the data are to be used in the data mining exercise, the report should address outstanding data quality issues and what possible effect this could have on the results.

- Reconsider how to deal with any observed type of noise
- Correct, remove, or ignore noise
- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking results of a survey where some questions were not asked or not answered. This might result in a value of 99 for unknown data. For example, 99 for marital status or political affiliation. Special values could also arise when data is truncated—e.g., 00 for 100-year-old people or all cars with 100,000 km on the odometer.
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data).

No specific cleaning had to be done, the entire data extraction process is described in section 1.1.

3. CONSTRUCT DATA

This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.

- Check available construction mechanisms with the list of tools suggested for the project • Decide whether it is best to perform the construction inside the tool or outside (i.e., which is more efficient, exact, repeatable)
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data construction (i.e., you may wish include/exclude other sets of data)

To input our data into BERT we must use word embeddings. The HuggingFace Transformers library is an ideal choice for generating word embeddings for LLMs.

For non-answerable questions we will train Bert to predict a null token or “not answerable” string, which may involve augmenting the current data by replacing null answers with a string or token.

We do not plan to change the Context, Question, Answer format already present in Torque.

3.1. Derived attributes

Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. An example might be: area = length * width.
Why should we need to construct derived attributes during the course of a data mining investigation? It should not be thought that only data from databases or other sources should be used in constructing a model. Derived attributes might be constructed because:

- Background knowledge convinces us that some fact is important and ought to be represented although we have no attribute currently to represent it
- The modeling algorithm in use handles only certain types of data—for example we are using linear regression and we suspect that there are certain non-linearities that will be not be included in the model
- The outcome of the modeling phase suggests that certain facts are not being covered

Derived attributes

- Decide if any attribute should be normalized (e.g., when using a clustering algorithm with age and income, in certain currencies, the income will dominate)
- [Optional] Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization)
- How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction).]
- Add new attributes to the accessed data

We do not need to have any derived attributes for our data.

Before adding Derived Attributes, try to determine if and how they ease the model process or facilitate the modeling algorithm. Perhaps "income per person" is a better/easier attribute to use than "income per household." Do not derive attributes simply to reduce the number of input attributes. Another type of derived attribute is the single-attribute transformation, usually performed to fit the needs of the modeling tools.

Single-attribute transformations

- Specify necessary transformation steps in terms of available transformation facilities (for example, change a binning of a numeric attribute)
- Perform transformation steps

As previously mentioned, we plan to use word embeddings to ensure our dataset is compatible with BERT. This essentially turns each string into a vector that can be used with BERT to train and improve it.

Transformations may be necessary to change ranges to symbolic fields (e.g., ages to age ranges) or symbolic fields ("definitely yes," "yes," "don't know," "no") to numeric values. Modeling tools or algorithms often require them.

3.2. Generated records

Generated records are completely new records, which add new knowledge or represent new data that is not otherwise represented (e.g., having segmented the data, it may be useful to generate a record to represent the prototypical member of each segment for further processing).

- Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).

Not Applicable

4. INTEGRATE DATA

These are methods for combining information from multiple tables or other information sources to create new records or values.
Merging tables refers to joining together two or more tables that have different information about the same objects. At this stage, it may also be advisable to generate new records. It may also be recommended to generate aggregate values.
Aggregation refers to operations where new values are computed by summarizing information from multiple records and/or tables.

- Check if integration facilities are able to integrate the input sources as required • Integrate sources and store results
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data integration (i.e., you may wish to include/exclude other sets of data)

This data has no need for merging or aggregation. Instead, we would like to highlight BERT's current performance by running inference on some of our questions. Some sample inputs and outputs are listed below:

Input 1:

Context: The coalition won in 2002 on a wave of euphoria after 24 years of rule by Daniel Arap Moi, but now is in a precarious position because of growing public dissatisfaction. Both camps agree that the existing constitution is outdated and oppressive but have failed to reach consensus on the new one, which has been the subject of debate since 1997.

Question: What event has already finished?

Output 1:

BERT Response: 24 years of rule

Expected keywords: won, wave, rule

Input 2:

Context: The security cabinet took the decision on December 6, a day after a suicide bomber from the radical Palestinian movement Islamic Jihad killed five Israelis at the entrance to a shopping mall. Under the deal, Israel should also allow convoys of trucks to travel between Gaza and the West Bank from January 15.

Question: What is likely to happen after the decision was taken?

Output 2:

BERT Response: suicide bomber from the radical palestinian movement islamic jihad killed five israelis at the entrance to a shopping mall

Expected keywords: allow, travel

These examples show the limitations of BERT inference prior to fine tuning. It seems as if BERT is relying on phrases that come after the keyword "after" in the context, or assuming similar results based on the structure of the sentence, but further testing will be needed to confirm any hypotheses.

5. FORMAT DATA

Formatting transformations refers primarily to syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.
Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Rearranging attributes

- Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

No such attribute arrangement is required as the tools planned for use take the input and outcome fields as separate parameters.

Reordering records

- It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute.

No records reordering is required. Instead, it is recommended to shuffle the data randomly.

Reformatted within-value

• These are purely syntactic changes made to satisfy the requirements of the specific modeling tool • Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data)

All the attribute values are numeric in the aggregated unified table except for the "Context," "Question," "Answer," and "Category" fields, which will need to be encoded to multidimensional numeric attributes if the modeling tool requires numeric input.

6. DATASET DESCRIPTION

Provide a general description of the final dataset (for instance, in terms of number of samples and number of features).

The final dataset consists of 3,000 samples and 4 features. This provides a comprehensive overview of the data structure, indicating a relatively moderate dataset size with a manageable number of features for analysis or modeling purposes.

There are 748 examples of non-answerable questions. We may decide to replace some of these examples with answerable questions or add more answerable questions to the dataset if we find that training could be improved.