Datasheet for candidates_labeled.pkl and candidates_unlabeled.pkl

NICO SCHMIDT, MARVIN MÜLLER, KLAUS SCHMIDT

1. MOTIVATION

A. For what purpose was the dataset created?

We created a dataset to train a model for the classification of statements about the future.

2. COMPOSITION

A. What do the instances that comprise the dataset represent?

The dataset contains candidates which are single sentences that could possibly be statements about the future.

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

There is one instance consisting of labeled (candidates_labeled) and unlabeled (candidates_unlabeled) candidates.

C. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

It contains all possible instances and is not a sample of instances from a larger set.

D. What data does each instance consist of?

The candidates instance consists of processed text in the form of sentences, where the sentences include a match with a previously applied regex filter (with future relation) and were cleaned by applying more regex filters. The candidates_unlabeled-set additionally includes 5000 sentences that weren't pre-filtered by applying future regexes, but also cleaned.

E. Is there a label or target associated with each instance?

Every candidate sentence is associated with a label. A label can be 0 or 1 stating if the respective candidate is a statement about the future (0) or not (1).

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

The candidates_labeled-set is split before the active learning process into a training- and a test-set by a 60-40 ratio. This ratio ensures that both sets include enough examples of each class. Furthermore active learning approaches usually use a bigger test-set than other supervised learning methods.

G. Are there any errors, sources of noise, or redundancies in the dataset?

It was tried to remove most sources of noise in the dataset through regular expressions. However there still might be noise in the dataset, caused by the huge variance of possible typesetting errors and different character encodings. Duplicate candidates in the dataset were completely removed to reduce the risk of overfitting.

H. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained and does not rely on any external source to be used.

I. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

The provided dataset is purely extracted from WARC files and hence was part of the publicly accessible internet.

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

The dataset could potentially involve the mentioned types of harmful content. This is, because major parts of the data were possibly not regulated by any superordinate instance, such as website moderators or automatic content filters when it was created.

K. Does the dataset identify any subpopulations (e.g., by age, gender)?

It is very unlikely that any subpopulations could be identified via the mentioned characteristics as the dataset contains only single sentences mostly without context. Furthermore all sentences were shuffled.

L. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

It may be possible to identify individuals, if they mentioned personal information or internet usernames that can in turn be linked to them using external sources.

M. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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)?

The dataset could possibly include sensitive data. This is, because the data was not filtered by any keywords except temporal determinations. Any other statements about orientations, beliefs or opinions can therefore be part of the dataset.

3. COLLECTION PROCESS

A. How was the data associated with each instance acquired?

The data was directly extracted from WARC files and accordingly directly observable through the whole extraction process.

B. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

The data was collected with the use of the WARC-DL pipeline, provided by [1]. Further deterministic methods such as regular expressions, content length checking and language detection were used.

C. Over what timeframe was the data collected?

The data was collected from WARC files containing web pages over a period of approx. ten years.

D. Were any ethical review processes conducted (e.g., by an institutional review board)?

The extracted data was not reviewed by any ethical means.

E. Were the individuals in question notified about the data collection?

No, they were not notified.

F. Did the individuals in question consent to the collection and use of their data?

Since the Internet Archive's data collection access is granted for research purposes, no further request for consent was made.

4. PREPROCESSING/CLEANING/LABELING

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The dataset was pre-processed by several strategies to obtain only specific sentences. Next to filtering the data by English language and sentence detection, regular expressions were used to extract only possible candidates for sentences with statements about the future. Afterwards the sentences were cleaned from noise, such as leading special characters or remaining HTML snippets, not belonging to the actual sentences. The candidates_labeled-set was also manually annotated by setting the affiliated labels to 0 or 1.

B. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

No, but with the use of the WARC-DL pipeline a new "raw"-dataset can be created at any time.

C. Is the software used to preprocess/clean/label the instances available?

Yes, the scripts are available under the following links:

- https://github.com/ja25opir/nostradamus-project/tree/main/scripts/data_cleaning.py
- https://github.com/ja25opir/nostradamus-project/tree/main/scripts/data_preprocessing.py

5. USES

A. Has the dataset been used for any tasks already?

So far the dataset was only used to perform model training.

B. Is there a repository that links to any or all papers or systems that use the dataset?

The dataset was previously only used in this paper and the associated repository:
https://github.com/ja25opir/nostradamus-project

REFERENCES

1. N. Deckers, “WARC-DL,” https://github.com/webis-de/WARC-DL (2022).