DataSheet Created by Obeta Ifeanyichukwu Malachy

(Email: angeloobeta@hotmail.com), GitHub: angeloobeta, Tel: +2348103405047

INTRODUCTION

In order to understand this documentation for the description of the datasets/datasheet; there are some important key points to note when reading through the body. There questions and answers that describe the motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution and maintenance of the datasets/datasheet. The questions are in blue text while answers in purple. It should also be noted that the questions have a capital letter Q preceding them while the answers A. it should also be noted that word **datasets** is used interchangeably with **datasheets** in this documentation.

The datasets are based on the Igbo language which is one of the three major ethnic group in Nigeria a country located in the western part of Africa. The lgbos (sometimes pronounced as Ibo) inhabitants every land in the world; there isn't any part of the continent that you wouldn't find an Igbo man; weather its America (North and South America), Europe, Asia, or Australia. Igbos are know all over the world because they are very influential people. It's said that anywhere that you go in the world that you can't find an Igbo man, you need to flee from there.

DESCRIPTION

The datasets are from the popular Igbo folk tales which are stories passed down from generations to generation. The stories are one of the ancestral methods teaching the upcoming generations lessons based on morals, values, ethics, wisdom etc. the Igbo culture for a better up-bringing of the generations to come. The stories are most often told my grandfather or mother after supper/dinner when children gather together to learn from the wisdom of the grand fathers and mothers. The stories are mainly for entertainment sake while some are told to remind the listeners the importance of keeping up with good morals.

MOTIVATION

Q: 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.

A: The datasets (corpora) were created to enable research on NLP (Natural Language Processing) find patterns and inference which include utilizing the dataset in areas like including, syntax, sentiment polarity, morphology, phonology (and phonetics), and the lexicon; given a piece of Igbo text from the datasheet/dataset, Which will enable the ML algorithm to efficiently and effectively learn. It was created intentionally with those tasks in mind, focusing on folk tales as a place where morphology, phonology (and phonetics) and lexicon affect/sentiment are frequently expressed.

Q: Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A: The datasets weren't created but was collected by Obeta Ifeanyichukwu Malachy at University of Nigeria, Nsukka (UNN) on behalf of the AI4D-African Language Dataset Challenge.

Q: 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.

A: Funding was provided single handedly by Obeta Ifeanyichukwu Malachy who is the dataset collector.

COMPOSITION

--Q: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
--A: The instances are plain text of Igbo folk tales collected from various different sources including novels, hearings.

Q: 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 (e.g., 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 (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A: The dataset is a sample of instances. It is (presumably) intended to be a random sample of instances of folktales which is from different sources including Igbo story books.

Q: What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

A: Each instance are raw data which haven't been processed (the only processing that it might have possible passed through was when each text was weighed to check if they were correctly spelled and checking for errors (in terms of grammatical errors were found corrected, apart from that no filtering were done.

Q: Is there a label or target associated with each instance? If so, please provide a description. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A: Everything is included. No data is missing.

Q: Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

➔A: There is no recommended data split since according to the challenge (AI4D-African Language Dataset Challenge) the dataset should be representative and balanced and useful for downstream NLP tasks. It all depends on what the consumer want to achieve with the datasets.

Q: Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., 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 (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A: There are no errors, sources of noise or redundancies in the datasets. The datasets isn’t really self-contained because it links to different story books of Igbo folk tales by numerous authors, the stories are linked to the authors; any restriction what would be associated with them are in terms of copy write permission.

Q: 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)? If so, please provide a description.

A: The datasets doesn’t contain any confidential data that’s protected by legal privilege they are all tells.

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

A: None of the datasets contain any data in any form that might be offensive, insulting, threatening or might otherwise cause anxiety.

Q: 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? If so, please describe how.

A: No it isn’t possible because all the characters in the datasets are fictions.

Q: 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)? If so, please provide a description.

A: The datasets contains some of the ethnic origins of the Igbo (tribe) people of the Nigeria.

Any other comments? No for now.

COLLECTION PROCESS

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

A: The data was acquired through directly observation (raw text, except the labels were extracted by the process described below. The data was collected by reviewing different books that contain Igbo folk tales.

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

A: The mechanisms or procedures for the collection of the whole data was through hardware apparatus, manual curation and software. The dataset was scanned with the help of HP scanner; the text where extracted through the help of software (Google text recognition software) and was finally stored in a text file.

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

A: The datasets aren't samples of a larger sets.

Q: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A: Those involved in the collection of the datasets are colleagues at work; for now they haven't been compensated because they did the collection of the data through their own freewill. But if at the end of the submission process and there is any available fund; they would be called to be compensated as a sign of appreciation.

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

A: The datasets were collected within the interval of two months that the AI4D-African Language Dataset Challenge was announced.

Q: Were any ethical review processes conducted (e.g., 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.

A: No

Q: Does the dataset relate to people? If not, you may skip the remaining questions in this section.

A: The dataset relates to people which their origin can't be traced because all the names mention are all myths and legends in the Igbo folk tales which are fictional.

Q: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

A: The data was collected from one of the popular Igbo story books known by the name **Omalinze a collection of Igbo folk tales**.

Q: 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.

A: No as mentioned earlier the datasets were sourced from a set of Igbo folk tales.

Q: 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.

A: No (see previous question).

Q: 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).

A: N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., 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.

N/A.

Any other comments? No for now

PREPROCESSING/CLEANING/LABELING

Q: 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)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

A: Some preprocessing errors were caught during the collection stage and these errors were in terms of grammatical errors. The errors were corrected immediately there were detected during the review of the datasets.

Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

A: Yes. The dataset itself contains all the raw data.

Q: Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

A: Yes, the software that was used is Google-text-recognition (https://cloud.google.com/vision/docs/ocr)

Any other comments? No for now.

USES

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

A: No, this is the first instance.

Q: 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.

A: No (please see the previous question)

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

A: The dataset could be used for anything related to natural language processing NLP, modeling or understanding language reviews which include on finding patterns and inference which include utilizing the dataset in areas like including, syntax, sentiment polarity, morphology, phonology (and phonetics), and the lexicon; given a piece of Igbo text from the datasets, Which will enable the ML algorithm to efficiently and effectively learn. It was created intentionally with those tasks in mind, focusing on folk tales as a place where morphology, phonology (and phonetics) and lexicon affect/sentiment are frequently expressed..

Q: 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 future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

A: None.

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

A: None.

Any other comments? No for now.

DISTRIBUTION

Q: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

A: Yes, the dataset is publicly available on the internet.

Q: How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A: The dataset is distributed on github.com and can be downloaded at:

https://www.github.com/angeloobeta/Igbo-datasets where it's hosted. The dataset does not have a DOI and there is no redundant archive.

Q: When will the dataset be distributed?

A: The dataset was first released in November, 2019.

Q: 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.

A: The collected data copyright belongs to the authors of the books unless otherwise stated. There is no license.

Q: 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.

A: No.

Q: 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.

A : No

Any other comments? No for now.

MAINTENANCE

This section should be completed once the dataset has been constructed, before it is distributed. These questions help the dataset creator think through their plans for updating, adding to, or fixing errors in the dataset, and expose these plans to dataset consumers.

Q: Who is supporting/hosting/maintaining the dataset?

A: Obeta Ifeanyichukwu Malachy is supporting/maintaining the dataset.

Q: How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A: The owner/curator can be contacted via the email address: angeloobeta@hotmail.com

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

A: Since it's the initial release (v0.1) there have been no later release because this is the first time the datasets is been used for NLP.

Q: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

A: This will be posted on the dataset webpage where a README would be made available to describe the changes that were done on the datasets.

Q: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were 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.

A: N/A.

Q: 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.

A: The dataset has already been updated; older versions are kept around for consistency.

Q: 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.

A: The dataset is hosted on github as an open source project, anybody that wants to contribute can contact the collector about incorporating fixes/extensions through a github pull and push request.

Any other comments? No for now.