Data sheet Documentation

These are the sections and questions you need to answer when submitting your datasheet. Please note, some questions might not apply to your dataset. If they do not you can delete the question or respond with 'No'. See this paper for more info.

Please provide as much information as possible about your data set.

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

1. For what purpose was the dataset created?
   Answer: The purpose of creation of this dataset is to help data practitioners in Tanzania to practice their NLP skills in order to solve problems in organization and companies that involve swahili text dataset.

2. Was there a specific task in mind?
   Answer: Yes I want to work on swahili dataset to solve problem in different industry (such as Telecommunication) related to text classification , topic modelling, text summarization, sentiment analysis and fake news detection.

3. Was there a specific gap that needed to be filled?
   Answer: Yes In Tanzania currently most of dataset available are structured dataset involve continuous and categorical data types. Unstructured or *open text dataset* are not often available in Tanzania.

4. Who created this dataset
   Answer: I created this dataset by myself, Davis David by collect news from different news websites.

5. What support was needed to make this dataset?
   Answer: No support was provided to make this dataset,I use my own resources to collect dataset.

6. Any other comments?
   Answer: We need increase number of open swahili text dataset in Tanzania, last month I contacted fellow Data scientist who currently learn NLP for swahili dataset but she didn't have even a single open text dataset. This is a problem and we can solve it by creating more open swahili dataset.

Composition

1. What do the instances that comprise the dataset represent?
   Answer: Swahili News across different Tanzania news websites.

2. What data does each instance consist of?

Answer: The dataset involve 3 features which are

(a) **Titles**: This is the title of a particular news.

(b) **Category:** This is a category of a particular news. We have 5 news categories which are:-

(i) Habari za Kitaifa (in English National news)
(ii) Habari za Kimataifa (in English International news)
(iii) Habari za Biashaa (in English Business news)
(iv) Habari za Burudani (in English Entertainment news)
(v) Habari za Michezo (in English sports news)

(c) Content: This is a full description of a particular news.

The target associated with this dataset is the Category features means we can classify news according to their categories.

3. Does the dataset contain data that might be considered confidential?
   Answer: No

4. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
   Answer: No

5. Any other comments?
   Answer: I was able to collect 13,557 swahili news across different swahili news website.I'm plan to collect more swahili news with different category such as news related to healthy.

Collection Process

1. How was the data associated with each instance acquired?
   Answer: Swahili news dataset was collected across different Swahili news websites hosted within and outside the country.

2. Over what timeframe was the data collected?
   Answer: The dataset was collected over a period of three weeks.

3. What mechanisms or procedures were used to collect the data
   Answer:
   I have a list of websites to collect news with different categories.
   I use jupyter notebook, python programming language with different python packages such as pandas package to organize the collected dataset, BeautifulSoup package to perform web scraping from different websites.

4. What was the resource cost of collecting the data?
   Answer: Resources involved in data collection area ,High speed internet access and a laptop with good computational power.

5. Who was involved in the data collection process?
   Answer: I was the only one involved in data collection process.

Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done?
   Answer: Yes**,** scrapping news from different websites can result collecting the same news. I use pandas to identify duplicated news. After collect more than 15,000 news I found 1,300+ were duplicated news. Duplicated news were removed and remain with 13,557 unique news content.
   Also I believe other cleaning steps should be considered in this dataset.

2. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data?
   Answer: No

3. Is the software used to preprocess/clean/label the instances available?
   Answer: No, I just you pandas python package to organize the dataset

Uses

1. Has the dataset been used for any tasks already?
   Answer: No

2. Are there tasks for which the dataset should not be used?
   Answer: The dataset should not be used for any Political agenda or activities.

3. Any other comments?
   Answer: This dataset can be used for text classification, text summarization and topic modelling

Distribution

1. How will the dataset will be distributed?
   Answer: The dataset will be distributed via Github Repository.

2. Does the dataset have a digital object identifier (DOI)?
   Answer: No

3. When will the dataset be distributed?
   Answer: The dataset will be available in the Github Repository around March 2020.

4. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
   Answer: No

Maintenance

1. Who is supporting/hosting/maintaining the dataset?
   Answer: I will be maintaining the dataset **(Davis David).**

2. How can the owner/curator/manager of the dataset be contacted?
   Answer: My Email address is **davisdavid179@gmail.com**

3. Is there an erratum?
   Answer: No

4. Will the dataset be updated?
   Answer: The dataset will be updated to correct labeling errors and add new news. I will be responsible to update the dataset(and I have plan to get support from other data practitioners in Tanzania on this process).The updates will be communicated via Github Repository.

5. Will older versions of the dataset continue to be supported/hosted/maintained?
   Answer: The older version of the dataset will continue to be supported, hosted and maintained via it's Github repository.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
   Answer: Yes, I welcome others to extend and contribute to the dataset. If anyone is interested he/she can contact me via my personal email davisdavid179@gmail.com .

   The contributions will be validated/verified by identify the source of the dataset(news websites or platforms) if is correct, language used in the dataset.(the language must be a Swahili language) and the contents of the dataset don't have any kinds of errors.