Machine Learning for Humanitarian Data: Tag Prediction using the HXL Standard

Vinitra Swamy
Microsoft
AI Frameworks
viswamy@microsoft.com

Elisa Chen
University of California, Berkeley
Division of Data Sciences
cheneli@berkeley.edu

Anish Vankayalapati
University of California, Berkeley
Division of Data Sciences
vananish@berkeley.edu

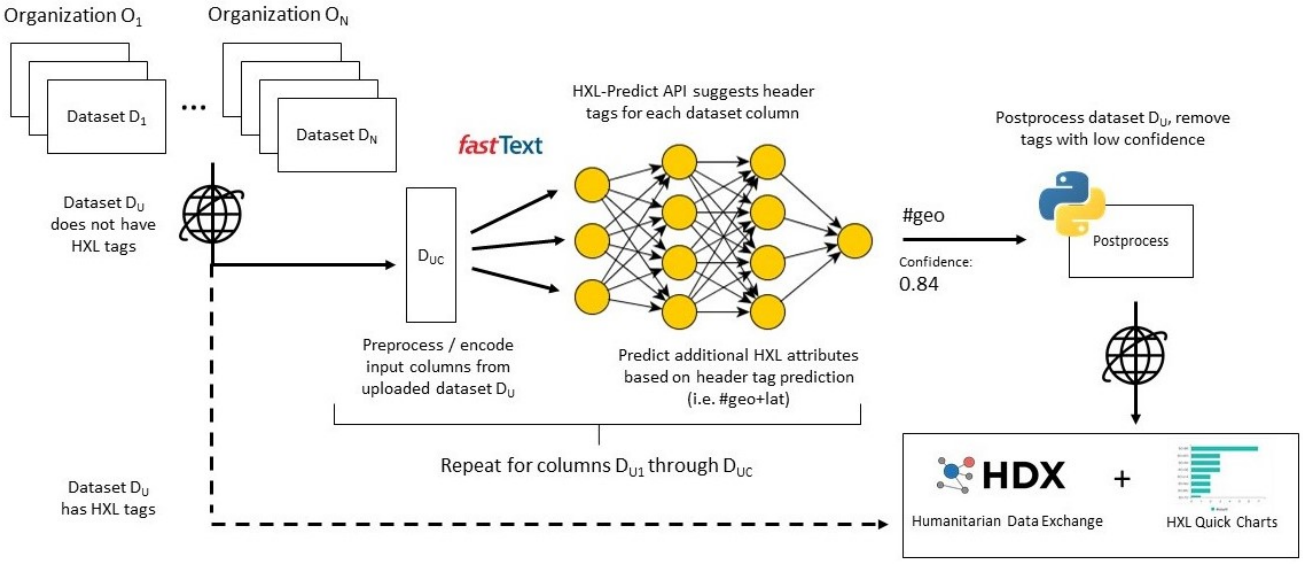
Abhay Aggarwal
University of California, Berkeley
Division of Data Sciences
abhaykaggarwal@berkeley.edu

Chloe Liu
University of California, Berkeley
Division of Data Sciences
ruochen99@berkeley.edu

Vani Mandava
Microsoft Research
Data Science Outreach
vanim@exchange.microsoft.com

Simon Johnson
UN Office of Humanitarian Affairs
British Red Cross
simonbjohnson@gmail.com

Figure 1: HXL Tag Prediction Workflow



ABSTRACT

Advances in natural language processing and machine learning (ML) enable the automation of humanitarian tasks that would traditionally require the expertise of a human expert. The labor-intensive process of data labeling requires crisis responders to spend valuable hours wrangling data instead of assisting with relief efforts. We present a machine learning model to predict tags for datasets from the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) with the labels and attributes of the Humanitarian Exchange Language (HXL) Standard [1, 2]. This paper details the methodology used to predict the corresponding tags and attributes for a given dataset with an accuracy of 94% for HXL header tags and an accuracy of 92% for descriptive attributes. Compared to previous work, our workflow provides a 14% accuracy increase and is a novel case study of using ML to enhance humanitarian data.

KEYWORDS

machine learning, humanitarian action, text tagging

ACM Reference Format:
Vinitra Swamy, Elisa Chen, Anish Vankayalapati, Abhay Aggarwal, Chloe Liu, Vani Mandava, and Simon Johnson. 2019. Machine Learning for Humanitarian Data: Tag Prediction using the HXL Standard. In *KDD '19: Workshop on Data for Social Impact, August 04-08, 2019, Anchorage, AK*. ACM, New York, NY, USA, 3 pages.

KDD '19, August 04–08, 2019, Anchorage, AK

1 INTRODUCTION

The Humanitarian Data Exchange (HDX) is an open data platform where recognized humanitarian organizations can upload their datasets to be stored with the United Nations [3]. This platform is operated by the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) for government organizations to have a central data source to make decisions about the distribution and implementation of aid efforts [1]. HDX utilizes the Humanitarian Exchange Language (HXL) Standard to tag data columns for standardization and interoperability [2]. Uploaded datasets that have been tagged with the HXL standard can utilize built-in analysis and interactive visualization tools that are generated from the HDX platform [4].

The goal for the collaboration between Microsoft, UN OCHA, and the UC Berkeley Division of Data Sciences is to build a predictive model to add HXL header tags to over 5000 untagged datasets, as well as create a tool to predict tags in real-time for new datasets. Our proposed model extends and formalizes work over the last year by predicting HXL tags (i.e. #org, #affected) as well as corresponding HXL attributes (i.e. #affected+refugee+f, representing affected female refugees), and draws upon FastText embeddings to pre-process input header names, column data, and metadata [5, 6]. In addition, we have implemented an API that allows users to tag their data efficiently and is being integrated into the HXL Proxy [4, 7].

2 PROBLEM BACKGROUND

Due to the manual effort involved, tagging datasets with the HXL standard is time-consuming for fieldworkers and can be done subjectively, making data analysis difficult. The Journal of International Humanitarian Action emphasizes the uses for big data in crisis response, lending to the need of data interoperability [8]. The Surge Information Management Support (SIMS) disaster response protocol from the Red Cross dictates that least two remote data specialists are required to guide a deployed team of two field responders for a crisis [9]. Any time the deployed team spends on data wrangling tasks is time that can be better spent contributing directly to relief efforts.

We propose a method to predict data tags upon upload (Figure 1) to save time and make data analysis tools (including visualizations and aggregations) more effective. Recent visualization tools with geo-linked maps and diagrams have proven useful in tracking epidemics and aiding humanitarian causes [10].

3 RELATED WORK

The model described in this paper builds off the work of Henrik Sjokvist, a volunteer Data Scientist at HDX [11]. Sjokvist was able to produce a proof of concept for a machine learning application based around 3 major pipelines:

(1) **Data Cleaning:** This consisted of converting all input into strings, splitting into individual words while removing excessive whitespace and converting to lowercase.
(2) **Feature Extraction:** This consisted of utilizing fastText to generate word embeddings for each word in the column headers [12]. The resulting vector would consist of 300 floating point numbers between -1 and 1.
(3) **Modeling:** Sjokvist obtained accuracy results of around 80% using a Multilayer Perceptron (MLP) classifier, which trains a supervised learning algorithm on a dataset to fit a non-linear function [13].

3.1 FastText

FastText is a library for text classification and vectorization developed by Facebook [6, 12]. It uses a hierarchical classifier instead of a flat structure, which speeds up the text classification process by several orders of magnitude. We utilized FastText embeddings to generate vectorized representations for the headers and datapoints in the input dataset and "warm-start" our learning representation process. The pre-trained FastText model used specifically for this project is English word vectors, which were trained mainly on Wikipedia and allow for a simplified data structure.

3.2 HXL Standard

The HXL standard is a data standard that aims to improve information sharing during humanitarian crises without adding extra reporting burdens [2]. Inspired by social media, HXL uses hashtags to categorize data in humanitarian datasets. These hashtags, examples of which include #country or #contact, are added in the row between the headers and the first row of data. If the data does not match an existing hashtag, organizations can also add their own customized hashtags and attributes. For the purpose of data interoperability, the proposed model aims to predict standard HXL tags from the HXL Standard dictionary [14].

4 METHODOLOGY

4.1 Web Scraper

The initial step in building the process for automating data-tagging process is through scraping the datasets which are used to train the model. For this purpose our team built a webscraping script primarily using the HDX Python Library, which is a Python API for interacting with HDX Data Portal [15]. The outcome of our webscraping process is a dataframe that separates the headers, the tags, the attributes, the data, the dataset name and the organization in distinct columns. We chose to filter out datasets of similar structures as to not overfit our data.

4.2 Model

The dataframe was preprocessed by lower-casing all words and removing certain punctuation marks that would skew the prediction. Our preprocessing uses FastText to vectorize each feature to create embeddings that would act as the inputs to the MLP Classifier. We embedded headers, organization name and the first seven rows of data for each column, which were all flattened into a single data structure for the final classifier. Examples of headers and predicted tags are included in Table 1.

Table 1: Example Input and Predicted HXL Output

Input Header	Data points	Predicted HXL Tag
year	2005, 2007, 2009	#date
percent funded	406504, 2217307, 2939092	#value+funding
center id	HM0096, MOH035, HM0097	#loc_id

During our model construction, we experimented using n-grams and bag of words techniques to vectorize the data, but achieved faster results using embeddings from FastText. The accuracy of the model was tested against multiple SciKit-Learn classifiers including the K-Nearest-Neighbors Classifier, Random Forest Classifier, MLP Classifier and Naive-Bayes Classifier [16]. MLP Classifier reached the highest accuracy on the test set and was chosen as the final classifier for the model.

The hyperparameters for the model were determined using cross-validation methods building off of the previous work, and used ReLU activation layers, a learning rate of 0.001, epsilon 1e-08, and hidden layer sizes of 150. The model was trained and tested against a test size of 33% of the input sample of 3659 datasets.

The model outputs the three most likely tags for the given header (along with the accuracy of the classifier as compared to the already-given tags of the test datasets). Tags with a confidence level of less than 0.5 were discarded and left blank in the predicted result. The output of the model is a pickle (pkl) file that is then input into the Flask API [7].

5 RESULTS

Using the header, organization name and the data content as features in the model, we obtain an accuracy of 95% using the MLP Classifier. The accuracy for predicting attributes was 92% using the headers as features. This is a roughly 15% improvement compared to the previous model which had an approximate 80% accuracy on its test set. Table 2 illustrates how the model performed against different classifiers.

Table 2: Tag-Predict Model prediction accuracy by classifier

Classifier	Accuracy	Hyperparameters
K Nearest Neighbors	90.4%	K = 3
MLP Classifier	94.3%	hidden layers = 150
Random Forest Classifier	67.7%	max-depth = 5
Naive-Bayes Classifier	62.5%	

6 FUTURE WORK

6.1 HXL Dashboard

The integration of HXL tags in data crowd-sourcing platforms will lend datasets credibility within the wider development network and enable the development of features including a UI for graphical data comparison, quick analysis, and interoperability. The API can generate predicted tags for the datasets uploaded by the users that currently support CSV and JSON files [7].

6.2 Attribute Prediction

A follow-up model predicts the additional HXL attributes (denoted by a "+" sign in the HXL Standard) for each header, to make the labeling more descriptive. The current work that has been done on this take a form similar to the HXL tag predict model that is described above, drawing on the headers, organization name, and data points to predict the attributes. Future work involves refining the attribute prediction model and using the predicted tags to limit the classifier input to identify possible attributes.

6.3 ONNX: Open Neural Network eXchange

In the spirit of model and data interoperability, future exploration involves exporting and continuing to optimize our model for real time inference performance using the ONNX (Open Neural Network eXchange) model format [17]. We intend to use ONNX Runtime for inference and have seen preliminary positive results [18].

7 ACKNOWLEDGMENTS

This applied data science research was supported by UN OCHA (Office for the Coordination of Humanitarian Affairs), Microsoft AI and Research, and the UC Berkeley Division of Data Sciences. We thank Anthony Suen for hosting this project as part of the UC Berkeley Data Science Discovery Project initiative. We also thank Mohammed Musbah of Microsoft Research Montreal for his helpful natural language processing expertise on this project. The basis and inspiration for this work was done during the summer of 2018 by Henry Sjokvist.

REFERENCES

[1] Un ocha. https://www.unocha.org/.
[2] Humanitarian exchange language (hxl) | a simple standard for messy data. http://hxlstandard.org/.
[3] Humanitarian data exchange. https://data.humdata.org/.
[4] Hxl proxy. https://proxy.hxlstandard.org/about.html.
[5] humanitarian-data-collaboration/hdx-python-model: Jupyter notebook of the model. https://github.com/humanitarian-data-collaboration/hdx-python-model.
[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
[7] humanitarian-data-collaboration/hxl-tag-api: Flask api to predict hxl tags. https://github.com/humanitarian-data-collaboration/hxl-tag-api.
[8] Junaid Qadir, Anwaar Ali, Raihan ur Rasool, Andrej Zwitter, Arjuna Sathiaseelan, and Jon Crowcroft. Crisis analytics: big data-driven crisis response. *Journal of International Humanitarian Action*, 1(1):12, 2016.
[9] Sims: Red cross. http://rcrcsims.org/.
[10] Rob Lemmens and Carsten Kessler. Geo-information visualizations of linked data. 2014.
[11] Ocha-dap/tag-predict: Poc for hxl tag prediction using fasttext word embeddings and mlp classification. https://github.com/OCHA-DAP/tag-predict.
[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
[13] Scikit-learn mlpclassifier documentation. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. (Accessed on 05/31/2019).
[14] Hxl hashtag dictionary. http://hxlstandard.org/standard/1_1final/dictionary/.
[15] Hdx python library. https://github.com/OCHA-DAP/hdx-python-api.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[17] Onnx. https://onnx.ai/.
[18] microsoft/onnxruntime: Onnx runtime: cross-platform, high performance scoring engine for ml models. https://github.com/microsoft/onnxruntime.