Columbia University Data For Good Scholars

Cost of Human Rights Violations Project

Spring 2020 - Summer 2020 update

Overview

The following document gives an overview of the work done by the Extension Project team of the Columbia University Data For Good (DFG) Scholars project from April to August 2020.

There are two main lines of work: (1) using unsupervised learning methods on proxy statements to separate irrelevant from relevant sections, investigate areas of existing and new materiality, and create a foundation upon which labeling proxy statements, if useful, would be an easier and less costly endeavor; and (2) using supervised learning methods on 10-Ks (both those excerpts labeled by SASB and others collected by the team) to flag and measure instances of risk disclosure for all General Issue Categories (GICs) of the Human Capital Management dimension, and the Supply Chain Management GIC of the Business Models & Innovation dimension by firms in industries for which those categories were not considered material for SASB's 2018 standards, and provide a model from which to conduct transfer learning onto proxy statements.

The goals of each of these workstreams and the approaches we took to accomplish them are listed below. The direction we've been moving in shows promise. By the end of 2020, we are on track to bring this work to a stage where it can start to support SASB's Human Capital Management project, as well as its other SASB standards revision projects.

The notebooks containing the code for the approaches below are saved at the link below:
https://drive.google.com/drive/folders/1spqS5oHw6gkFQxRIyJ929LvunnWXkX0I?usp=sharing

Unsupervised learning

Goals

Our goals for conducting unsupervised clustering on proxy statements were as follows:

1. Create a relevance classifier to identify only those sections of proxy statements that contain content relevant to the SASB standards, cutting down on the volume of documents that need to be analyzed by human analysts.
    - Approach: Cluster paragraphs of proxy statements to separate irrelevant from relevant paragraphs
2. Find potential new categories of materiality
    - Approach: Cluster relevant paragraphs and map clusters to SASB' s Human Capital Management dimension and Supply Chain Management GIC topic areas; if there are clusters that contain content that are outside the GICs within the Human Capital Management dimension and Supply Chain Management GIC, then consider whether such clusters could represent potentially new GICs
3. Provide a launching point for labeling proxy statements, so that SASB can apply a broader range of techniques to parse data
    - Approach: Create approximate category labels using clusters

Methods

After pre-processing, we used a K-means clustering algorithm and an LDA-mallet model to cluster paragraphs of proxy statements.

Pre-processing

For each proxy statement, we split the statement by line breaks into paragraphs, removing all excerpts shorter than 300 characters and duplicate paragraphs. For each of these excerpts, we removed all non-character text, converted all letters to lowercase, and used the Porter Stemmer from the NLTK package to stem all words. For the K-means model, we trained word embeddings using Word2Vec on the processed text of this labeled dataset, and used these embeddings to create 300-dimensional vectors for each word. The vectors were averaged for each word in the excerpt to create a single representation for each excerpt. For the LDA model, we used generic embeddings from Word2Vec that were pre-trained on a Google search dataset.

K-means clustering

We used a K-means model as one of our two clustering approaches, because of the simplicity of the algorithm and the tendency to give relatively good results on text clustering. After the pre-processing steps described above, we trained a Word2Vec word embeddings model on the proxy dataset, and created a 300-dimension vector representation of each paragraph using these embeddings. (These were averages of the embeddings of each word in the paragraph, rather than using Doc2Vec).

An elbow plot was used to determine the optimal number of clusters (6), and the clustering produced about 2 clusters whose paragraphs seemed highly suggestive of risk disclosures, and

5 clusters that seemed to contain irrelevant parts of the proxy statements. Here are some of the example paragraphs associated with cluster 2, the most promising of the clusters in terms of risk disclosures:

- Human Rights Policy. In February 2019, we adopted a human rights policy. This policy reinforces our commitment and responsibility to respect all human rights, including those of our employees, suppliers, vendors, subcontractors and other partners, and individuals in communities in which we operate. Our policy addresses promoting health and safety, eliminating compulsory labor and human trafficking, abolishing child labor, eliminating harassment and unlawful discrimination in the workplace and providing competitive compensation.
- Prison labor (both voluntary and involuntary) is often deployed in a manner that involves prisoner mistreatment and is frequently compared to **modern slavery**. Although companies benefit from low overhead expenses when inmates work for the company or its suppliers, companies have experienced public backlash, boycotts, and long-term brand name and reputation harm from a connection to prison labor;

Other approaches attempted:

Tf-iDF: Using a tf-idf approach, rather than word embeddings, yielded 7 clusters that had much less differentiation in their likelihood of containing risk disclosures:

- Topic 1: plan, termin, particip, benefit, employ, agreement, chang, award, payment, control
- Topic 2: audit, committe, independ, account, financi, public, firm, regist, statement, servic
- Topic 3: share, stock, common, ownership, benefici, compani, power, vote, outstand, director
- Topic 4: board, director, committe, govern, meet, independ, member, nomin, risk, corpor
- Topic 5: compani, compens, fund, manag, busi, execut, perform, committe, offic, year
- Topic 6: award, stock, grant, option, share, restrict, vest, exercis, plan, date
- Topic 7: vote, broker, propos, share, meet, proxi, instruct, nomine, annual, quorum

Next steps

The K-means model's parameters other than the number of clusters is not optimized using a grid or randomized search, so using these approaches to tune these parameters would be a logical next step.

Once we've used all possible approaches to achieve maximum comprehensibility of the clusters, we will select those that are most likely to contain risk disclosures and cluster them in order to unearth different General Interest Categories (GICs) within them. Through iterative clustering and elimination, we hope to eventually find clusters that can be used to further all

three of our goals for unsupervised learning: to separate irrelevant from relevant material, identify potential new clusters of materiality, and provide tentative labels for excerpts.

To determine with more certainty which clusters to eliminate, we plan on assessing the frequency of appearance of certain terms provided to the team by the Independent Advisory Group that are relevant to the Human Capital Management dimension and Supply Chain Management GIC. Clusters with high frequencies of these terms would be considered promising for the next round of clustering, while clusters where these terms and others collected from SASB's description of the GICs on its website are largely absent would be considered irrelevant and eliminated.

LDA-Mallet model

The LDA model was chosen because it is a standard for most text-based analyses; the mallet implementation was chosen because it is an efficient implementation that can be easily adjusted. The model produces 19 clusters, which are represented by their top key terms in Appendix A.

These clusters are not as clearly differentiated into relevant and irrelevant material as those used by the K-means model. As a result, the LDA-mallet model is not as useful in providing tentative labels for paragraphs of proxy statements as the K-means model.

Other approaches attempted

Prior to using word embeddings, we used count vectorization and tf-idf vectorization, which produced even more undifferentiated clusters - many of them had very similar top keywords.

Next steps

The word embeddings used for the LDA model are generic from Word2Vec and are pre-trained on Google search results, rather than embeddings, which are trained specifically on proxy statements. A logical next step is to replace these word embeddings with ones trained on the proxy dataset. There are also several parameters within the model we are not tuning, and the next step would be to tune these using a grid search.

In order to give the clusters more definition, we plan to have a similar cycle of cluster elimination and re-clustering as what we proposed with the K-means model. We anticipate with fewer irrelevant entries, the clusters will be able to more clearly differentiate between relevant and irrelevant material, and will be able to begin separating into categories that are more easily labeled as relevant to different GICs. In order to determine which clusters to eliminate with more certainty, we plan on performing the same keyword count as described for the K-means model.

Overall next steps

1. Tune all parameters for each model
2. Perform iterative cycles of cluster elimination and re-clustering

    a. Gauge relevance (irrelevance) for cluster retention (elimination) with more certainty using frequency of CSAG-suggested terms
3. Replace existing generic pre-trained word embeddings in LDA-mallet model with those trained on proxy statements

Supervised learning

Goals

1. Build an industry-agnostic classifier that can be used to identify reporting in “white spaces” in materiality map, using both current (2020) and “emergent” standards of materiality.
    - Approach: Training a classifier to have high recall and precision on SASB’s labeled dataset for the Human Capital Management dimension and Supply Chain Management GIC, and then applying the classifier to a broader corpus (in terms of both industries covered and time period) of new 10-Ks from firms in both industries that were and were not suggested to disclose on human capital and supply chain related metrics according to the 2018 standards.
2. Create a model that can provide a starting point for proxy statement classification.
    - Approach: Train a convolutional neural net on the labeled training data for human capital and supply chain management, and then use transfer learning to apply it to proxy statements to see the classification that arises.

Methods

Pre-processing

For each excerpt, we removed all non-character text, converted all letters to lowercase, and used the Porter Stemmer from the NLTK package to stem all words. We trained word embeddings using Word2Vec on the processed text of this labeled dataset, and used these embeddings to create 300-dimensional vectors for each word. The vectors for each word were averaged to create a single representation for each excerpt.

XGBoost

Our XGBoosted logistic regression model gives about 73 percent accuracy in identifying positive and negative (disclosure/no disclosure) cases in the labeled Human Capital Management dataset. Since we are targeting a recall threshold, rather than accuracy, we adjusted the probability threshold to consider any probability above 0.2 to be a positive case (rather than 0.5), which yields 93 percent recall and about 64 percent precision. A few summary statistics from after the threshold adjustment are below:

```
on test:
final recall score is:
 0.9274413529732679
final precision score is:
 0.6376594148537135
final accuracy score is:
 0.7336403296170625
final AUC score is:
 0.7530795949340827
```

Next steps

We hope to improve the recall and precision of the model by: 1) ensuring that the tuning was comprehensive; 2) removing certain stop words (informed by those that appear across most clusters in unsupervised learning methods); and 3) applying frequency rules based on Advisory Group's -suggested terms to convert some excerpts that are labeled negative to positive (i.e., switch label to positive if the excerpt contains a certain frequency of modern slavery-related words).

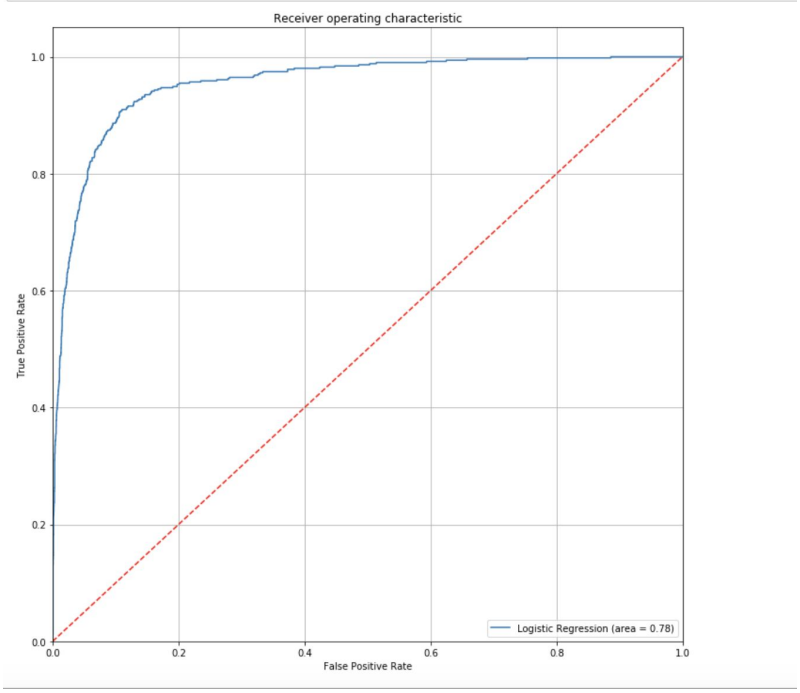
Logistic regression

The logistic regression achieved a recall of about 92 percent with a precision of 45 percent, with the threshold adjusted so that probabilities greater than 5 percent were considered positive and those less than 5 percent were considered negative. The accuracy and AUC metrics seem to be a reasonable range for a decent model, but underperform the XGBoost model.

```
Classification Report
              precision    recall  f1-score   support

           0       0.99      0.88      0.93      4647
           1       0.45      0.92      0.60       510

    accuracy                           0.88      5157
   macro avg       0.72      0.90      0.77      5157
weighted avg       0.94      0.88      0.90      5157
```



Other approaches attempted

A count-vectorization and tf-idf based approach (before using word embeddings) yielded lower precision for the same level of recall.

Next steps

Although the improvement of this model will be lower in priority than improving the XGBoost model (which currently yields higher precision at the same recall level) we plan on replacing the generic 10-K based word embeddings with embeddings that were trained on the same dataset. Afterwards, we may see if further parameter tuning or any of the improvements suggested for the XGBoost model will meaningfully increase the performance of this model. It is likely we will continue to use this model as a point of comparison, but that moving forward most efforts will be directed at the XGBoost model and fitting a convolutional neural net for transfer learning.

Neural nets

We fitted two types of neural nets on the dataset (plain two-layer net with the same word embeddings as used in the XGBoost and logistic regression models, and another net with a single LSTM layer using a function to create embeddings from within keras), and attempted to fit a third (convolutional neural net, or CNN, the most promising for transfer learning to proxy statements).

The results from the first two types of neural nets we attempted to fit were underwhelming, yielding an accuracy of about 0.68 in each case. We used an Adam optimizer and, as with the XGBoost model, a target of positive/negative flags for disclosure. We tuned the number of epochs and the learning rate using a grid search for the basic neural net and the net with an LSTM layer.

We coded up a two-layer convolutional neural net and a gridsearch to tune its parameters, but ran into hurdles trying to fit the net - the process consumed a lot of our local machines' resources and took quite a while. However, having a fitted CNN is particularly important moving forward as we pursue transfer learning on the proxy statements, so we are searching for ways to overcome this hurdle.

Other approaches attempted

We tried a few different ad-hoc changes to the nets to yield better accuracy, including adjusting the number of nodes in each layer, and using an SGD optimizer instead of Adam, all of which yielded worse results.

Next steps

For the basic neural net, we did not explore the full range of possibilities for the number of layers and nodes in each that we could have, and for the LSTM-based net, additional layers would have likely been helpful. We can also likely increase the fit by tuning a few additional parameters. For the CNN, we will try to modify the net, the gridsearch, or the data and explore the possibility of using a virtual machine to be able to fit the net in a reasonable time frame. Once the net is fitted, we can remove the last layer and apply it to our proxy statement dataset.

SVM

We chose to try to build an SVM classifier on the labeled data as well, but it yielded discouraging results even after making several improvements to the pre-processing (replacing count-based vectorization with generic pre-trained 10-K word embeddings). The maximum AUC was about 0.57. Since other models were more successful, we decided to forgo further improvement on this model in favor of devoting more attention to the others.

Overall next steps

1. The convolutional neural net seems the most promising for transfer learning, but we encountered processing power limitations while trying to fit it. We will try to give Google Colab (Google's open-source and cloud-based version of Jupyter Notebook another try

using TPUs (Tensor Processing Units) or try to modify either the net's parameters or the dataset. Once we have fitted the CNN, we will remove the last layer and apply it to the proxy statement dataset to see the classifications it yields, as a first step toward the goals of our unsupervised learning exercise: to separate relevant from irrelevant proxy statement text, find new and existing clusters of materiality, and to offer a launching point for labeling, if of interest.

2. Currently the logistic regression and XGBoost models seem to give the best results in terms of recall and accuracy. We will proceed with improving the XGBoost model by making sure the parameter tuning we've performed was comprehensive enough and investigating the possibility of blending it with another model. Once we are satisfied with the recall and the precision of this model, we will apply the classifier against a wider corpus of 10-Ks that we are in the process of downloading, which will span several years in time.
    a. On the latest 10-Ks, we will apply the classifier to see where risk disclosures appear by industry, specifically to determine whether disclosures in the human and supply chain management-related GICs are appearing with equal volume in those industries where published SASB standards deemed those GICs immaterial.
    b. On the full span of 10-Ks, we will see where disclosures related to the relevant GICs have been increasing over time to detect "emergent materiality."
3. We will implement frequency-based rules using the terms provided to us by CSAG in order to flag additional excerpts from the labeled dataset as relevant, and use setflag the terms provided to us by CSAG. This will likely help us as we apply these classifiers to a broader range of 10-Ks, including those in industries that the labeled dataset does not include, but may have more evidence of materiality in recent 10-Ks or show trends toward emergent materiality over time.

Conclusion and Fall 2020 Plan

The extension team's plan for Fall 2020 has three major components:

1. Continue refining unsupervised clustering approaches (using iterative elimination and re-clustering) until we arrive at clusters that contain content related to human capital and supply chain management-related GICs and potential new labor-related topics.
2. Hone supervised learning models (primarily XGBoost) to build an industry-agnostic classifier to apply to a broader range of 10-K statements to find disclosures present or beginning to appear in industries for which labor GICs were previously considered immaterial.
3. Fit a CNN on labeled 10-K dataset and apply to proxy statement dataset as a complementary method to (1) in identifying human capital and supply chain management-relevant portions of proxy statements and as a complementary method to

(2) in identifying evidence of materiality in industries previously without labor-related standards.

Appendix

Appendix A: LDA-mallet model clusters:

```
Remove Topics: 0,2,3, 14
Keep Topics: 7, 9, 11, 13
[(0,
  '0.171*"committe" + 0.072*"corpor" + 0.070*"govern" + 0.059*"board" + '
  '0.043*"nomin" + 0.035*"recommend" + 0.027*"consid" + 0.021*"member" + '
  '0.019*"determin" + 0.017*"candid"'),
 (1,
  '0.284*"director" + 0.222*"board" + 0.046*"meet" + 0.038*"elect" + '
  '0.036*"independ" + 0.036*"member" + 0.026*"annual" + 0.017*"chairman" +
'
  '0.017*"sharehold" + 0.015*"serv"'),
 (2,
  '0.184*"share" + 0.106*"stock" + 0.077*"common" + 0.030*"outstand" + '
  '0.023*"number" + 0.022*"ownership" + 0.021*"benefici" + 0.018*"issu" +
'
  '0.018*"class" + 0.017*"holder"'),
 (3,
  '0.249*"compens" + 0.038*"execut" + 0.032*"program" + 0.023*"pay" + '
  '0.020*"committe" + 0.018*"ceo" + 0.017*"employe" + 0.016*"total" + '
  '0.016*"incent" + 0.016*"annual"'),
 (4,
  '0.064*"agreement" + 0.051*"chang" + 0.049*"termin" + 0.048*"employ" + '
  '0.035*"control" + 0.027*"payment" + 0.026*"provid" + 0.024*"event" + '
  '0.024*"term" + 0.023*"period"'),
 (5,
  '0.055*"requir" + 0.044*"section" + 0.034*"amend" + 0.033*"tax" + '
  '0.032*"applic" + 0.029*"act" + 0.024*"exchang" + 0.023*"rule" + '
  '0.021*"code" + 0.020*"law"'),
 (6,
  '0.231*"execut" + 0.213*"offic" + 0.086*"chief" + 0.038*"financi" + '
  '0.035*"presid" + 0.024*"corpor" + 0.020*"compani" + 0.017*"posit" + '
  '0.017*"includ" + 0.016*"vice"'),
 (7,
  '0.067*"manag" + 0.062*"risk" + 0.041*"review" + 0.041*"polici" + '
  '0.040*"includ" + 0.039*"respons" + 0.031*"committe" + 0.018*"oversight"
+ '
  '0.018*"relat" + 0.018*"complianc"'),
 (8,
  '0.047*"experi" + 0.034*"serv" + 0.022*"busi" + 0.021*"presid" + '
  '0.019*"industri" + 0.018*"manag" + 0.018*"corpor" + 0.017*"sinc" + '
  '0.017*"technolog" + 0.017*"oper"'),
```

```
 (9,
  '0.072*"proxi" + 0.049*"stockhold" + 0.047*"annual" + 0.037*"materi" + '
  '0.032*"statement" + 0.031*"sharehold" + 0.031*"receiv" + 0.027*"notic"
+ '
  '0.022*"meet" + 0.020*"internet"'),
 (10,
  '0.388*"compani" + 0.115*"year" + 0.047*"decemb" + 0.035*"fiscal" + '
  '0.027*"end" + 0.026*"includ" + 0.022*"addit" + 0.018*"provid" + '
  '0.017*"februari" + 0.015*"prior"'),
 (11,
  '0.097*"audit" + 0.070*"committe" + 0.069*"account" + 0.066*"financi" +
'
  '0.065*"independ" + 0.041*"public" + 0.039*"firm" + 0.037*"report" + '
  '0.035*"servic" + 0.029*"statement"'),
 (12,
  '0.047*"serv" + 0.027*"bank" + 0.022*"univers" + 0.021*"member" + '
  '0.020*"sinc" + 0.017*"presid" + 0.016*"director" + 0.014*"board" + '
  '0.011*"chairman" + 0.010*"busi"'),
 (13,
  '0.146*"vote" + 0.075*"meet" + 0.049*"proxi" + 0.045*"propos" + '
  '0.033*"annual" + 0.023*"nomine" + 0.020*"person" + 0.019*"broker" + '
  '0.018*"present" + 0.017*"sharehold"'),
 (14,
  '0.114*"award" + 0.073*"stock" + 0.070*"grant" + 0.062*"option" + '
  '0.049*"vest" + 0.044*"valu" + 0.038*"unit" + 0.035*"date" + '
  '0.035*"restrict" + 0.034*"exercis"'),
 (15,
  '0.038*"person" + 0.035*"transact" + 0.030*"relat" + 0.029*"interest" +
'
  '0.021*"includ" + 0.020*"secur" + 0.019*"busi" + 0.017*"expens" + '
  '0.015*"affili" + 0.014*"parti"'),
 (16,
  '0.021*"busi" + 0.018*"continu" + 0.013*"includ" + 0.011*"develop" + '
  '0.011*"oper" + 0.011*"growth" + 0.010*"support" + 0.010*"leadership" +
'
  '0.010*"believ" + 0.010*"success"'),
 (17,
  '0.113*"perform" + 0.058*"base" + 0.036*"target" + 0.024*"annual" + '
  '0.023*"cash" + 0.022*"incent" + 0.022*"achiev" + 0.022*"goal" + '
  '0.019*"salari" + 0.019*"period"'),
 (18,
  '0.083*"fund" + 0.061*"invest" + 0.043*"manag" + 0.033*"sinc" +
0.028*"llc" '
  '+ 0.027*"trust" + 0.023*"truste" + 0.020*"capit" + 0.020*"incom" + '
  '0.019*"servic"'),
 (19,
  '0.140*"plan" + 0.069*"particip" + 0.044*"benefit" + 0.042*"employe" + '
```

```
'0.032*"amount" + 0.027*"retir" + 0.022*"year" + 0.021*"paid" + '
'0.019*"contribut" + 0.019*"elig"')]
```