Using Data Science to Facilitate Civic Action

Eric Kingery
Public Good Software
Chicago, USA
ekingery@publicgood.com

Q. McCallum
Independent Researcher
Chicago, USA
research-papers@qethanm.cc

Stephan Brown
Public Good Software
Chicago, USA
stephanrb3@publicgood.com

***Abstract*—Public Good Software (PGS) makes the news actionable: we connect readers, publishers, NGOs, and brands to address social causes, including the UN Sustainable Development Goals. The common element is news content, which we analyze to pair a given article with a relevant social cause. In turn, we provide a set of actions readers can take to help the cause.**

News content analysis is a labor-intensive process. We employ text-matching, natural language processing, and machine learning to quickly and efficiently pair articles with relevant social causes. In this paper we describe the evolution of the PGS platform, which employs data science to facilitate civic action and engagement.

***Index Terms*—Machine Learning, Natural Language Processing, Journalism**

Introduction

Public Good Software (PGS) matches publishers' news content to social causes. In matching an article to a social cause, we provide readers the opportunity to make a positive impact through actions including fundraising, advocacy, and education. The high volume and unpredictable timing of the news makes manual content analysis slow, inconsistent, and labor-intensive. We also incur a high marginal cost of labor to add new publishers. Manual analysis, simply put, does not scale. We therefore employ text-matching, natural language processing, and machine learning to analyze and classify news content.

This has proven to be a challenging problem, requiring us to evolve our business model, technology, and product offerings as we expand our understanding of the problem space. We divide this evolution into two phases, the first being an exploration of our Taxonomy for Social Good, and the second being a system to rate an article's similarity to a cause. Before proceeding, we introduce the PGS Impact Unit, which connects readers, publishers, NGOs, and brands through news articles that are relevant to social causes.

PGS Impact Unit

The PGS Impact Unit is delivered via an HTML <script> tag and a corresponding tag, which a news publisher places in their articles through their content management system (CMS). Figure 2 is a diagram of our technology platform.

The Impact Unit engages readers by giving them the opportunity to take actions such as learning more about the cause, volunteering their time, donating money, or sharing information about the cause with others. This not only helps the cause directly, it gives publishers, NGOs, and brands insight into what the readers care about and the different ways they are willing to interact with the cause.

Phase 1: A Taxonomy for Social Good

The Original Take Action Button

The first media-focused PGS product took the form of a Take Action Button. Content publishers could embed the static button in their news sites [1]. Publishers could manually place the button and choose the landing page relevant to each article, or they could allow PGS to analyze the content and automatically choose the landing page. The analysis involved scraping keywords from the article and feeding them to an Elasticsearch percolate query [2], which searched for relevant organizations and social causes that were already registered within the PGS system..

The original Take Action Button left us with three key takeaways:

- The manual matching and placement process yielded good matches of content to causes, but was labor-intensive and placed too high of a demand on newsrooms.
- The button redirected readers away from the publisher's site, which was a disincentive to placement and adoption by many media organizations.
- The keyword scraping / search solution often returned poor matches.

Figure 1. The original Take Action Button

Journalistic Article Analysis

hose takeaways from the original Take Action Button guided the next step in the evolution of our technology platform. We explore this stage at length in a previous paper, Automating, Operationalizing and Productizing Journalistic Article Analysis [3]. When this stage was complete we had the ability to store, process, analyze, and classify any article according

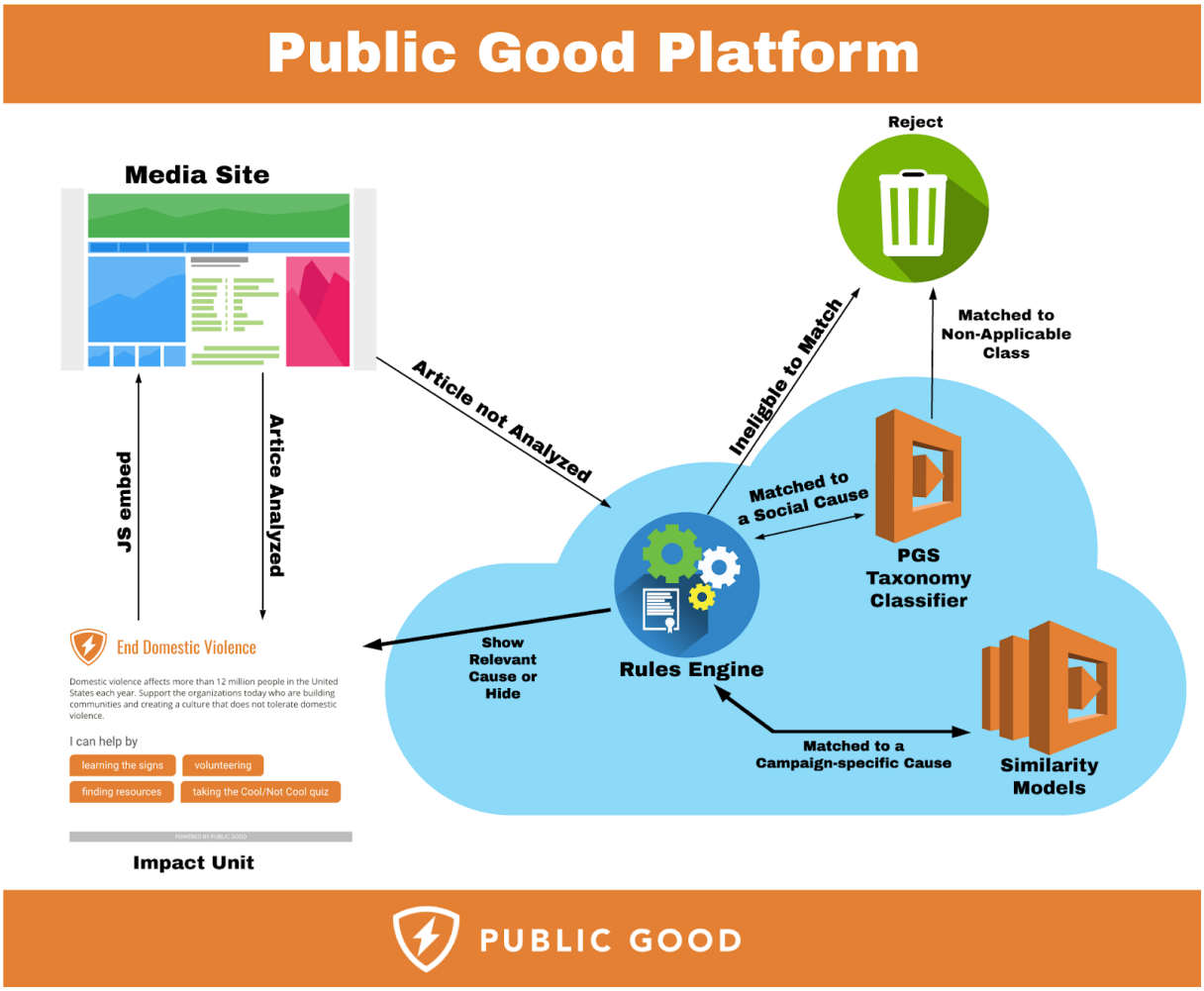

Figure 2. Outline of the PGS Platform

to our custom PGS Taxonomy of 27 social causes (discussed below). These technological advancements allowed us to improve the product and user experience. We gained the ability to show the reader the name of the cause in question before they clicked the button to take action. This stage was a valuable step forward in our journey towards making the news actionable in a viable, cost-effective manner.

After deploying our first machine learning (ML) classifier, we introduced a deterministic method for analyzing articles known as the Rules Engine. Through a mix of procedural code and regular expressions, the Rules Engine allows us to match articles based on keywords, hide the Impact Unit based on media partner requests and cause-related preferences, and otherwise accommodate business rules that should not be left to a predictive model.

The Rules Engine empowers PGS team members who are not software engineers to launch and configure both network-wide and publisher specific rules in real-time. Furthermore, the Rules Engine narrows the pool of articles requiring ML classification, which reduces our exposure to false positives and false negatives.

PGS Taxonomy of 27 Social Causes

We originally used data from the National Taxonomy for Exempt Entities (NTEE) to classify NGOs, and eventually developed our own taxonomy to match news articles with social causes. Our custom social cause taxonomy was a key aspect in the evolution of both our product and the technology platform.

The NTEE data proved inadequate due to its size, inaccuracy, and inconsistency. We also observed misalignment between the stories in the news and the way NGOs classified themselves within NTEE. We needed to create fewer categories that would better describe and integrate the target populations appearing in the news with the underlying social causes addressed by the NGOs.

We started with the UN Sustainable Development Goals (SDGs) [4] as a model. The SDGs proved to be a comprehensive yet succinct foundation upon which to base our custom taxonomy. We supplemented the SDGs with other taxonomic sources and additional dimensions of data which allowed us to closely match the social causes appearing in the news with relevant NGOs addressing those causes.

The PGS Taxonomy has 27 imperative social causes such as ”Support LGBTQ Rights,” ”End Domestic Violence,” ”Fight Racism,” and ”Save the Environment,” as well as 10 general-purpose / non-cause-related categories such as ”Lifestyle,” ”Sports,” and ”Financial.” The following diagram shows how we mapped the two taxonomies to each other:

UN Sustainable Development Goals	PGS Taxonomy
No Poverty	End Poverty
Zero Hunger	
Good Health & Well Being	Support Public Health
	Support Mental Health
	Fight Cancer
	NA - Health - Other
Quality Education	Support Quality Education
Gender Equality	Support Women's Rights
Clean Water & Sanitation	
Affordable & Clean Energy	
Decent Work & Economic Growth	
Industry, Innovation & Infrastructure	NA - Business / Financial / Marketing / Advertisin
Reduced Inequalities	Support LGBTQ Rights
	Fight Racism
	Stand Against Hate
	Support Immigrant Rights
Sustainable Cities & Communities	Help Refugees
Responsible Consumption & Production	Protect Consumers
Climate Action	Save the Environment
Life Below Water	
Life On Land	End Animal Abuse
Peace, Justice & Strong Institutions	Fight Crime
	End Child Abuse
	Stop Gun Violence
	End Violence
	End Domestic Violence
	End Sexual Violence
	Stop Bullying
	Fight Terrorism
	Keep Government Honest
	NA - Crime / Legal - Other
	NA - Government - Other
	NA - Social Affairs
Partnership for the Goals	

Figure 3. A look at how our categories map to SDGs

The PGS Taxonomy enables us to reframe the UN Sustainable Development Goals to be more actionable by consumers reading the news. For example, SDG 10, Reduced Inequalities, is too broad for effective matching with news articles. The PGS categories “Support LGBTQ Rights,” “Stand Against Hate,” “Support Immigrant Rights,” and “Fight Racism,” all map to this goal. These categories proved to be broad enough and distinct enough to be accurately labeled by our ML classifier.

Presenting Impact Units on news articles relevant to an actionable social cause has allowed us to help policymakers, brands, and influencers gain insight into what consumers care about and which actions they are most interested in taking. We have worked with organizations such as The Gates Foundation to further explore these topics. This data helps policymakers and NGOs progress towards accomplishing UN SDGs when related issues gain consumer attention via the news.

PGS Taxonomy Proven in Production

Our first ML classification model used the Naive Bayes algorithm, trained on roughly 6,000 manually-labeled news articles from a subset of our publishers. This model achieved roughly 50% accuracy in production, compared with a random-choice accuracy of 2.7% (1/37). It succeeded on similar yet distinct categories, and often failed when classifying into broader categories.

We tried a variety of off the shelf tools (Google Natural Language API, IBM Watson Natural Language Classifier), algorithms, and tuning parameters to classify content across the 37 total categories in the PGS Taxonomy. We experimented with k-nearest neighbors, support vector machines (SVM), Vowpal Wabbit, fastText, and ensemble methods including random forests and voting classifiers which combined results from many of these methods. Our LinearSVC model demonstrated the best performance, with 65% accuracy. We therefore run the LinearSVC model in production.

While LinearSVC was a significant improvement over Naive Bayes, the pure accuracy metric did not reflect our business mission. We cared more about eliminating false positives (showing an Impact Unit that connected an article to an unrelated cause) than false negatives (hiding the Impact Unit when the article matched the cause). We also wanted to focus on the 27 social cause categories, not the 10 non-applicable ones. Furthermore, we needed a metric that would account for an article’s view count.

We developed two new KPIs as a result. Both are similar to the pure accuracy metric – number of correctly classified articles over total number of articles – but they focus on actionable articles:

- number of actionable-and-correct articles over total actionable articles
- number of actionable-and-correct-views over total actionable views

Based on our test data, our LinearSVC implementation scored 52.2% and 18.9% on those KPIs, respectively. It is important to note that the performance of the PGS technology platform is stronger than that of the classification model on its own. To ensure quality in production, the platform runs articles through the Rules Engine and manual QA before and/or after the automated analysis.

Furthermore, these KPIs do not reflect the overlap between some classes in our taxonomy, which means classifier results can be subjective. For example: an article about a hurricane could reasonably match ”Stop Climate Change,” and it could also match ”End Poverty” if the storm had greater impact on the homes of the area’s poor.

Similarly, the KPI results do not reflect the performance of the classifier within each category in the taxonomy. We found the classifier performs up to 30% higher on categories that had a more distinct vocabulary, such as ”Disaster Relief” and ”Fight Cancer.” These results indicate significant traction on the problem. Despite these mitigating factors, there remains room for worthwhile gains in the future.

Challenges

In 2018, our platform classified 4 million articles and served nearly 1 billion requests. This is a major accomplishment and our partners are satisfied with the quality and quantity of matches. However, we still strive for increased performance. During the course of our evolution, we have unveiled two primary technical challenges: matching accuracy and scale.

Several issues influence the classification model's accuracy. We have a large number of categories (37) compared to a fairly small training set (roughly 45,000 articles). Our training data is imbalanced, as some categories have several times more representation than others. Some categories overlap in cause and vocabulary. Additionally, content length varies and some articles were as short as 200 words with a very limited vocabulary. All of these factors compound one another, and ultimately affect the model's ability to classify articles into the appropriate category. We continue to investigate other techniques, such as neural networks, to address this as our training dataset grows.

Given the current classifier performance, we employ people to perform Quality Assurance (QA) on content. Their work must scale as the number of publishers, articles, and social causes increase. To make our QA team's work more efficient, we have built a set of workflow tools to best identify high priority and/or low confidence items that merit human review. We also generate keyword-based rules for new campaigns using ML and NLP-assisted manual intervention.

In addition to these technical challenges, our business model has evolved to focus on a wider variety of highly specific social causes. Our machine learning classifier and the fixed PGS Taxonomy have proven insufficient in capturing the nuances of, and matching on, these new social cause initiatives. This has led us to investigate a new machine learning approach.

Phase 2: Detecting Cause Similarity: More Like This

From Static Categories to Dynamic Social Causes

In 2018 we launched a new service offering: PGS partners with brands to match news content to social causes they are working on. See Figure 4 as an example, we partnered with Unilever to raise awareness for The Right to Shower brand [5].

The Right to Shower donates profits to mobile shower initiatives, providing people currently experiencing homelessness with access to basic hygiene services. The campaign focuses on the impact that access to these services has on individuals experiencing homelessness. This means The Right to Shower Impact Unit should be matched to articles that specifically cover mobile shower initiatives, as well as broader topics such as poverty, hygiene, and employment security. This single, relatively narrow campaign cross-cuts multiple UN Sustainable Development Goals such as No Poverty, Zero Hunger, and Clean Water. At the same time, it did not precisely match any of the social causes in the PGS Taxonomy. Our taxonomy proved insufficient when describing the real world. It became clear we needed a new system through which we could build a corpus of existing content related to a very specific social cause, and then evaluate new articles for similarity to that corpus. We named this system the More Like This engine.

The More Like This engine is under active development. We are currently researching two backing implementations, one based on the Gensim library [6] and another that uses transfer learning (via Google's BERT toolkit).

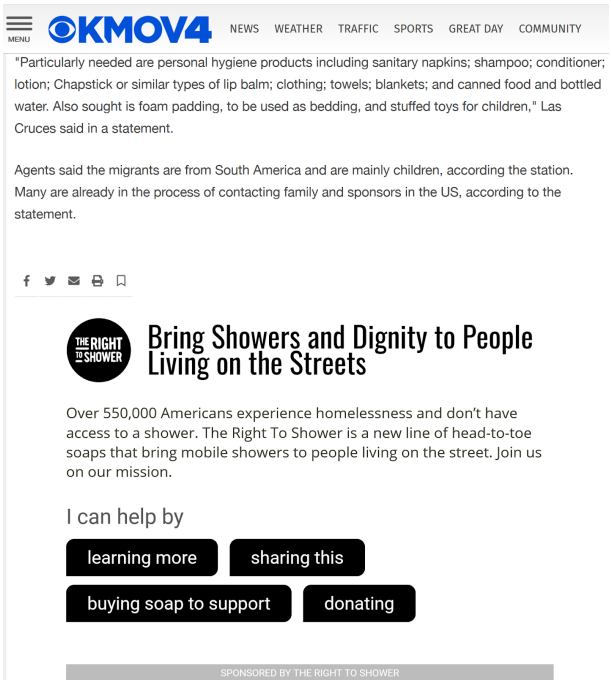

Figure 4. A look at the Impact Unit in The Right to Shower

Gensim

For our first attempt at the More Like This engine, we chose Gensim's builtin Similarities class. This class handles a lot of housekeeping for us, in that we can feed it a corpus of documents (technically, document vectors), test cosine similarity of those documents to each other, and then test new documents against all of those in the corpus. The Similarities class encapsulates a lot of document management so that we can focus on tuning our model and preparing it for production use.

The cosine similarity between two documents is in the range [-1, 1]. A score of 1 means that the two documents are exactly the same, and a score of -1 means that they are diametrically opposed. [7]

We used intra-corpus similarity as our baseline metric for new content: we tested all documents in the corpus against each other and took the average of the top-most-similar documents to determine a suitable minimum threshold value of cosine similarity. When we test a new, incoming document against the corpus, we mark it as cause-related if its similarity score exceeds the threshold and reject it otherwise. For our initial experiments on this test campaign, we settled on a cosine similarity threshold of 0.14.

Careful Re-Selection of Training Data

Our initial experiments in building a More Like This engine exposed some problems in our training data. The training set was purpose-built for a test campaign, based on a mix of result from manual QA and the Rules Engine. It comprised roughly 8,000 articles from our publisher network.

We initially noted that the Gensim-based system was more sensitive to news publishers' copyright notices and similar boilerplate text than our Linear SVC classifier had been. After we scrubbed out that boilerplate text, performance was still lagging. We performed a manual spot-check of the data and realized that a portion of our training data was not sufficiently related to our test campaign. There was enough noise in the training data as to have a material impact on the performance of the Gensim-based More Like This engine.

In an attempt to improve model performance, we sent our training data through another round of manual labeling. This shrank the test set to roughly 6,500 articles. What we've traded in training set size, we hope to make up in model performance.

Transfer Learning with BERT

We are also experimenting with transfer learning for the More Like This engine. Transfer learning is an ML methodology which uses a pre-trained model as a starting point for a model with a similar task or domain. By using a pre-trained model we cut down on development costs, reduce our training data requirements, and potentially boost the accuracy of our predictions. It also scales well, allowing us to retrain a model for each cause within minutes.

We chose Google's BERT for this task due to its excellent performance on NLP tasks. BERT is the first deeply bidirectional, unsupervised language representation. It is pre-trained using a text corpus from Wikipedia. BERT's bidirectionality is powerful enough to beat out other state of the art models on 11 different NLP tasks [8]. Integrating BERT into the PGS platform requires adding a softmax layer designed to predict whether an article is a match to a specific social cause.

CONCLUSION

The Public Good team employs data science to meet the challenge of making the news actionable. Our product has evolved, from a manual placement requiring manual analysis, to an automated workflow using multiple ML models while integrating human intervention in high-leverage situations. The development of a custom PGS taxonomy, mapped to the UN Sustainable Development Goals, was key to advancing our technology platform and product offerings. The underlying techniques have also evolved, from a static taxonomy and Naive Bayes model, to cause-specific models and transfer learning. Keeping humans in the loop has been critical to our success and has required that we maintain models and rules that are transparent and interpretable.

We also recognize the potential to use semantic understanding – such as, noting the difference between "dog bites man" and "man bites dog" – to improve our system. We are investigating knowledge-based approaches and Siamese LSTM networks, which have demonstrated some degree of semantic understanding. We expect our technology will continue to evolve to better capture semantic meaning within news articles.

By making the news actionable, Public Good provides insight on consumers' views of social causes to policymakers and NGOs. They use this information, looking backward, to see which causes resonate most with consumers. Looking forward, this information helps them decide how and where to apply their efforts to addressing those causes. While our technology has evolved, our ultimate goal remains the same: to revolutionize the way consumers, NGOs, and brands interact with media to make a positive social impact on our world.

ACKNOWLEDGMENT

We thank Daniel Ratner (CEO), Michael S. Manley (CTO), and Jet Traverso (Product Specialist), of Public Good Software, for useful discussions and assistance related to this paper.

REFERENCES

[1] "How some news outlets let readers 'take action' with a click." [Online]. Available: https://news.wttw.com/2018/03/26/how-some-news-outlets-let-readers-take-action-click

[2] "Elastic search reference." [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-percolate-query.html

[3] E. Kingery, M. S. Manley, and D. Ratner, "Automating, operationalizing and productizing journalistic article analysis," *CoRR*, vol. abs/1710.08522, 2017. [Online]. Available: http://arxiv.org/abs/1710.08522

[4] "Sustainable development knowledge platform." [Online]. Available: https://sustainabledevelopment.un.org/

[5] "The right to shower." [Online]. Available: https://www.therighttoshower.com/mission

[6] "Gensim: Topic modelling for humans." [Online]. Available: https://radimrehurek.com/gensim/similarities/docsim.html

[7] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4, 2008, pp. 9–56.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805