Text Analytic Research Portals: Supporting Large-Scale Social Science Research

Lisa Singh,*§ Colton Padden,* Pamela Davis-Kean,† Rabin David,*
Virinche Marwadi,* Yiqing Ren,* and Rebecca Vanarsdall*
*The Massive Data Institute, Georgetown University
Washington, D.C., USA
†Institute for Social Research, University of Michigan
Ann Arbor, MI, USA

***Abstract*—Large-scale organic data generated from newspapers, social media, television, and radio require an expertise in infrastructure management, data collection, and data processing in order to gain research value from them. We have developed text analytic research portals to help social science researchers who do not have the resources necessary to collect, store, and process these large-scale data sets. Our portals allow researchers to use an intuitive point and click interface to generate variables from large, dynamic data sets using state of the art text mining and learning methods. These timely variables constructed from noisy text can then be used to advance social science research in areas such as political science, economics, public health, and psychology research.**

***Index Terms*—text analytic tools, variable construction, research portal, organic data, social media**

I. Introduction

Researchers at Georgetown University and University of Michigan have built a shared text analytic research portal that enables social scientists to generate structured variables from text using state of the art natural language processing and data mining. Specifically, we provide access to aggregate variables constructed from millions of social media posts, newspaper articles, television transcripts, and open-ended survey questions for different domain, e.g. US Presidential election. Over time, we anticipate this portal serving as a hub for constructing variables from organic data sets using methods that may not be as mainstream to those in disciplines outside of computer science. To date, we have developed prototypes for multiple research portals, including ones for the US presidential elections [5], Covid-19 [6], gun violence, forced migration [8], and parenting/child behavior [3], [4].

The portals[1] are accessible to researchers through an intuitive web interface and the variable construction components are accessible through web forms. Researchers specify a data set, date range, units of analysis, algorithm, and algorithm parameter. The portal then uses this information to generates variables. Algorithms we support for generating different types of variables, include word and emoji frequencies (n-grams), sentiment analysis, storyline/event detection, topic analysis, and location detection. Future additions will include emotion and stance detection [2].

§Contact author: Lisa Singh - lisa.singh@georgetown.edu

[1]Portals at: https://portals.mdi.georgetown.edu

II. Architecture

In order to run algorithms, e.g. topics, sentiment, word frequencies, etc., on the large, organic data sets, we rely on a variety of services and technologies. The Google Cloud Platform is used to ensure scalability, reliability, and cost-effectiveness of our application and pipelines components. The web-interface and accompanying API are hosted on Google App Engine. Job and user specific metadata are stored in a Postgres database. Build artifacts, collected data, and computed aggregates are stored in Cloud Storage. Long-running data collection jobs run on compute instances, batch-processing runs on ephemeral Dataproc clusters, and scheduling and orchestration is done with Apache Airflow.

Processing components have been designed to be generic and extensible to support a variety of data sources and to expedite the introduction of new data sets and algorithms. The use of libraries and architectures that support scalable infrastructure allow us to run the same code on both small, and very large data. Given the volume and diversity of data we maintain, we conducted a series of experiments to determine that using custom distributed Spark processing is significantly more cost effective with only a small loss in performance when compared to using other Google distributed capabilities, like Big Query. The difference in cost results from the data processing and analysis, not the data storage.

A. Interface

Figure 1 shows the components of the data processing request initiated by a researcher. The web-interface allows researchers to request processing jobs on targeted data. Job parameters are supplied to further refine the data by language and keywords and to provide algorithm specific configurations. Jobs are added to a queue and are scheduled to periodically run with Apache Airflow. Apache Airflow is a workflow management platform where ephemeral clusters are created, jobs are run, and computation results are put into CSV files. Because we batch jobs to save on resource costs, after the results are compiled, they are emailed to the researcher.

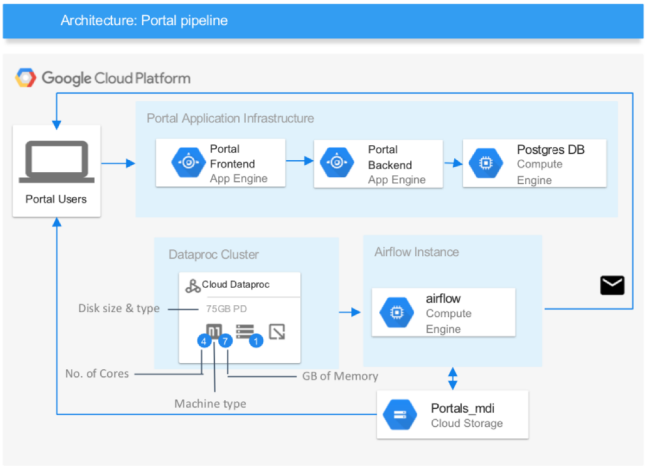

Fig. 1. Data processing requests components

B. Data Collection

The majority of our data are collected through APIs, RSS feeds, web scraping, and downloading of other publicly available data. Social media conversation is limited to that which is available publicly, and the scope of data is specific to project-specific keywords and topics. In the case where data are streamed, long-running processes collect data and periodically batch them for upload. The unaltered raw data is archived in cloud storage, and staged for further pre-processing and normalization.

C. Data Normalization & Pre-processing

These organic sources are noisy. For example, social media posts may be grammatically inconsistent and contain misspellings, abbreviations, and emojis. Therefore, text pre-processing is important for many algorithms, including sentiment and topics. Some pre-processing options for text data include: tokenization, stopword and flood word removal, emoji conversion, contraction expansion, stemming, lemmatization, part-of-speech tagging, and language detection. When multiple sources are used for the same data set, attributes are normalized for consistency. Data is partitioned by date, language, and geographic region, and converted into a standard format like Parquet and JSON.

D. Features

While different types of data portals and data hosting technologies have existed for a while, e.g. DataVerse[2] and DataKind[3], the text analytic research portals presented here differ in that their goal is not just to share data, algorithms, and code, but to emulate the process social scientists follow to construct variables supporting their research. The MDI technical team worked with an interdisciplinary team of researchers to design the portal so that the process of constructing variables mimicked the process social scientists would use for more traditional, smaller scaled studies. This includes having ways to assess reliability by sampling and labeling subsets of data, and running algorithms with custom parameters, e.g. using custom stopword lists. Through an iterative feedback process with portal users, we continually improve the interface design, usability, and functionality.

[2]https://dataverse.org/
[3]https://www.datakind.org/

III. Case Example

The steps required to submit a processing job vary depending on the project, data set, and algorithm. However, a commonality exists between these tasks. After selecting an algorithm (e.g. sentiment analysis), the user is presented with options to choose a source of data, specific filtering options for that source, a date range, and optionally, a language. Additional filtering options that are presented depend on the data source selected.

For the Twitter source, the user chooses whether he/she wants to limit the scope to a specific set of hashtags, keywords, or handles – selections corresponding to search terms that were used during data collection. The user also selects the type of post: tweets, retweets, and quotes. Figure 2 shows an example for the sentiment analysis task on our Covid-19 portal. For the Reddit data, the user is presented with an option to limit the conversation to specific subreddits. When the news articles source is selected, the option to choose specific publishers or geographic regions of these publishers are presented to the user.

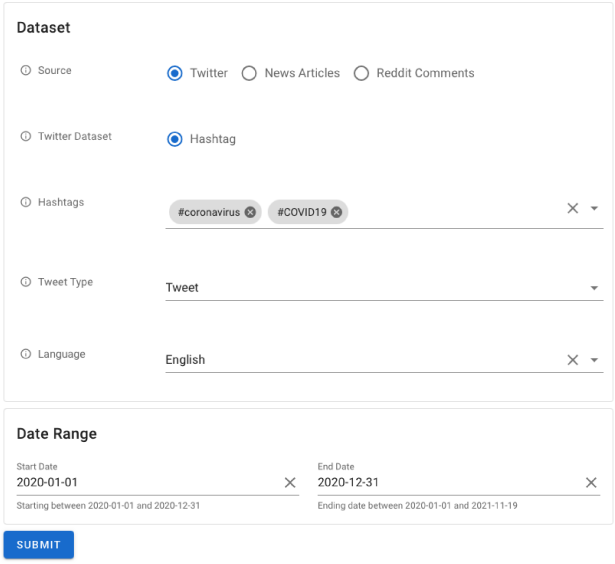

Fig. 2. Example form for submitting a sentiment analysis processing job using a Twitter data source.

After the form is submitted, the job is added to a queue and an entry is added to the *My Jobs* page on the user's profile. The result of our example case is a comma separated file with daily sentiment scores generated from three sentiment algorithms on the selected Twitter data. Once the job has completed, the user is notified via email and the entry is updated on the user's *My Jobs* page (see Figure3). This page maintains a list of all

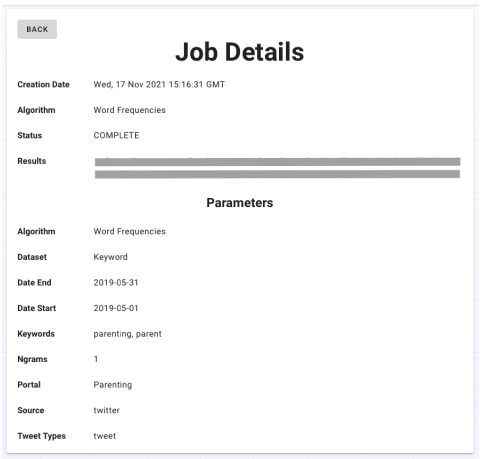

Fig. 3. A user's *My Jobs* page for a specific job

previously submitted jobs, their statuses, the parameters that were selected, and links to the results.

IV. Challenges

While we continue to add new portals, there are a number of challenges we face. First, working with different types of organic data that have limited consistency has required a significant amount of work on pre-processing options and normalization to standardize data in different ways for different portals. Second, the data associated with portals ranges in size, from 2 to 3 gigabytes to 40 to 50 terabytes. This variance in data size requires an infrastructure that is adaptable, scaling based on the amount of data being processed. Next, not all algorithms are readily adaptable to run in a distributed computing environment. This means that we either rewrite the algorithm or only allow certain algorithms to run on smaller data sets. Finally, our infrastructure must remain cost effective. This means we are continually working to optimize how we efficiently use an ephemeral infrastructure and effectively use storage of intermediate stages of the processed data to avoid re-computing values that are computationally expensive.

V. Conclusion

To date, we have approximately 200 terabytes of data and data collection continues daily. As MDI adds new data sources for different projects, random samples and aggregate variables constructed from these noisy, organic sources are made available for the research community through our text analytic research portals.

The adoption of the research portals has been promising, with a number of social scientists incorporating the results of the processing-pipelines as variables in their research analysis. The capability of using organic text data has proven to be an important avenue for improving our understanding of human beliefs, behaviors, and emotions. We also hope the portal helps democratize access to text analysis in domains where researchers may not be able to afford this type of compute infrastructure.

While our initial focus has been on organic data, we anticipate that over time we will support robust, large-scale variable construction based on data blending of both organic and design data, merging both timely data with higher quality data.

Many future improvements and features are planned, including support for new algorithms for existing tasks and new tasks as they are released and expanding processing from purely text-based approaches ones that include image and video analysis. We also anticipate supporting robust, large-scale variable construction based on data blending of both organic and design data to reduce the amount of missing data that exists for many societal scale problems. The addition of interactive visualizations and other data products will provide researchers with more advanced data exploration and analysis capabilities. Inclusion of streaming processing for near real-time data analysis will support online analysis of events such as debates. Finally, population weighting to better model demographic representations[1], [7] of different social media sites and understand who is sharing content is also an important future direction.

Acknowledgment

This research is supported in part by National Science Foundation awards #1934925 and #1934494, the Google Cloud Research Credits program award GCP19980904, the National Collaborative on Gun Violence Research (NCGVR), and the Massive Data Institute (MDI) at Georgetown University. We thank our funders for supporting this research.

References

[1] Liu, Y., & Singh, L. (2021). Age Inference Using A Hierarchical Attention Neural Network. In Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM).

[2] Kawintiranon, K., & Singh, L. (2021). Knowledge Enhanced Masked Language Model for Stance Detection. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).

[3] Ryan, R., Davis-Kean, P., Steimle, S., Waters, N. (2021). "Sharenting" on Social Media: Praise versus Criticism of Mothers and Fathers Online. Society for Research in Child Development 2021 Biennial Meeting.

[4] Davis-Kean, P., Ryan, R., Singh, L., Waters, N., Steimle, S., Liu, Y. (2021). The Kids are Not All Right: Educational Inequalities in the Time of COVID-19. Society for Research in Child Development 2021 Biennial Meeting.

[5] Bode, L., Budak, C., Ladd, J. M., Newport, F., Pasek, J., Singh, L., Soroka, S. and Traugott, M. W. (2020). Words that matter: How the news and social media shaped the 2016 Presidential campaign. Brookings Institution Press.

[6] Singh, L., Bode, L., Budak, C., Kawintiranon, K., Padden, C., & Vraga, E. (2020). Understanding high-and low-quality URL Sharing on COVID-19 Twitter streams. Journal of Computational Social Science, 3(2), 343-366.

[7] Jackson, M. McKinstry, J., McPhee, C., Raghunathan, T., Singh, L., Traugott, M., Turakhia, C. & Wycoff, N. (2021). Understanding who uses Twitter: State-level estimates of those on Twitter. MOSAIC Methods Brief: November 2021. Measuring Online Social Attitudes and Information Collaborative.

[8] Singh, L. Wahedi, L., Wang, Y., Kirov, C., Wei, Y., Martin, S., Donato, K., Liu, Y., Kawintiranon, K. (2019). Blending Noisy Social Media Signals with Traditional Movement Variables to Predict Forced Migration. ACM International Conference on Knowledge Discovery and Data Mining (KDD), Anchorage, Alaska.