Untapped Energy

Datathon For The Oil and Gas Industry
October 12 - 14, 2018

REPORT

Social License

Marc Boulet
Karen Morton
Christopher Lloyd
Wes Baird
Ivone Wardell
Joyce
Tessa Peterson
Jessica Liu
Keenan Viney
Alex Havers
Andrey Sivyakov

Overview

Every project, big or small, is comprised of people. People responsible for the project (Industry), people impacted by the project (Public), and the people responsible for ensuring it's completed responsibly (Government) all have a role to play in the life of a project. Combined, these groups act as the stakeholders in determining the fate of any project, but do we truly understand why do some projects go forward, while others do not? This concept is known as "Social License".

Initial insights and observations

- Not enough years of data to perform machine learning or predictive analytics
- Unstructured text files are hard! - So much of the "data" collected was not collected in a format designed for this type of analysis (e.g. letters of comment, Factiva news articles, media inquiries)
- Opportunity to link data in new ways and visualize geospatially

Actions items and the path forward

Dashboard prepared to link media assessment with application descriptions

Objective

Our objective is to explore what successful and unsuccessful Social License looks like through data of perspective of major oil and gas infrastructure projects by examining formal (regulator and federal government statistics) and informal (traditional and social media) data.

Technical Summary

- ## External Data
- **Hearing Statistics –** Data from Canada's federal energy regulator from 2007-2018, The National Energy Board. The dataset provides information about the approval process for large and small oil and gas pipeline and facility (e.g. processing plants, etc.) and some international power line applications. The data give information on dates of significant stages of the regulatory approval process, decision types, outcomes, environmental assessment, participants and the complexity of the decision. There are links to applicable hearing folders for ease of access to download or view documents. Data dictionary included within the excel file.
- **Letters of Comment –** 8500 letters of comment pertaining 51 NEB energy infrastructure applications, taken from NEB regulatory document index.

- **Factiva News Files** – Canadian News searches from Factiva global news database. Search files in .txt and .htm formats. Factiva Database News Search Notes outline the search terms used
- **Federal Election Results** – Data on election results for every riding in Canada from 2006 to our most recent federal election (2015) with political party affiliation of every elected Member of Parliament. Data comes from the House of Commons website. Data dictionary and notes included in the excel file.
- **Census Data** – Census data from Statistics Canada for 2011 and 2016, in several formats and sorted by: i) Federal Electoral District, or ii) Forward Sorting Area (first 3 digits of postal code). The data include population and dwelling counts, age characteristics, marital status, family characteristics, household and dwelling characteristics, language, etc. Data Dictionaries are included in the excel files. Metadata and other explanations are provided in metadata files from the Statistics Canada download.
- **Media Requests** – Data from Canada's federal energy regulator from 2006-2018, The National Energy Board. The dataset (largely text) provides information requests to the regulator from media. The data give the date of the request, the media outlet, the request, the information provided. More recent data entries include a general theme or topic of the request. Names of reporters and NEB staff removed for privacy reasons.
- **Social Media** – Twitter data pulled from 9 hashtags for current major pipeline projects and "pipeline" from mid-August. The Twitter free API provides data for 10 days, therefore regular downloads were created up until the datathon event.

- # Tools and Technologies

- Power BI
- Python
- R
- Looker
- Google Big Query
- Excel
- Power Query
- Azure platform (Power BI is part of the platform)
- SAS
- Tableau
- QGIS

- # Analysis Techniques
- **Power BI** – clean, merge and visualize data, creating dashboards
- **Python** – programming language for data analytics and web-scraping data
- **R** - programming language for data analytics. Sentiment analysis of text performed using this language (also using the Syuzhet package within R)
- **Looker** – to transform 3 million rows+ of census data for digestion in a useable format
- **Google Big Query** – the database where census data was held for Looker use
- **Excel** – some data collected within Excel
- **Power Query** – used to append and modify multiple excel and csv files
- **SAS** – for formatting data for use with Python
- **Tableau** – for dashboarding and visualization of data
- **QGIS** – for geospatial analysis and visualization

Key Learnings and Takeaways

Sentiment analysis of Twitter data for TransMountain project

- Tweets suggest positivity around the delay of the pipeline
- Traditional news media Tweets are not neutral on the TransMountain issue
- Context is important for sentiment analysis!
- Lexicon is important – understand what words/terms have been programmed to reflect "positivity" within sentiment analysis program. Oil & gas needs a sentiment dictionary (e.g. "crude" is a neutral term)

Media Inquiries & Hearing Statistics

To link hearing statistics with media data, common keys need to be agreed upon (e.g. date, location) to enable the merging of data for comparative analysis. The data model is the key! Once this was established, a great PowerBI dashboard proof of concept was possible. All of the goodness comes at the end (20 minutes of fun after 12 hours of cleaning).

Factiva News Data

When you work with unstructured data, it is best to get as close to the raw source as possible (e.g. scraping directly from the website via API). The Factiva searches were essentially text copies of the web page and valuable metadata was lost. In addition, we found the text copies of articles were not consistently formatted by the Factiva search and save, making algorithm programming for data extraction too problematic. Manual cleaning of attributes was very time-consuming, therefore only a subset of the data was utilized in the PowerBI proof of concept.

Geospatial Analysis of Hearing Data and Census Data

We were able to spatially join the project pipelines with the electoral districts of Canada. From this, we were able to understand the difference in census metrics between electoral districts with approved pipelines, pipeline projects in limbo, and districts with no pipeline projects. Projects that end up in limbo are, on average: in areas with higher unemployment, lower average and median income, higher post-secondary graduates. The data reveals a correlation between the status of pipeline projects with occupation of Canadians along the pipeline corridor. We found that for approved pipeline projects, there were higher rates of employment in the mining and extraction sectors. For unapproved/in limbo pipelines, these regions had lower rates of employment in the mining and extraction sectors.

Letters of Comment

We do know that a subset of data is underrepresented in the analysis of letters of comment, due to formatting (e.g. handwritten letters, and image-based letters). We recognize that, at the time, computer sentiment analysis was likely not considered as an output of the filing. Web forms would facilitate separation of stakeholders' comments and other data for better analysis.