**Regis University**

**Sentiments Toward Energy Sources and Sustainable Technologies:**

**A Twitter Analysis**

**Jameson Schryer**

**MSDS 696: Data Science Practicum II**

**Professor John Koenig**

**1 May 2022**

**Background**

Worldwide, nations are facing aging electrical infrastructure. The innovations in the past two decades have paved the way for a transformation in the way energy is generated, distributed, and consumed. Historically, the electrical distribution model has operated in providing electricity from centralized powerplants to consumers (Black&Veatch, 2020). The advent of solar panels and battery storage have flipped this model in some countries and will continue to do so worldwide. Solar and battery technologies allow consumers to generate their own energy for use and provide potential avenues for them to sell energy back through the grid. This model of operation is known as Distributed Generation (EPA, n.d). This is mutually beneficial for consumers and utility providers. Customers may lower their utility bills, and utility providers receive savings on the reduced need for infrastructure projects and generation.

Utility providers are also challenging traditional energy sources and are investing in renewable energy. Though many of these changes have occurred due to regulatory reasons, investing in renewables and diversifying energy portfolios can aide in operating cost for utility providers. Energy supply diversity also leads to political independence, as countries become less dependent on regions exporting hydrocarbons. The most notable benefit of renewables is the reduction in air pollution at local to national levels (Vakulchuk, 2020).

In the past decade, we have seen many advancements in technologies that help cut down fossil fuel use. Electric vehicles have become more accessible to the public…. Heat pumps and smart thermostats are now available for consumers. Utility companies are now providing rebates to customers who participate in programs using these technologies. Rebates are also available to those who participate in demand side management, where consumers opt in to occasionally seed control of their thermostat to off-set overloaded systems (SWEEP, 2020).

Another societal shift that has occurred in the past decade is the rise of social media. Social media has become the new public square, allowing users to express opinions in a way like never before. One major platform is Twitter, where users can express their thoughts in bite sized 280-character blurbs. As of May 2020, there were roughly 6,000 tweets being posted per second (Sayce, 2020). Twitter can provide a robust dataset of text when accessed via an API or scraped using a package. Data collected from Twitter has been used to model natural disasters, pandemics, stock fluctuations, etc (O’Connor, 2010). Although Twitter is far from being representative of public opinion as a whole, it can provide insights into public sentiment on a variety of topics.

**Research Question**

With this rapidly changing energy landscape, how have consumer’s opinions changed on energy related topics and technologies in the past decade? Can changes in these opinions be visible on Twitter?

**Data**

The purpose of this project was to collect tweets from Twitter regarding key words related to energy, analyze the data, and observe any trends made apparent. A list of energy related topics was compiled from several sources (Stanford Univ., n.d.). As the project progressed, it was clear that some of the topics originally chosen did not have enough data to be used. These were dropped, and the years of interest were narrowed down to every other year for the past decade. The topics this project focused on were: Electric Vehicles, Energy Rebates, Clean Energy, Renewables, Solar Panels, Geothermal, Hydro Electricity, Coal Power, and Natural Gas. The years focused on were: 2011, 2013, 2015, 2017, 2019, 2021.

Originally, the data was going to be scraped from Twitter using the official API and a developer license. It was found that the API only provided tweets from the past two years. The Python package snscrape was used instead, due to it’s ability to scrape all of Twitter’s history, as well as it’s intuitive commands.

Several packages were tested for calculating sentiment scores. A robust data cleaning section of the script was developed. After several tests, the vader sentiment analyzer for Python was chosen. This package was chosen for its unique aptitude for analyzing common phrases and lexicon, as opposed to a simple labeled list (Hutto, 2014). Because of it’s ability to handle phrases, the majority of the data cleaning regiment was dropped.

All data for this project was hosted in a Mongo Database. Mongo was specifically selected for it’s NOSQL architecture. Using a NOSQL solution was preferred, as this project required much storage space for the scraped tweets (MongoDB, 2022).

**Methodology**

To begin this Data Science Project, millions of tweets were needed to be scraped from Twitter. To accomplish this, a Python script called *Scrape All.py* was written. These packages were imported for the purposes of scraping and storing the tweets. The first step of this script established a connection to a local Mongo DB database, and assigned the connection to a variable.

Two lists were then written to lay out the years and topics to be explored. These were:

***year\_list = ['2011','2013','2015','2017','2019','2021']***

***topic\_list = ['NaturalGas','ElectricVehicles','CleanEnergy','Renewables', 'SolarPanels', 'Geothermal', 'EnergyRebates', 'HydroElectricity', 'CoalPower']***

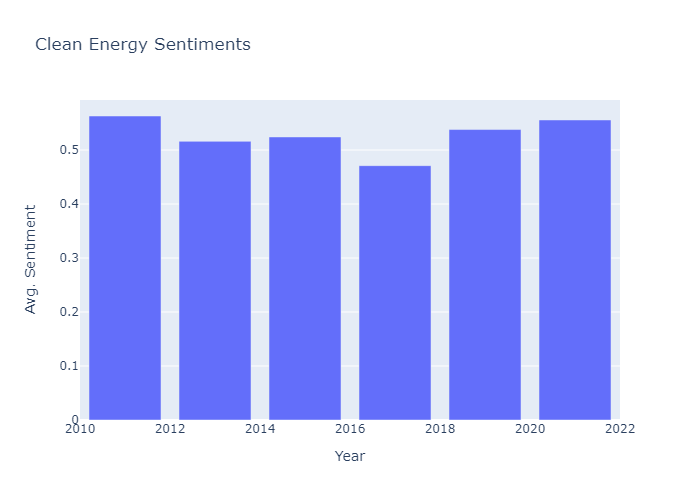
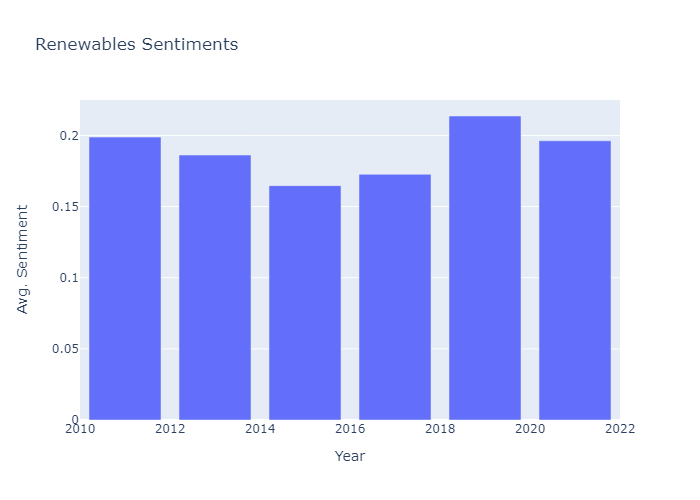
These lists were then referenced in a loop statement. The loop was set up to iterate up to 100,000 times for each topic and year combination, and dynamically scrape and append tweets to a Mongo DB collection for later use. As each topic was scraped, a new Mongo DB collection was created and named for the respective topic. The final product of this script was a Mongo DB database named *GreenEnergy&Efficiencies*. This database contained collections for each topic and tweet data within those collections.The fields pulled for this project were: the tweet date, tweet content, like count, user location, and user. To check if the tweets were scraped properly, a Jupyter notebook was written to access individual Mongo DB collections, convert them to a dataframe, and provide basic analytics on tweet counts. This notebook was called Explore Collections Individually.

The next phase of this project required a script to access the data in Mongo DB, clean it, perform a sentiment analysis, append the composite sentiment score to each tweet record, then re-upload the data into Mongo DB. This script was called *Append Sentiment.py*. To begin, a list of the collection names was written and a loop statement to reference it. The loop statement was written to iterate through the collection names and establish a connection to each collection. The data from the collection was then inserted into a pandas dataframe. The tweet text was then cleaned in preparation for sentiment analysis. The cleaning applied to this data was: digits, web addresses, and duplicates. The final product was a new field called *clean text.* After the data was processed, vaderSentiment SentimentIntensityAnalyzer was imported. The newly cleaned records in the *clean\_text* fieldwere ran through the sentiment analyzer, and a composite sentiment was appended to each record in the dataframe. With the composite sentiment scores appended, the dataframe was inserted into a new collection named after the previous collection, and a new Mongo DB called *GreenEnergy&Efficiency\_Sentiments*.

Next, a script was written to loop through the list of collections in the *GreenEnergy&Efficiency\_Sentiments* database and generatepdfs of visualizations depicting the mean sentiment value of tweets regarding each topic. This script was called *Visualization Loop.py.*  To begin, a new loop statement was written using the list of collections used in the previous script. A new connection was established for each collection name. The data from the collection was then inserted into a new pandas dataframe. The year of the tweet was extracted from the date field and was placed in a new field called *year*. Next, all records were grouped by year and the mean of all tweets from that year were calculated. The results were inserted into a new dataframe called *year\_df.* The *year\_df* dataframe was then used in a statement to create a bar plot in plotly, depicting the average sentiment value for each year. Lastly, visualizations for each topic were sent to a png and pdf in a local folder.

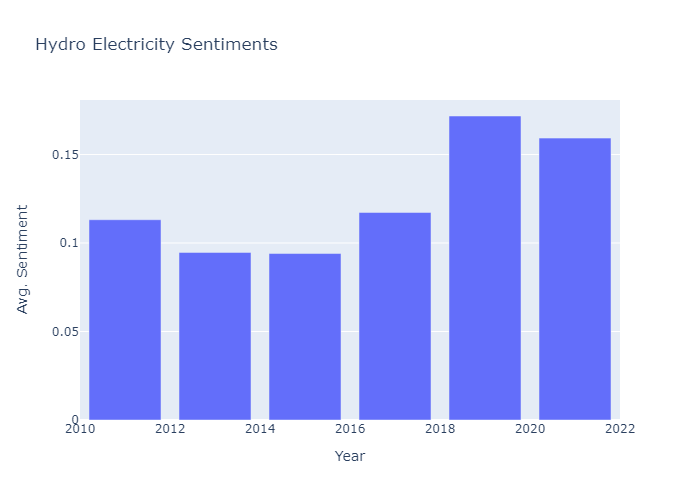
Lastly, to explore an anomaly in the results regarding values for Coal Power, a Jupyter notebook called Explore\_Coal Texts was written. This notebook explored the geography and user count of Tweets.

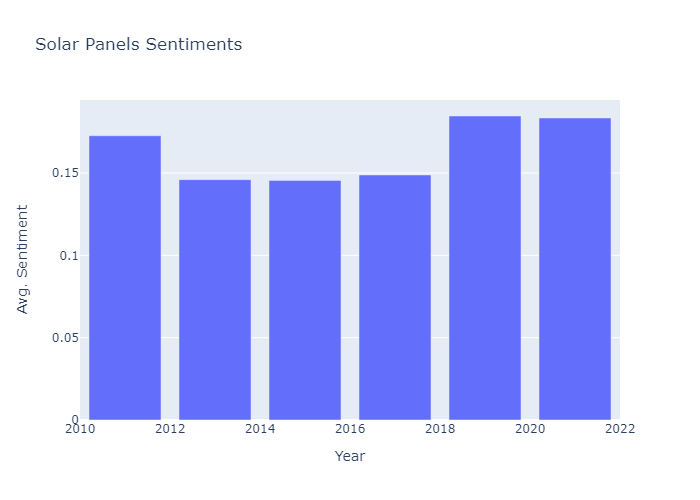
**Results**

****

**Fig. 2** Sentimentstoward Renewables lowered in the middle of the decade, but rose is 2019-21

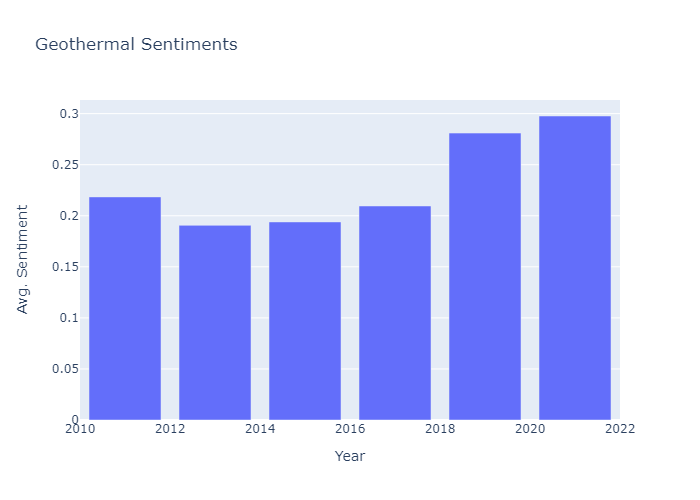
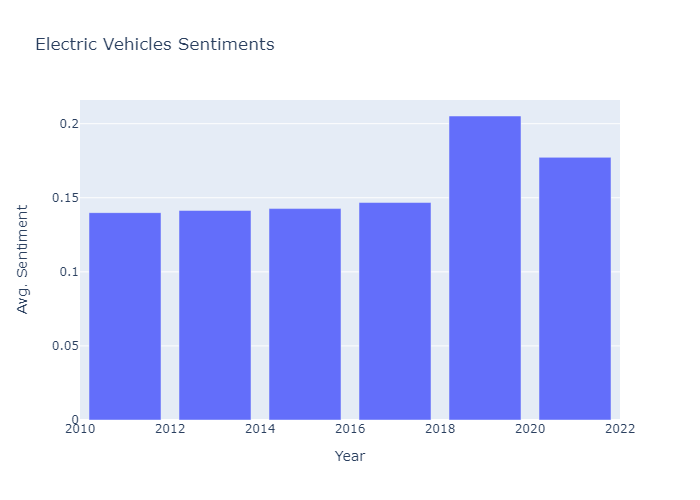
**Fig. 1** Sentimentstoward Clean Energy have remained stable, with a small dip in 2017

****

****

**Fig. 4** Sentimentstoward Hydro Electricity were low at the start of the decade and significantly improved in 2019-21

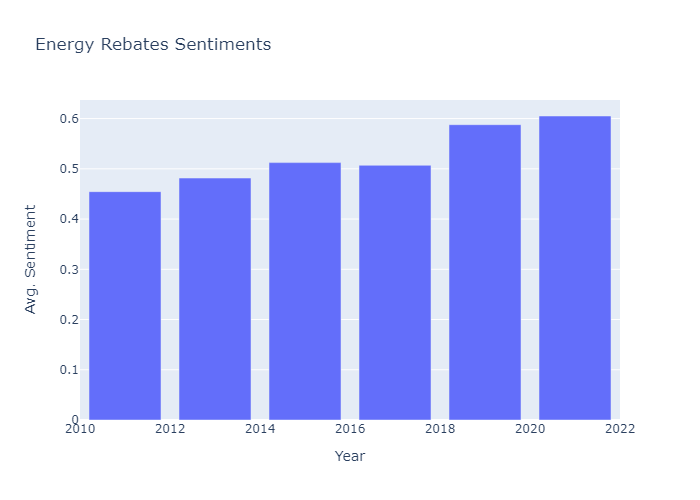
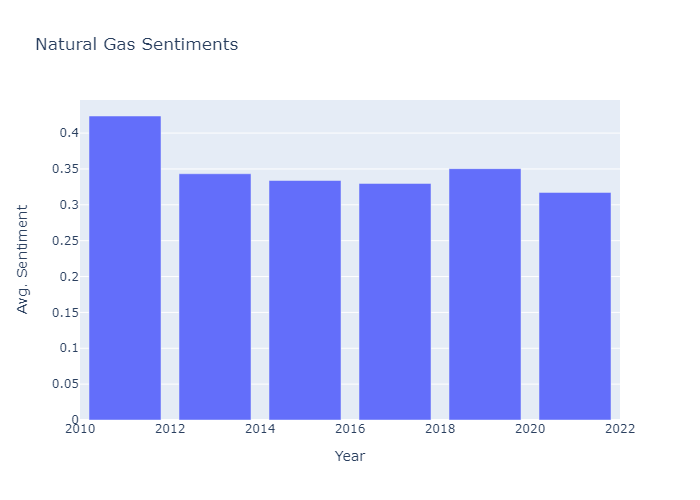
**Fig. 3** Sentimentstoward Solar Panels lowered in the middle of the decade, but rose is 2019-21

****

**Fig. 6** Sentimentstoward Electric Vehicles remained stable early in the decade, but rose in 2019-21

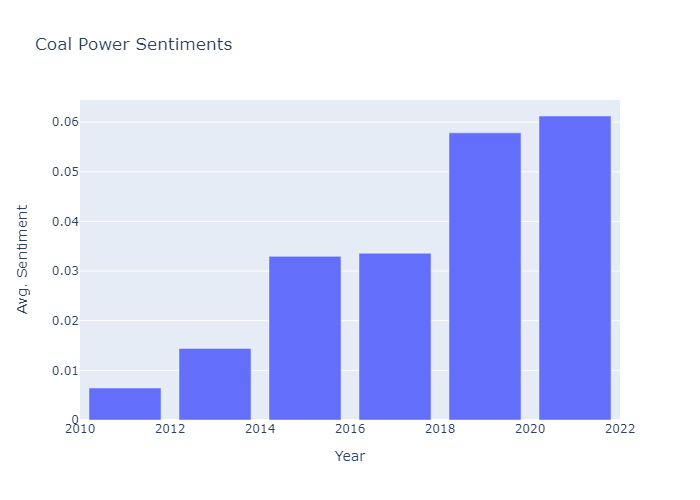
**Fig. 5** Sentimentstoward Geothermal Energy started low early in the decade, but steadily rose

.12310………………..

****

**Fig. 8** Sentimentstoward Natural Gas had an initial drop, then remained stable

**Fig. 7** Sentimentstoward Energy Rebates steadily increased over the decade

****

**Fig. 9** Sentimentstoward Coal Power have dramatically risen over the past decade

**Conclusions**

Through the methodology set forth in this project, sentiment changes can be observed over the past decade for a host of energy related topics. Many of these trends match the expectations of increasing or staying neutral. However, Coal Power trended in an unexpected positive fashion. After further exploration into the accounts with the tweets with the highest sentiments toward Coal Power, it was found that many of the tweets were in fact anti-coal but used positive language. This indicates there is much more work that can be done to improve this project.

If this project was to continue, several changes would be implemented. Tweets belonging to governments, companies, and organizations would be removed. Time would be invested to develop a custom sentiment classification model. The results of this model would undergo rigorous accuracy testing. There would be further exploration of the correlation between sentiments and likes. Lastly, there would be an exploration between tweet location and sentiments on various topics.

**Works Cited**

Black&Veatch. (2020). *Electric Report: Repowering the Power Industry*. Retrieved April 2, 2022, from https://www.bv.com/sites/default/files/2020-10/20%20SDR%20Electric%20PDF%20FINAL.pdf

EPA. (n.d.). *Distributed Generation of Electricity and its Environmental Impacts* . EPA. Retrieved April 10, 2022, from https://www.epa.gov/energy/distributed-generation-electricity-and-its-environmental-impacts

Vakulchuk, R., Overland, I., & Scholten, D. (2020). Renewable energy and geopolitics: A review. *Renewable and Sustainable Energy Reviews*, *Volume 122*.

SWEEP. (2020). *DSM Policy & Programs Overview*. Southwest Energy Efficiency Project. Retrieved April 16, 2022, from <https://www.swenergy.org/utilities/dsm>

Sayce, D. (2020, December 16). *The number of tweets per day in 2020*. David Sayce. Retrieved April 16, 2022, from https://www.dsayce.com/social-media/tweets-day/

O’Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). *AAAI Publications*. Fourth International AAAI Conference on Weblogs and Social Media. Retrieved April 10, 2022, from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10

Stanford Univ. (n.d.). *Research areas*. Energy. Retrieved April 10, 2022, from https://energy.stanford.edu/research/research-areas

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225. Retrieved April 10, 2022, from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>

MongoDB. (2022). *What is nosql? NoSQL databases explained*. MongoDB. Retrieved April 16, 2022, from https://www.mongodb.com/nosql-explained