Earvin Sahayathasan

500709422

April 23, 2020

Twitter Sentiments on Nuclear Energy / Power

Introduction

The nuclear energy business is surrounded by confusion which people are polled towards pro or against this form of energy. The opinion of the people in a community typically determine the longevity of a business and if one side is much more heavily favored compared to others, governments and major corporations are forced to change their systems. This practice is also applied to nuclear plants with some nuclear plants being shutdown due to public opinion in the masses. Although, these cases exist the public view is still heavily divided, providing evidence to both pro and against using nuclear energy. In a study from the National Public Opinion Survey on Nuclear Energy, 65% of U.S public expressed support for nuclear energy.

Literature Review

Article about twitter more specifically demographics on who uses it:

https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Sizing Up Twitter Users

According to the article, Sizing UP Twitter Users by the Pew Research, the current demographics of the twitter community were taken for last year in the US. These statistics will give an understanding of who is commonly using the application and what percentage of people are likely to be speaking about these issues online. Last year, approximately 10% of users had nearly 80% of the tweets for adult users stated by the article. Thus majority of active users have low tweets posted but still use to read others tweets, predominantly the 10% tweeters. This heavily affects twitter opinions on debates because the more traction a user can provide the greater impact in society it affects. The average age of a twitter user is very well spread with the biggest age group, 30 – 49 having nearly half the activity on the site. The gender is almost split evenly where the low active female users have a 2 percent edge compared to the male low active users. Majority of the twitter users that are active have a high school diploma with 50 to 59 % of users and the major ethnicity group using the site is White who are around 60 – 64 % of the active users. These statistics help understand the twitter community and who will tweet about these topics by knowing about their general backgrounds and age groups.

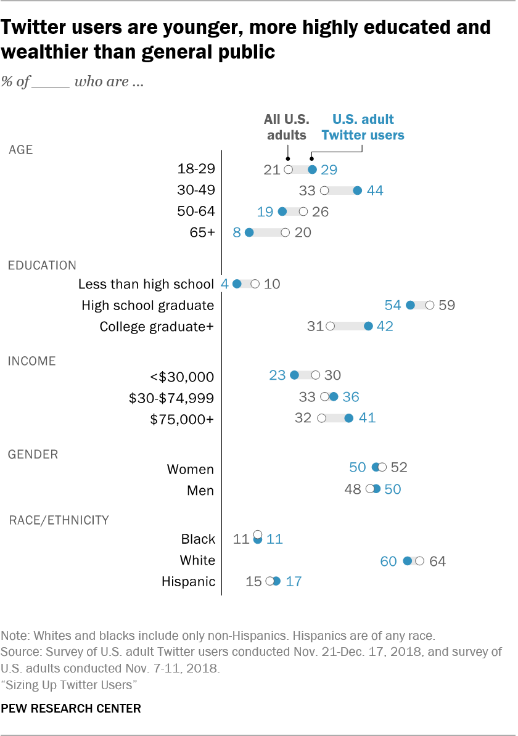


Figure 1: Demographics of Twitter Users by Age, Education, Gender and Race

Article about nuclear energy and what it is:

https://www.nei.org/fundamentals/what-is-nuclear-energy, What Is Nuclear Energy?

The objective of the project is to understand the position on nuclear energy and to do such a thing a certain knowledge of nuclear energy must be understood. The article, What is Nuclear Energy, says that nuclear energy is retrieved from splitting atoms in a reactor to heat water into steam which turns a turbine and generates electricity. The standard states how currently US citizens run on electricity that 20 % of all the energy comes from nuclear energy, this implies how powerful the energy source is comparative to the amount of resources it needs. The nuclear energy system uses elements to perform the atom splitting and the element commonly used is uranium. The uranium is not carbon emitting thus it does not provide the negative effects that of which a fossil fuel does directly to the environment. However, uranium and nuclear energy are still not the best eco-friendly solution because other waste is provided from these nuclear plants.

Article about positives of nuclear energy:

Power to Save the World: The Truth about Nuclear Energy by Gwyneth Cravens

The nuclear power ideology has been given a negative view in the past few decades by a fair amount of people. However, there are activists today that have proof and statistics stating that the problem is not the energy or power plants but rather the incompetence. In fact, the book, Power to Save the World: The Truth about Nuclear Energy, states how not only are nuclear energy not harmful but rather helpful and promotes an eco-friendly option for energy resources. The book states how power plants have grown in popularity in recent ages due to the immense energy that is produced from a single location which is more than enough to energize the cities surrounding it. The energy produced will produce leakage but if disposed properly will not have as major of an impact if mishandled. The nuclear energy trend has surfaced internationally where in France, 80% of energy is nuclear electricity. Japan is also changing its resources towards nuclear energy while seeing the benefits as well, and in the USA the booming increase in nuclear reactors and nuclear power is growing till this day.

Article about negatives of nuclear energy:

The Pros and Cons of Nuclear Power by Ewan McLeish

The nuclear power and energy is more notable for there negative impacts to society. The major problems that people bring up is that the energy hurts the environment, is harmful to people and can massively affect nearby citizens when near power plants. The first issue that was discussed in the book, The Pros and Cons of Nuclear Power, is the ecosystem harm it causes, where when handled properly still has affects to the environment and when handled poorly which is fairly common could result in a grave harm to the nearby ecosystem. The leakage that occurs from nuclear power plants happens fairly often leading to shutdowns of entire plants due to these incompetence’s. Typically, workers that are exposed to the nuclear energy are wearing full portative suits and respirators to protect from breathing in these harmful chemicals but when leaked people are inhaling these directly causing major incidents, some fatal. Thus, the growth of the nuclear power grew with the protest against it because the more that existed the more hate it received.

Article about people most impacted by nuclear energy:

Values, Perceived Risks and Benefits, and Acceptability of Nuclear Energy by Judith I. M. de Groot

The impact that nuclear energy has to people in their neighbourhoods and past historic events alter people’s opinions on the usage of it. The article states how depending on the mindset of individuals they trend toward a certain opinion on nuclear energy. For example, people that are pro nuclear energy tend to have and egoistical belief where they typically look at the positives that the energy has for them and how it could make life easier. This correlates to the previous positives to nuclear energy since it provides a significant amount of energy from a single plant. The negative people typically are more ecological and altruistic because they believe that the harms that are caused to the environment are too impactful and cannot be beneficial. The people that do not believe in the nuclear energy are typically fixed on their position on the topic as well even if evidence can be provided to show positives in the environment and people’s well being. The reason people are firm to their beliefs is due to the past, which in the article acknowledges the harm that is caused from harnessing the nuclear energy. The most notorious act of human cruelty was the bombing in Japans using the nuclear energy to flatten the city and leave high remnants of radioactivity causing even more fatalities. Thus, currently people’s positions on nuclear energy revolve around these ideologies.

Article about keywords and selecting them:

Meertens, R. W., & Stallen, P. J. (1981). Opinions on nuclear energy: Evaluations, beliefs, and attitudes. Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 36(1), 45–55.

The typical topics of discussion when trying to extract information from text and messages is to know what are the keywords or ideals that cover the main categories for the discussion. In the case of nuclear energy, the article goes on to mention what are the main constructs of the nuclear energy discussion. The typical topics are the society/societal attributes, technical and economic attributes of nuclear energy for the positive opinions or the pro-nuclear energy believers. The topics for those that oppose typically revolve around bio-friendly, environmental and human rights when discussing why they think it should not be used. The arguments and words may be spoken eloquently or simply as possible but the thought behind people’s opinions typically are rooted in any of these topics. When looking for keywords especially when to group individuals, it will be helpful to correlate their sayings with what their stance on the topic is.

Article about Text Classification:

https://dl.acm.org/doi/abs/10.1145/1835449.1835643

Short text classification in twitter to improve information filtering

Text classification is a tool used to classify a message, an opinion or any textual, non-numeric dataset into classifications based on the conditions predefined by the user. In the case for the twitter messages, the classification would be the question to classify if a user opposes or upholds the arguments against nuclear energy. This can only be done using text classification because all the arguments being made are in text and will have to classify them into one of the three groups which are agree, disagree or not applicable. Text classification will determine using the keywords to deduce what tweets classify an individual into whichever category. In the article, a user had performed text classification using twitter as well and had gone about it by taking short texts from the massive raw dataset and classifying each segment based on the conditions provided. This was done because the more text that is provided into a single classification, the more variables are add in play which may defer the actual result rather than doing separate classifications and combining the results to see which result is heavily favored and that will be the classification. Twitter themselves use this approach when labeling tweets in categories such as News, Events Opinions and so on.

Article about Predictive Analysis:

https://www.jstor.org/stable/23042796?seq=1

Predictive Analytics in Information Systems Research

Predictive analysis essentially is predicting a future result based on the previous results and analytical behaviour. In the article it mentions how that predictive analysis uses statistical and other methods that generate data predictions as well as methods for future predictions. This means that not only does predictive analysis provide results but it can also provide a means of getting continual future results as long as the conditions for the situation do not change. This flexibility and applicability is why many applications and large companies invest into predictive analysis, to see their longevity. The variety of applicability for predictive analysis makes this a universal field because it can be used to proves theories and test models, test experiments, predict outcomes, show timelines and solve important matters depending on the usage. Thus, when applying the proper result causes more influence in the field as well. However, since it is a prediction the results are only a high probability of what occurs and is not necessarily a guaranteed result because the one thing that cannot be predicted are radical, instantaneous changes that alter prior conditions in a set. Although, taking all this into account predictive analysis will be used to help find out what type of individuals are voting for which side based on the results that are found from the text classification.

Article about dataset cleaning or outliers:

https://academic.oup.com/nar/article/24/2/316/2359286

Cleaning the GenBank Arabidopsis Thaliana Data Set

Dataset cleaning is a tool used for predictive analysis and classifying information before the methodology and analysis on a dataset. The tool essentially removes any irrelevant data that could sway the results from the necessary outcome. Dataset cleaning is not simply removing the data that is not correlated or does not have the necessary requirements for the problem being tackled but rather taking the least amount of data possible to result in the most accurate result. Not only does this help the problem solving but it also makes efficiency much higher because less data being run through makes the methodology move faster. In an article that used the dataset cleaning on a similar topic which run through text classification similar to this twitter problem. When cleaning the data, they checked for irrelevancy first removing 15 % of the data provided that did not satisfy the proper results. The Next step was to remove the data that conflicts each other meaning they had the same role, the same outcome and thus is the same data. When this occurs, the data becomes redundant and with datasets its is ideal to have alternative splicing according to the article. However, some data may be corrupt by faulty transferring or misinformation so when performing these corrections it causes a more reliable dataset to provide a proper outcome.

Article about influence on opinion based on status specifically twitter status:

https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Sizing Up Twitter Users

https://dl.acm.org/doi/abs/10.1145/1935826.1935845

Everyone's an influencer: quantifying influence on twitter

The point of tracking influence from twitter followers is because when a large personality voices an opinion, it causes a huge change in how votes and events can occur. So, when an influencer of some sort voices their stance on nuclear energy the masses will react to the influencer’s opinions. As mentioned from the previous article only 10% of people use twitter avidly on the site. Thus those 10% have more of an influence on people’s opinions. Thus, when a major influencer campaigns for a side on a topic, typically the influencer will gain more followers agreeing with their opinion, tipping the scale in their favour. However, this causes a problem in itself because when people do not completely understand their stance on a topic it causes backlash if it was wrong. In another article, it stated that in a 2-month interval in 2009, influencer with over 1.6 million followers had 74 million diffusion events that occurred. The result became conclusive that all the activity being produced was from major influencers with a huge following. When taking analytics into account there were instances where opinions based on influence matters since a person with no influence and social interaction cannot have the equivalent impact compared to a major influencer. Thus, when choosing attributes of importance people with higher influence should be a major factor alongside key terms for the text classification.

Dataset

The data source is called: Judge emotions about nuclear energy from twitter. This dataset was from an online open source website: <https://www.figure-eight.com/data-for-everyone/>. The methodology and approach on how the dataset will be used is shown below through the Approach. The dataset contains headers describing the contents below which are the tweet, their stance on the topic whether it be negative or positive referred to as sentiment and finally the confidence summary which is a tab on whether the tweet feigns any relevance to the topic at hand. Typically, three categories is not enough but since this is a poll and a predictive test the words used in the tweets are plentiful of information to deduce whether a person is towards or against nuclear energy. The tweets range from 20 characters to up to 120 characters which is the max number of characters in a tweet. This allows for some manual constraint without having to force a limit myself. The sentiment tab shows not only negative and positive but information as well. This means that information or neutral stances will have to be taken into account amongst the choices. This is how the confidence summary and the sentiment tab correlate since people that do not have an opinion may be discarded but since there is a clear indicator in which someone has relevance using the confidence summary it will be easy to remove the outliers. This is what the dataset offers and this is how it will be used to perform actions.

Approach

Remove outliers from the dataset that are irrelevant to the comparison, remove further outliers that are redundant and can be assimilated with other data present to make the process faster

Then using Text Classification, pull out certain words that show the most prominent affect to the Classification of either Negative, Positive or Neutral public opinion on nuclear energy

From the words that were classified predict the words that will be tweeted to show consistency with the nuclear energy stance, post predictive analysis to test upon a random part of the dataset using either 10-fold or test vs trial data sets

Conclusively answer the questions that pertain to the research questions that are given.

Step 1: Remove outliers from the dataset that are irrelevant to the comparison remove further outliers that are redundant and can be assimilated with other data present to make the process faster

To remove the outliers the confidence summary will be used initially and thus can easily be done to remove the clear-cut outliers. The confidence summary will remove any sentiment that is either not related to nuclear or is an “I can’t tell” indicator.

The next step is to use the outlier test to look from specific conditions such as:

* Tweet is in English
* Tweet passes the confidence summary test
* Tweets that are exactly the same will be removed

This cleaning process allows the results to not be swayed or compromised for a specific faction.

This was achieved by performing pre-screening and outlier checks using Weka which is an open-source application that performs cleaning of data that is repeated to ease the cleaning process initially. This was shown when the data that was taken which was shown had been significantly decreased from the one that had been changed. Both datasets have been uploaded to the GitHub respectively for analysis. The technique that was used was to remove correlating lines which either were retweets or simply did not meet the confidence summary.

After the initial screening, the data had been cleaned again in the R code shown in the lab as well. The dataset now has raw data tweets that were unreadable and cluttered. To fix this the R corpus technique was used. The data had been cleaned and removed of all data punctuation and spacing issues leaving just the plain text to be text mined and sentiment analysed for the full classification.

Step 2: Use Text Classification, pull out certain words that show the most prominent affect to the Classification of either Negative, Positive or Neutral public opinion on nuclear energy

When performing Text Classification, the tweets must be oriented with their corresponding result meaning that whatever sentiment corresponds with the tweet that is what will get upvoted on the classification. The tally will show the entire scoring of Negative, Positive and Neutral from the entire dataset. Although not only do the tweets need to assigned a classification, the tweets must be checked for correlation so that the system may not be biased towards an opinion.

This was essentially performed using the Text Mining and Sentiment analysis code. To find the words the data was first cleaned from before using the Text Mining code and then using a function in R called tdm, the dataset was fully capable of being directly text mined. The text mining had already had the sentiment correlated in the dataset making the check for post analysis simple. However, to achieve sentiment the words needed to be found that had the most impact in the dataset. This was done using the wordcloud method which scoured the dataset of all the individual words that were not considered articles of English to find what resonated commonly amongst the tweets. Below are some tables and graphs showing the wordcloud and tdm functions results.

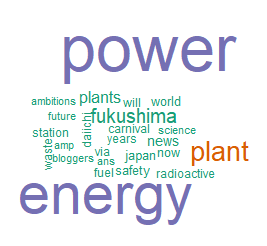


Figure 2: Result of Wordcloud for Frequency of Usages in Dataset Correlating to Sentiment

|  |
| --- |
|  |
|  |
|  |

Figure 3: The R Table which showed Sentiment Scores for each Tweet in the Dataset

Step 3: From the words that were classified predict the words that will be tweeted to show consistency with the nuclear energy stance

From the classifications, extract the words that are commonly associated with each of the classifiers and use them for a general tool to determine the classifications of future tweets. This will be done by using the classification method mentioned above where segments of the tweets will be further broken down till a consistent trend towards conclusive stances from words can be defined. From these words predictive analysis is being performed and will be tested by running the dataset though various tests both trial datasets and 10-fold.

The testing would have liked to be done on a different dataset to perform different results however the results would have varied so what was done was to use the 10 fold analysis on the dataset and take a portion of the dataset and correlate the results with the sentiment that had been assigned to the data originally in its entirety. To find the test dataset, it will be disclosed in the GitHub as well.

Step 4: Conclusively answer the questions that pertain to the research questions that are given

From the data accumulated till this point, the data should provide a conclusive answer to the problem at hand. Thus, a report encapsulating the entire process and results will be shared via GitHub. The git hub is currently open for inspection however the project is in the early stages so limited amount of code is updated.

1. What are the most common words to provoke positive opinion? Negative opinion?

Most Common Negative Word was Fukushima. This makes sense since the logic is that each statement that corresponded to this was negative connotations. The messages typically were opposing of nuclear power and would never support the usage seeing the destruction.

Most Common Positive Word was Power. This implies that the positive tweets were about using the power source for good and utilizing it for the better of people. This also shows how the importance of energy makes such a volatile power needed because of how strong and impressive the power is.

There were other words that associated with negative and positive statements but these were the main ones.

1. What problems are shown to most impact negative opinion?

The words that negatively impact the opinion typically originate around the past and advert affects from nuclear energy. The glaring word was Japan which when correlated to nuclear energy and power typically was negative. This was due to the fact that of the mass bombing that had occurred in the attack on Hiroshima that utilized nuclear energy in a negative sentiment. In the literature review, it was stated how past events typically have a longing impact in current day results which truly is relayed with these sentiments from these tweets.

1. Is there any tweets that do not pertain to nuclear energy such as outliers that will not help with the general consensus?

In the early stages of the process, there were various unrelated outliers that were recognized in the dataset by having no sentiment. When the data retrieved a N/A data in R. The data must be nulled and thus removed from the dataset which occurred in the second screening of the cleaning data stage. However, the Weka application removed most null applicants prior to the actual performance of the text mining. Some messages that were unrelated in the dataset were tweets that had no substance or context to the topic at hand and thus deemed null due to irrelevance to the topic.

Github Link: <https://github.com/earvinsahayam/CKME-136>

Bibiography

1. American Psychological Association, American Psychological Association, psycnet.apa.org/record/1981-33788-001.
2. Bakshy, Eytan, et al. “Everyone's an Influencer: Quantifying Influence on Twitter.” Everyone's an Influencer | Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 1 Feb. 2011, dl.acm.org/doi/abs/10.1145/1935826.1935845.
3. Cravens, Gwyneth. Power to Save the World: the Truth about Nuclear Energy. Alfred A. Knopf, 2008.
4. G., Peter, et al. “Cleaning the GenBank Arabidopsis Thaliana Data Set.” OUP Academic, Oxford University Press, 1 Jan. 1996, academic.oup.com/nar/article/24/2/316/2359286
5. Groot, Judith I. M. De, et al. “Values, Perceived Risks and Benefits, and Acceptability of Nuclear Energy.” Risk Analysis, vol. 33, no. 2, 2012, pp. 307–317., doi:10.1111/j.1539-6924.2012.01845.x.
6. McLeish, Ewan. The Pros and Cons of Nuclear Power. Rosen Central, 2008.
7. Shmueli, Galit, and O. Koppius. “Predictive Analytics in Information Systems Research.” SSRN Electronic Journal, 2010, doi:10.2139/ssrn.1606674.
8. Sriram, Bharath, et al. “Short Text Classification in Twitter to Improve Information Filtering.” Short Text Classification in Twitter to Improve Information Filtering | Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1 July 2010, dl.acm.org/doi/abs/10.1145/1835449.1835643
9. “What Is Nuclear Energy?” Nuclear Energy Institute, [www.nei.org/fundamentals/what-is-nuclear-energy](http://www.nei.org/fundamentals/what-is-nuclear-energy).
10. Wojcik, Stefan, and Adam Hughes. “How Twitter Users Compare to the General Public.” Pew Research Center: Internet, Science & Tech, Pew Research Center, 2 Jan. 2020, www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.