Characterizing nuclear energy conversations on Twitter

Sean Deloddere
sean.deloddere@aalto.fi

Tutor: Aqdas Malik

Abstract

Nuclear energy could play an important role in containing global warming. However, it is a controversial topic. Public opinion has an impact on technological potential and policy. While there has been prior research on the public opinion on nuclear energy before, there has been no research of this kind using Twitter yet. The use of social media platforms as data sources for research has increased recently; however, as a whole, it is still fairly new. The goal of this paper is to use data gathered from Twitter to characterize nuclear energy conversations. The data is gathered through the Twitter API and pre-processed. Using topic-modelling the most prevalent topics are gathered. Sentiment analysis is performed to analyse the sentiment of the conversations. 5 topics were extracted from the tweets, 'Energy production and climate' was the largest topic. Sentiment was slightly positive and slightly objective.

KEYWORDS: nuclear energy, public opinion, Topic Modelling, Sentiment Analysis, NLP, Twitter

1 Introduction

With 188 parties currently partaking in the Paris Climate Agreement, showing their commitment to containing global warming, nuclear energy could come to play a significant role in the future [1]. Nuclear energy is a clean energy source, meaning it has low CO_2 emissions [2]. In 2018, nuclear energy produced 5% of global energy, more than hydro, wind and solar combined, making it the largest clean energy source available [3]. This could make it an ideal candidate for countries attempting to shift their energy productions towards lower carbon emission alternatives. Nevertheless, nuclear energy has been a polarizing topic since its conception [4]. From safety concerns, its link with nuclear weapons and the issue of radioactive waste, to its high energy returned on investment (EROI) and environmental impact, nuclear energy has been a point of discussion for decades now.

Furthermore, it is not only experts having these discussions, regular people often do too. Unfortunately, common knowledge about nuclear energy is not always accurate. When asked, around 70% of students in a British survey responded that they thought less use of nuclear power would reduce global warming [5]. Nuclear energy is also generally seen as risky and dangerous, referring to major accidents such as the ones in Chernobyl and Fukushima [6]. However, when inspecting deaths per Terawatt-hour, a metric to measure safety of energy production, nuclear energy accounts for a death rate of only 0.07, compared to , e.g., the 24.6 death rate of coal [7]. For reference, 1 Terawatt-hour is the annual energy consumption of 27.000 people in the EU. This makes nuclear energy one of the safest methods of energy production [8]. Other studies show that people worry about developing cancer due to their proximity to a nuclear power plant, or that nuclear scientists do not have enough knowledge to handle nuclear waste, while both claims have been disproved [9]. There is also no evidence that countries with nuclear energy programs are more likely to seek or acquire nuclear weapons [10].

Researching the public opinion on nuclear energy has been done before; however, this research was based on polling and surveys. These surveys usually gather the opinions of a few thousand people at most. The rise of social media in the last decade has opened up opportunities for research on a totally different scale. With its 330 million users worldwide, and an average of 500 million tweets, short messages which are limited to

280 characters, being published every day, Twitter serves as the perfect medium for research into conversations [11]. Due to this considerable amount of users, Twitter has become a very powerful tool to gain real-world insight. In recent years, much research has been done leveraging the large amount of data twitter provides as a tool to examine how people in general discuss certain topics [12, 13, 14, 15, 16, 17, 18]. This large amount of data can be automatically analysed effectively using natural language processing (NLP) techniques, which will be discussed later on in the paper. The machine learning techniques are perfect for these applications as they grow more accurate with more data input.

The goal of this research is to analyze the current public discourse on nuclear energy by defining the most prevalent topics being discussed in nuclear energy conversations and evaluating the sentiment polarity and subjectivity for each of these topics.

2 Methods

2.1 Data collection

In this phase, tweets were collected using Twitter's Application Programming Interface (API), Developer [19]. We used Python version 3.8.3 for both the data gathering as well as processing and analyzing [20]. We employed Tweepy, a library from Python, to interact with the API [21]. There are multiple methods to extract tweets from twitter, the API provides options to either collect historic tweets, that are already on the platform, or stream relevant new tweets. In this paper, the latter was used. From October 28 to November 29, 39400 tweets were extracted using 'nuclear power', 'nuclear energy', 'atomic power', 'atomic energy', 'fission power', 'nuclear fission' and 'thermonuclear energy' as relevant key words. We detected whether a tweet was a retweet or not, and did not save retweets to avoid duplicate bodies of text in the data set. Only English tweets were selected.

2.2 Data cleaning and pre-processing

To optimally prepare the data for NLP analysis, standard pre-processing techniques were performed [22]. Employing re, a Python library, the text was transformed to lower case, and hashtags, punctuation, apostrophes,

numeric values, quotation marks, newline characters, hyperlinks, subsequent spaces and spaces at the beginning of a tweet were removed [23]. The text was then tokenized; this is the process of splitting the text into smaller pieces, in this case into separate words. This was done utilizing Python library sklearn [24]. The result is a Document-Term Matrix with a bag of words for each tweet. This matrix was then transposed to obtain the Term-Document Matrix used in topic modelling. The most frequent words in the data set, as well as for each topic, were displayed using wordcloud and matplotlib, both Python libraries [25, 26].

2.3 Topic Modelling

The first form of processing performed on the data was topic modelling. This was conducted using latent Dirichlet allocation (LDA), a probabilistic model to organize collections of discrete data, e.g., text, based on latent topics [27]. Most similar research also uses LDA and it is proven to be effective [28]. Python library gensim was employed for performing LDA [29]. LDA requires a number of topics as input and returns that amount of lists of words with weights that are representative of a topic. The topic label itself is assigned based on these words and their weights. Several different models were trained with different inputs, including the complete Term-Document Matrix, only nouns, only nouns and adjectives, and only nouns and adjectives without the most common words. Selecting only nouns and adjectives from the Term-Document Matrix was achieved employing Python library nltk [30]. For each of these models a number of topics ranging from 2 to 20 was evaluated based on semantic similarity of the list of words and comparison of representative tweets for each topic. Ultimately the model trained on only nouns and adjectives without the most common words, 'nuclear', 'power' and 'energy', with input number of topics as 5 was chosen. This model was then employed to determine for each tweet the probabilities of it belonging to each topic. Tweets were classified as belonging to a certain topic according to the highest probability.

2.4 Sentiment Analysis

In order to get a grasp of how people felt about the topic of nuclear energy sentiment analysis was performed. Python library TextBlob was employed for this purpose [31]. We inspected the sentiment polarity and

subjectivity of each tweet. TextBlob determines the polarity and subjectivity utilizing the AFINN lexicon, in which each word has been assigned a polarity and subjectivity score for each of its meanings [32]. The mean of the values of each meaning is then returned, and the mean of the values of all the words in the tweet constitute the polarity and subjectivity values of the tweet itself. The polarity value ranges from -1.0, most negative, to 1.0, most positive. The subjectivity value ranges from 0.0, most objective, to 1.0, most subjective [33]. Polarity and subjectivity of the entire data set as well as of each topic independently were evaluated. Determining the polarity and subjectivity of each topic was done through assigning weights to each tweet within a topic based on the probability determined by the LDA model and calculating weighted means.

3 Results

3.1 Word Frequency

Before any processing, the 39400 tweets that were gathered contained 51144 unique words, out of 551479 words in total. The 200 most occurring words are illustrated in Fig. 1. After processing, the 20 most occurring words are 'nuclear' (31206 times), 'power' (20663 times), 'energy' (13365 times), 'plant' (3591 times), 'plants' (2692 times), 'new' (2403 times), 'just' (2101 times), 'like' (2064 times), 'need' (1942 times), 'atomic' (1860 times), 'wind' (1790 times), 'solar' (1714 times), 'waste' (1587 times), 'world' (1506 times), 'people' (1475 times), 'trump' (1459 times), 'green' (1417 times), 'going' (1336 times), 'years' (1321 times) and 'climate' (1300 times).

3.2 Topic Modelling

The words contributing to the topic model for each topic and their subjectively assigned label are displayed in Fig. 2. The largest topic that was detected was labeled 'Energy production and climate', and contained tweets that discussed nuclear energy in the context of climate change and alternative energy sources. The second largest topic, labeled 'Trump and international conflict', contained tweets referring to Trump's nuclear policy, the Iran nuclear deal and Pakistan's nuclear weapons. The third largest topic was labeled 'Power plant near nature reserves' and contained tweets mostly petitioning against power plants being built near nature re-

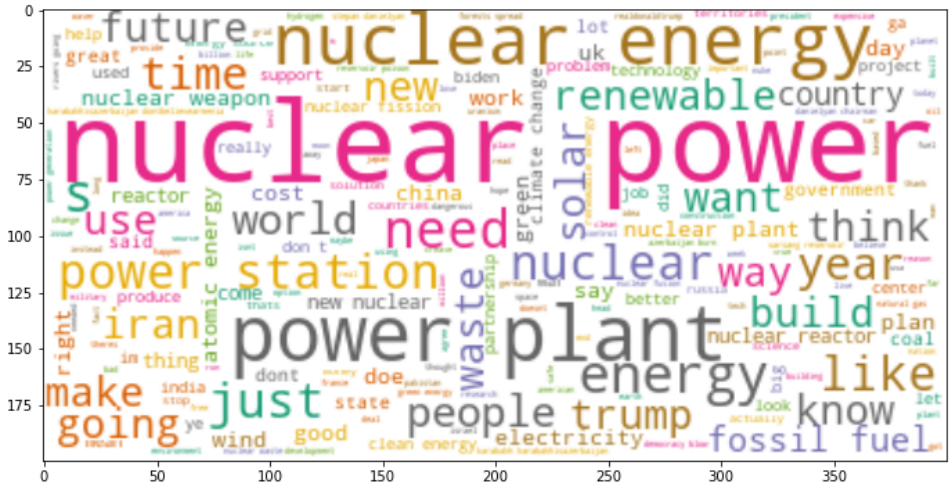

Figure 1. Wordcloud of top 200 words.

serves, particularly near RSPB Minsmere, a nature reserve in England. The forth largest topic was labeled 'science' and contained a variety of tweets discussing technicalities of nuclear energy or using terms such as 'fission' and 'fusion'. The smallest topic was labeled 'conflict between Azerbaijan and Armenia' and contained tweets about a recent quote of Stepan Danielyan, Chairman of the Center for Partnership for Democracy, allegedly urging the spread of nuclear waste in territories of Karabakh. For each topic the amount of tweets it contains and a representative tweet (with 95% or higher probability of belonging to the specific topic) was selected and displayed in Table. 1.

3.3 Sentiment Analysis

The mean polarity sentiment of all tweets was evaluated at 0.0695, indicating a slight positive sentiment. There were 16591 positive tweets (polarity score > 0), 8386 negative tweets (polarity score < 0) and 9202 neutral tweets (polarity score = 0). The mean subjectivity of all tweets was 0.379, indicating the tweets are slightly more objective than subjective. The mean polarity and subjectivity of each topic are displayed in Table. 2. 'Energy production and climate' and 'Power plant near nature reserve' have a more positive sentiment while 'Azerbaijan and Armenia conflict' has a more negative sentiment, when compared to the total data set. 'Energy production and climate' and 'Trump and international conflict' are the most subjective topics, while 'Azerbaijan and Armenia conflict' and 'Science' are the most objective topics. A scatter plot of the polarity and subjectivity score of all tweets is illustrated in Fig. 3.

Topic Label	Tweets/ Topic	Representative Tweet
Energy production and climate	13162	@fmeikle Germany is full of ironies - massive build out of wind and solar (good) but emissions still very high. Because shut nuclear early, and burning lignite for power (bad) - to keep industry with reliable supply.It's a nonsense for a supposedly rational country
Trump and international conflict	7803	Just a reminder that Sen. Jim Inhofe (R) blasted Trump Energy Sec. Dan Brouillette Friday for forcing out National Nuclear Security Admin chief Lisa Gordon-Hagerty, saying it showed Brouillette "doesn't know what he's doing in national security matters."
Power plant near nature reserve	3660	@SZCConsortium Not welcome right up against the best nature reserve in the UK PLEASE sign & retweet v.important petition below against proposed new nuclear power station right on border with Minsmere - station would have a big impact on the reserve itself
Science	5402	Current forecast is sunny so I'm shining orange. Current temperature is 69.3 degrees, current humidity is 67%. Random sunny fact: The energy created by the Sun's core is nuclear fusion.
conflict between Azerbaijan and Armenia	2649	@bbcazeri Stepan Danielyan, Chairman of the Center for Partnership for Democracy: "Blow up the Sarsang reservoir, poison the rivers going to Azerbaijan, burn all forests, spread the waste of the nuclear power plant" in territories of Karabakh." #KarabakhisAzerbaijan #DontBelieveArmenia

Table 1. Topic Clusters and their most representative tweet.

Topic Label	Mean Polarity	Mean Subjectivity
Energy production and climate	0.0896	0.404
Trump and international conflict	0.0515	0.391
Power plant near nature reserve	0.0840	0.367
Science	0.0709	0.364
Azerbaijan and Armenia conflict	0.0169	0.249

Table 2. Polarity and Subjectivity score of each topic.

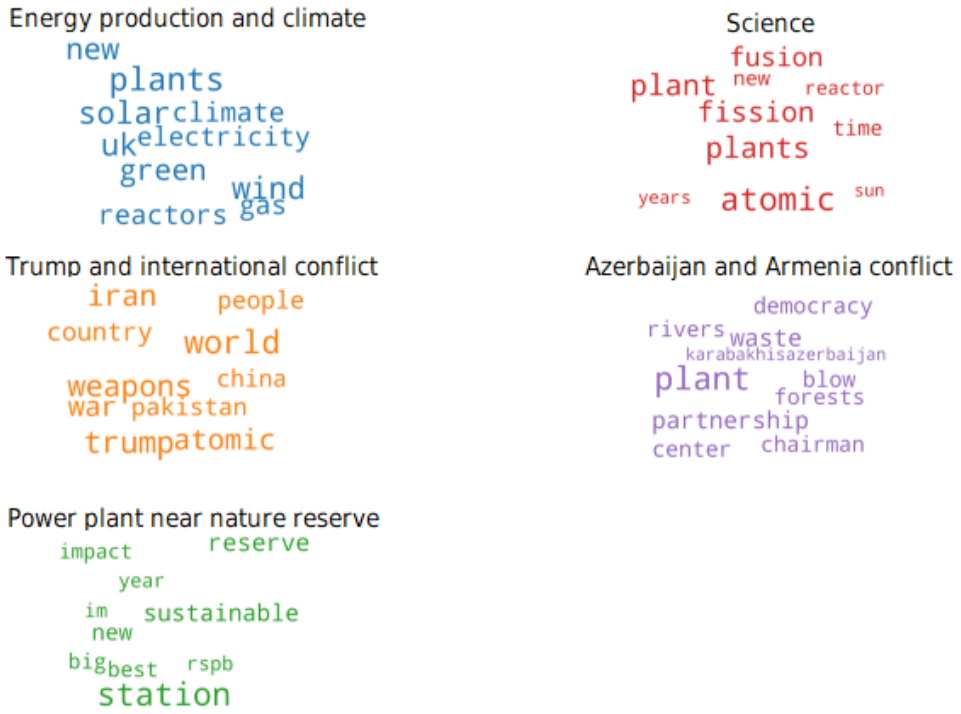

Figure 2. Wordcloud of top 10 words for each of the 5 topics.

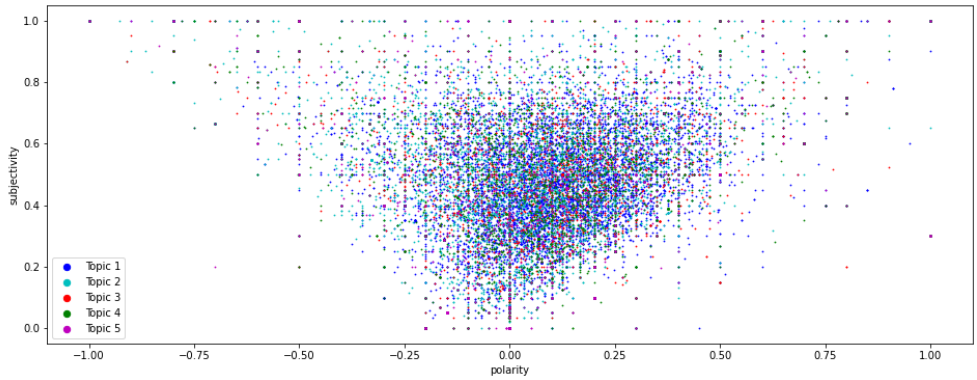

Figure 3. Scatter plot of subjectivity and polarity values of each tweet and to which topic they belong.

4 Discussion

Knowing the most prevalent topics in nuclear energy conversations and what the public's sentiment is about them can be helpful for decision makers in politics as well as in industry. Analyzing Tweets allows investigation of public discourse on a scale previously impossible. Gathering tweets over a 1 month period and analyzing them, we were able to distinguish 5 topics out of 39400 tweets and evaluate their polarity and subjectivity. Furthermore, utilizing machine learning approaches to group tweets into topics and analyse their sentiment avoids the introduction of bias from the researcher.

The topic modelling reflected general topics as well as topics that were

discussed in the media during the period of data gathering [34, 35, 36]. This demonstrates that it can be a useful and accurate tool when investigating public opinion. Energy production and climate is the most discussed topic, and solar, wind, new and green are some of the most important words in those conversations. Interestingly, safety or price was not among the most frequent words, nor a frequent topic in conversations.

While many tweets tend to be opinion-based, tweets about nuclear energy are more on the objective side. They are also generally positive, particularly when it comes to energy production and climate. However, tweets are short messages, and variance in polarity and subjectivity is high, as can be seen in Fig. 3. Topic modelling can be an effective tool to group similar tweets together, thus reduce complexity and allow meaningful differences in polarity and subjectivity to be extracted.

There are several limiting factors to this study. Firstly, even though Twitter has 330 million users, there is some representation lost for people with no access to internet or social media. Secondly, while the key words to gather the tweets are nuclear energy related, it is possible that tweets are gathered that are not necessarily related to a discussion on nuclear energy, e.g., song lyrics. Furthermore, while the model was able to accurately capture the topics that were discussed in the tweets, a longer study would have to be done to distinguish more general topics in the nuclear energy conversation. 'Energy production and climate' and 'Science' could be considered perennial topics when investigating nuclear energy conversations, while 'Trump and international conflict', 'Azerbaijan and Armenia conflict' and 'Power plant near nature reserve' are more indicative of topical subjects. This does not mean that these topics are uninformative. The first two could have a shared underlying topic of 'Politics' and the third 'Nature preservation', which are more general subjects independent of current affairs. Gathering data over a longer period of time would also result in more data in general, which would likely improve performance even further. In addition, more pre-processing could have been explored, such as removing spelling errors and lemmatization. This could result in a better performance of the LDA model and the sentiment analysis. Lastly, although the most appropriate input and number of topics for the LDA model was carefully chosen through thorough analysis of representative tweets and semantic similarity of contributing words of each topic, a more objective criteria could be created to select the optimal model.

References

[1] United nations treaty collection: 7. d paris agreement. https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en. Accessed: 12.11.2020.

[2] Kojo Menyah and Yemane Wolde-Rufael. Co2 emissions, nuclear energy, renewable energy and economic growth in the us. *Energy Policy*, 2010.

[3] Explore energy data by category, indicator, country or region. https://www.iea.org/data-and-statistics/data-tables?country=WORLD&energy=Balances&year=2018. Accessed: 12.11.2020.

[4] Shuji Takashina. *Nuclear Power in an Age of Uncertainty*, chapter 8, pages 211–224. 1984. Public Attitudes Toward Nuclear Power.

[5] Bronwen Daniel. How can we best reduce global warming? school students' ideas and misconceptions. *International Journal of Environmental Studies*, 2007.

[6] Younghwan Kim, Minki Kim, and Wonjoom Kim. Effect of the fukushima nuclear disaster on global public acceptance of nuclear energy. *Energy Policy*, 2013.

[7] Anil Markandya and Paul Wilkinson. Electricity generation and health. *THE LANCET*, 2007.

[8] Hannah Ritchie. What are the safest and cleanest sources of energy? *Our World in Data*, 2020.

[9] Shirley S. Ho, Tsuyoshi Oshita, Jiemin Looi, Alisius D. Leong, and Agnes S.F. Chuah. Exploring public perceptions of benefits and risks, trust, and acceptance of nuclear energy in thailand and vietnam: A qualitative approach. *Energy Policy*, 2019.

[10] Nicholas L. Miller. Why nuclear energy programs rarely lead to proliferation. *International Security*, 2017.

[11] Ying Lin. 10 twitter statistics you need to know. 2020.

[12] Sameh N. Saleh MD, Christoph U. Lehmann MD, Samuel A. McDonald MD, Mujeeb A. Basit MD, and Richard J Medford MD. Understanding public perception of coronavirus disease 2019 (covid-19) social distancing on twitter. *Infection Control & Hospital Edidemology*, 2020.

[13] Amir Karami, Alicia A Dahl, and Hadi Kharrazi. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 2018.

[14] Hea-Jin Kim, Yoo Kyug Jeong, and Min Song. Topic-based content and sentiment analysis of ebola virus on twitter and in the news. *Journal of Information Science*, 2015.

[15] Lauren E. Sinnenberg, Christie L. DiSilvestro, Christina Mancheno, Karl Dailey, Christopher Tufts, Alison M. Buttenheim, Fran Barg, Lyle Ungar, H Schwartz, Dana Brown, David A. Asch, and Raina M. Merchant. Twitter as a potential data source for cardiovascular disease research. *JAMA Cardiol*, 2016.

[16] Salvatore Pirri, Valentina Lorenzoni, Gianni Andreozzi, Marta Mosca, and Giuseppe Turchetti. Topic modeling and user network analysis on twitter during world lupus awareness day. *International Journal of Environmental Research and Public Health*, 2020.

[17] Amir Karami, Vanessa Kitzie, and Frank Webb. Characterizing transgender health issues in twitter. 2018.

[18] King-Wa Fu, Hai Liang, Nitin Saroha, Wion Tsz Ho Tse, Patrick Ip, and Isaac Chun-Hai Fung. How people react to zika virus outbreaks on twitter? a computational content analysis. *American Journal of Infection Control*, 2016.

[19] Twitter developper. https://developer.twitter.com/en. Accessed: 12.11.2020.

[20] Python software foundation. https://www.python.org/psf/. Accessed: 29.11.2020.

[21] Tweepy. https://www.tweepy.org/. Accessed: 12.11.2020.

[22] Xiaobing Sun, Xiangyue Liu, Jiajun Hu, and Junwu Zhu. Empirical studies on the nlp techniques for source code data preprocessing. 2014.

[23] Python re. https://docs.python.org/3/library/re.html. Accessed: 12.11.2020.

[24] Scikit-learn. https://scikit-learn.org/stable/. Accessed: 29.11.2020.

[25] Wordcloud. http://amueller.github.io/word_cloud/. Accessed: 29.11.2020.

[26] matplotlib. https://matplotlib.org/. Accessed: 29.11.2020.

[27] David M. Bei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. 2003.

[28] Carina Jacobi, Wouter van Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 2015.

[29] Gensim. https://radimrehurek.com/gensim/. Accessed: 12.11.2020.

[30] Natural language toolkit. http://www.nltk.org/. Accessed: 29.11.2020.

[31] Textblob. https://textblob.readthedocs.io/en/dev/. Accessed: 12.11.2020.

[32] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, 2011.

[33] Textblob documentation. https://textblob.readthedocs.io/en/dev/quickstart.html. Accessed: 29.11.2020.

[34] Trump 'asked for options on strike on iran nuclear site'. https://www.bbc.com/news/world-middle-east-54972269. Accessed: 12.11.2020.

[35] Fresh fears for minsmere as pm prepares special environment speech. https://www.eadt.co.uk/news/rspb-minsmere-sizewell-c-damage-1-6926669. Accessed: 12.11.2020.

[36] Armenian political scientist calls for ecological terrorism. https://defence.az/en/news/148530. Accessed: 12.11.2020.