Using NLP to Find Barriers to Humanizing Energy Transition

I learned about Omdena when I won the Intel Edge AI scholarship program and through other scholars who used to share resources with us. At the time the project that grabbed my attention was about employing NLP approaches to understand how citizens of different countries perceive their role toward de-carbonization and energy transition. Since my experience in NLP was limited to course-works, I was very excited to get involved in the project and expand my experience.

## **AI for Energy Problem**

The philosophy behind this project is to find ways that governments can benefit from to involve people in accelerating energy transition and taking advantage of sustainable energy resources. In fact, people play an important role besides state-of-the-art technologies and business boosters. How could we get these insights? The problem was quite complicated and there was no data provided to us and we were supposed to create our own dataset, analyze it and provide the world energy council (WEC) with insights about it. We had only two months to complete the project and there were lots of possible next steps and questions:

What should our output look like?

What search terms would be useful to scrape our data with?

What countries should be considered as our main focus?

Should we consider Non-English languages as well and analyze them?

How much data per country will be enough?

What would be the best sources for collecting data for each specific country?

What keywords should be used to scrape data so that it provides us with useful insight?

What would be better than keywords? Summaries? Bigrams? Topic modeling?

What should our output look like and what kind of insights will be beneficial to WEC?

A lot could be done with the data we were going to scrape. In order to progress and meet the deadline of the project we prioritized the steps we needed to take. We decided to go with English languages only and come up with good working models on them. Then, the next step would be to analyze non-English languages as well. For start, we decided to go with scraping the following resources: **Twitter, YouTube, Facebook, Reddit and famous newspapers**. We were supposed to provide WEC with insights for developed, developing, and under-developed countries and the emphasize was specifically on **developing**, and **under-developed** countries. The results discussed in this article obtained from *scraped tweet data* only obtained for specific countries including *USA, UK, India,* and *Nigeria* which cover the three categories of developed, developing, and under-developed countries.

## **Our Approach to the Problem**

I added stop-words, stemming, removed hashtags, punctuations, numbers, mentions and replaced urls with \_URL. I used TF-IDF vectorization for feature extraction of the articles. I am going to walk you through various steps have been taken to tackle the problem.

**Sentiment Analysis:**

For sentiment analysis, we first gathered data by scraping tweets using several specific keywords we found to be important for specific countries using google trends. Sentiment analysis of short tweets data comes with its own challenges and some of the important challenges are:

* Tags mean different things in different countries. #nolight can be Canadians complaining about the winter sunset; or Nigerians having a power cut.
* Tags take a side. For example, #renewables is pro-green and #climatehoax is not. So positive sentiment on #renewables might not really tell us much.
* The classifier model built on #climatechange and related tags does not work at all on the anti-green tags such as #climatemyth.
* Some anti-green tweets are full of happy emojis which makes the sentiments unreliable.
* The major tweeting countries are overwhelmingly positive. In fact, the distribution of the climate change related tweets across world is not uniform and the number of tweets across some countries are more prevalent in the dataset as compared to others [1].
* The interpretation of outputs. In fact, by just assigning labels to each tweet we will not be able to derive insights on the barriers to energy transition. Therefore, the interpretability of the model is very important.

As a result, the sentiment analysis on the tweets did not produce satisfactory results and we decided to test other models.

A close up of a map

Description automatically generated

**Figure1:** Number of climate change related tweets per country [1]

**Table1:** Classifier accuracy for sentiment analysis of tweets data (USA)

|  |  |
| --- | --- |
| Classifier Name | Accuracy |
| Logistic Regression | 0.68 |
| SVM-RBF | 0.69 |
| Decision Tree | 0.59 |
| Random Forest | 0.67 |
| SVM-Linear | 0.68 |
| MNB | 0.6 |

**Topic Modeling, Specifically Correlated Topic Modeling:**

Next step was to test topic modeling. Topic models have become a standard tool within quantitative text analysis for many different reasons. Topic models can be much more useful than simple word frequency or dictionary-based approaches depending upon the use case. Topic models tend to produce the best results when applied to texts that are not too short and those that have a consistent structure. Because tweets are short it will be hard to correspond them to a specific topic using models such as LDA.

Therefore, I used correlated topic modeling ([CorEX](https://github.com/gregversteeg/corex_topic/blob/master/corextopic/example/corex_topic_example.ipynb)) [2]. Topic modeling provides a way to compare the strength of different topics and tells us which topic is much more informative as compared to others. Since the data was very high dimensional, I needed to apply dimensionality reduction in order to remove noise and interpret the data. Permutation Tests is used to determine the optimum number of principal components required for PCA [3,4]. Through permutation tests and plotting the explained variance ratio, it appeared that the cumulative explained variance line is not perfectly linear, but it is very close to a straight line. Mean of the explained variance ratio of permuted matrices do not really differ from the explained variance ratio of the non-permuted matrix which suggests that applying PCA on correlated topic results are not helpful.

|  |  |
| --- | --- |
| A close up of a map  Description automatically generated | A close up of a map  Description automatically generated |
| **Figure 2.** Explained variance ratio | **Figure 3.** Mean explained variance ratio |

Well, is this good? No. This means each of the principal components contributes to the variance explanation almost equally, and there’s not much point in reducing the dimensions based on PCA.

The CorEx results showed that there are about 20 important topics and it was also showing the important words per topic. But how did I want to interpret the results? Data was very high dimensional and dimensionality reduction was not helpful at all. For example, if the word price is very important for one topic how do I want to understand the concerns of the people of that country? Is it fuel prices, electricity prices, ticket prices? There could be a combination of many different possibly related words in each topic and by just looking at them it would not be easy to find out what the data is telling us about.

|  |
| --- |
| A close up of a logo  Description automatically generated |
| **Figure 4.** Strength of each topic, showing how informative each topic is as compared to others |
| **A screenshot of a cell phone  Description automatically generated** |
| **Figure 5.** Results from the t-SNE dimensionality reduction algorithm of the vectors of a query word, and its list of most similar words. |

**Clustering (Kmeans & Hierarchical):**

Both Kmeans and Hierarchical clustering models lead to comparable results illustrating separate clear clusters. Because both models have comparable performance, we derived all results using Hierarchical clustering. Tweet data was collected for four different countries as discussed before and the model was applied to the data of each country separately to analyse the results. To summarize we only show the clustering results for India. But all the insights will be shown for all countries.

|  |  |
| --- | --- |
| A picture containing game  Description automatically generated | A close up of a logo  Description automatically generated |
| **Figure 6.** Dendrogram plot to visualize the history of groupings and figure out the optimal number of clusters | **Figure 7.** Results from the t-SNE dimensionality reduction |

## **Hierarchical Clustering Results**

### **Visualizations Using Scattertext**

Now that clear separate clusters could be achieved using clustering the next step would be to create meaningful visualizations and make sense out of data. [Scattertext](https://spacy.io/universe/?id=scattertext) is an excellent exploratory text analysis tool which allows cool visualizations differentiating between the terms used by different documents using an interactive scatter plot.

Two types of plots were created which was very helpful in interpreting the results. 1) Visualizing word embedding projections. This has been explored using Word Association with a specific keyword. The keywords tried includes the following: [***Access, Availability, Affordability, Bills, Prices***]. If the reader is interested, can try more keywords using the provided code in this study.  2) In another plot, the unigrams from the clustered tweets are selected, and plotted using their dense-ranked category-specific frequencies. We used this difference in dense ranks as the scoring function.

All the interactive plots are stored in an HTML file and are available in the GitHub repository.  If you click on the interactive version, the list of tweets with each specific term can be explored. All the following insights have been drawn looking into these plots. Please note that first hierarchical clustering is applied to the data and then the clustered tweets are given to Scattertext as an input. Note that further information can be gained by deep diving into these plots further, in case the reader is curious to explore them more. The data used for creating these results can be found [here](https://github.com/mkhoshle/AI-for-Energy/tree/master/pipedata-simon) and the notebook to apply clustering and create these scatter plots can be found [here](https://github.com/mkhoshle/AI-for-Energy).

The following shows the interactive versions of all plots for various countries:

***India:*** [Cluster 1 vs 2](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData12_India.html), [Cluster 1 vs 3](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData13_India.html), [Word Embedding: Prices](https://mkhoshle.github.io/AI-for-Energy/word_similarity_prices_India.html), [Word Embedding: Bills](https://mkhoshle.github.io/AI-for-Energy/word_similarity_bills_India.html)

***Nigeria:***[Cluster 1 vs 2](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData12_Nigeria.html), [Cluster 1 vs 3](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData13_Nigeria.html), [Word Embedding: Prices](https://mkhoshle.github.io/AI-for-Energy/word_similarity_prices_Nigeria.html), [Word Embedding: Access](https://mkhoshle.github.io/AI-for-Energy/word_similarity_access_Nigeria.html),

[Word Embedding: Affordability](https://mkhoshle.github.io/AI-for-Energy/word_similarity_affordability_Nigeria.html), [Word Embedding: Availability](https://mkhoshle.github.io/AI-for-Energy/word_similarity_availability_Nigeria.html)

***UK:*** [Cluster 1 vs 2](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData12_UK.html), [Cluster 1 vs 3](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData13_UK.html), [Word Embedding: Prices](https://mkhoshle.github.io/AI-for-Energy/word_similarity_prices_UK.html), [Word Embedding: Bills](https://mkhoshle.github.io/AI-for-Energy/word_similarity_bills_UK.html), [Word Embedding: Availability](https://mkhoshle.github.io/AI-for-Energy/word_similarity_availability_UK.html)

***USA:*** [Cluster 1 vs 2](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData12_USA.html), [Cluster 1 vs 3](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData13_USA.html), [Word Embedding: Prices](https://mkhoshle.github.io/AI-for-Energy/word_similarity_prices_USA.html), [Word Embedding: Bills](https://mkhoshle.github.io/AI-for-Energy/word_similarity_bills_USA.html), [Word Embedding: Availability](https://mkhoshle.github.io/AI-for-Energy/word_similarity_availability_USA.html)

#### **Visualising rank and frequencies across different categories for *India***

**A screenshot of text

Description automatically generated**

**Figure 8.** An example Scattertext plot showing positions of terms based on the dense ranks of their frequencies, for cluster 1 & 2. The scores are the difference of the terms’ dense ranks. The bluer terms are, the higher their association scores are for cluster 1. The redder the terms, the higher their association score are for cluster 2. See [Cluster 1 vs 2](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData12_India.html) for an interactive version of this plot.

#### **A screenshot of a cell phone Description automatically generated**

**Figure 9.** An example Scattertext plot showing positions of terms based on the dense ranks of their frequencies, for cluster 1 & 3. The scores are the difference of the terms’ dense ranks. The bluer terms are, the higher their association scores are for cluster 1. The redder the terms, the higher their association score are for cluster 3. See [Cluster 1 vs 3](https://mkhoshle.github.io/AI-for-Energy/ScattertextRankData13_India.html) for an interactive version of this plot.

#### **Word Embedding projection plots using Scattertext for *India***

#### **A screenshot of a cell phone Description automatically generated**

**Figure 10.** An example Scattertext plot showing word associations to term ***prices*** using Spacy's pretrained embedding vectors. This is used to see the terms most associated with the term ***prices.*** At the top right corner, we see the most commonly associated words with the term ***prices*** such as electricity. If you click on the interactive version, the list of tweets with the terms can be explored. See [Word Embedding: Bills](https://mkhoshle.github.io/AI-for-Energy/word_similarity_bills_India.html) for an interactive version of this plot.

#### **A close up of text on a white background Description automatically generated**

**Figure 11.** An example Scattertext plot showing word associations to term ***bills*** using Spacy's pretrained embedding vectors. This is used to see the terms most associated with the term ***bills.*** At the top right corner, we see the most commonly associated words with the term ***bills*** such as electricity, prices, energy, power. If you click on the interactive version, the list of tweets with the terms can be explored. See [Word Embedding: Prices](https://mkhoshle.github.io/AI-for-Energy/word_similarity_prices_India.html) for an interactive version of this plot.

#### **Some Insights Regarding Price & Energy Transition Concerns Across Different Countries obtained using Scattertext:**

**India:**

* Solar and wind don't necessarily mean cheaper prices as it did not cause so in Germany. When Germany went all on **renewables, energy prices** and carbon emissions went up.
* The electrical **prices can** drop for people who are sourcing power from the government owned renewable sources, because the prices are not going to vary with oil and natural gas.
* Renewable energy policy can lead to much **lower electricity prices,** stronger globally competitive economy, less import of fossil fuels and as a result less pollution.
* Putting tax on coal and making open access a reality are two potential action areas to make renewable energy **affordable.**
* Let **oil prices** increase and subsidies stop.
* Many requests to replace fossil fuels with cleaner fossil fuels such as stubbles from farmers.
* Cut **oil imports** and encourage **renewable energies.**
* A lot of complaints regarding **electricity shortage,** lack of electricity for hours or days, electricity cut, electricity and water supply.
* Fossil **fuels** are dirty, and **Nuclear power** is dangerous. Therefore, we need to make **renewable energy** work.

**Nigeria:**

* People complaining about no constant **electricity,** and zero business friendly policy.
* Enhancing the delivery of **Electricity** in the country.
* Whenever it rained **electricity supply** was **cut off** for days, **lack of electricity** every weekend daily and overnight, and **unstable electricity**.
* No **water** and no **Electricity.**
* The **electricity** sector is the third main consuming sector of oil.
* Lots of worries and trouble regarding **paying** **electricity bills**.
* **Access** to **electricity** is not for everyone.
* **Access** to **affordable** sustainable **renewable** energy.
* **Renewable** energy **water** and **waste** management are some of Nigeria’s major partnership areas with Ghana.
* Harnessing tidal or offshore wind energy which is a clean and **renewable** source.
* Lots of positive experiences and low prices with the usage of **Solar** power systems.

**UK:**

* Bringing down the **prices** of **electricity** and **gas**.
* Having **stable prices** for **electricity.**
* People prefer **higher prices** for **gas** than **electricity**.
* Need to think **beyond electricity** to affect the **energy transition.**
* **Renewables** disrupt the **electricity market** and politicians raising **electricity prices** to tackle **climate emergency** problems is an **awful policy**.
* A lot of requests on investment on **Renewable Energies**.
* The **transition to renewable** is being too **slow**.
* Lots of discussions on whether it is good to replace the **nuclear stations** with **renewables**.
* Whether the **zero-carbon economy** has any **economic benefit** for the UK.

**USA:**

* **Slowing down climate change.**
* **Market-based** solutions for **climate change.**
* **Renewable energy infrastructure** is **lame** and **unreliable**.
* **Renewables** increase **electricity prices** and **distort energy markets** with favorable purchase agreements.
* Many complaints regarding **gas prices**.
* **National security’s priority** should be on renewable energyInvesting in its infrastructure and jobs progs.
* Figure out how to **store renewable energy** and get rid of **excess CO** in the atmosphere.
* **Renewable energy** represents a significant economic opportunity.

## **Dispersion Plot**

A word’s importance can be weighed by its dispersion in a corpus. Lexical dispersion is a measure of a word’s homogeneity across the parts of a corpus. The following plot notes how many times a word occurs throughout the entire corpus for different countries including India, Nigeria, UK, and USA. According to the following dispersion plot, access to electricity is an important concern while this is not the case for the other three countries. How do we know that this access is related to electricity? Well, the answer is Scattertext plots shown in the previous section. Analyzing those plots together with the dispersion plot shows that the concern is mostly related to the electricity access.

Access to affordable renewable energy is a big concern in Nigeria and then India, while affordability of renewable energy is not a problem for people in the UK and USA. Affordability is a big concern for the people in Nigeria and people have difficulty paying their electricity bills. Energy, electricity, power and renewables are also the topic of most of the discussions in all the countries. But what aspects of each topic is of concern to each country? The answer is given in the previous section where we interpret the results of Scattertext plots.

A screenshot of a cell phone

Description automatically generated

Top Trigrams for Different Countries:

The data used for creating these results can be found [here](https://drive.google.com/drive/folders/194O5sEH6nY8tFo1sXxdsHvaDE9FgIyH-) and code to produce these results can be found [here](https://github.com/omdena/wec/blob/master/Task3-sentiment_analysis/Clustering/Visualizations/Visualizations.ipynb).

A screenshot of a cell phone

Description automatically generated

As can be seen from the top 20 trigrams for ***India*** the top concerns are: **Renewable energy, Renewable energy sector, Renewable energy capacity, Renewable energy sources, New renewable energy, and Clean renewable energy.** These top concerns specifically match the insights drawn from clustering in the previous section.

A screenshot of a cell phone

Description automatically generated

As can be seen from the top 20 trigrams for ***Nigeria*** the top concerns are: **Renewable energy, Renewable energy training, Electricity distribution companies, Renewable energy sources, Renewable energy solutions, Solar renewable energy, Renewable energy sector, Affordable prices, Power Supply, Climate change renewables, Public private sectors, Renewable energy industry, Renewable energy policies and Access to renewable energy.** These top concerns specifically match the insights drawn from clustering in the previous section.

A screenshot of a cell phone

Description automatically generated

As can be seen from the top 20 trigrams for the ***United-Kingdom*** the top concerns are: **Free renewable energy, Renewable energy sources, Using renewable energy, New renewable energy.** These top concerns specifically match the insights drawn from clustering in the previous section.

## **A screenshot of a cell phone Description automatically generated**

As can be seen from the top 20 trigrams for ***USA*** the top concerns are: **Clean renewable energy, Renewable energy sources, Supporting renewable energy, Renewable fuel standard, Transition into renewable energy, Solar renewable energy, New renewable energy, Using renewable energy, Need for quality products, and renewable energy jobs.** These top concerns specifically match the insights drawn from clustering in the previous section.

## **Collocated word clouds & Co-occurrence Network Implementation & Visualization**

This plot displays the networks of co-occurring words in tweets on different countries. Here, we visualize the top 25 occurring bigrams as networks. For this analysis, we replace urls with the word URL, lower case the words, and remove stop and collection words from the tweets.

A close up of a newspaper

Description automatically generated

**Figure 12. Collocate Clouds-India**

A picture containing different, table, various, photo

Description automatically generated

**Figure 13. Co-occurrence Network-India (First 25 Bigrams)**

A picture containing table, screen

Description automatically generated

**Figure 14. Collocate Clouds-Nigeria**

A picture containing map, table, different, various

Description automatically generated

**Figure 15. Co-occurrence Network-Nigeria (First 25 Bigrams)**

A screenshot of a cell phone

Description automatically generated

**Figure 16. Collocate Clouds-UK**

**A picture containing different, table, air, photo

Description automatically generated**

**Figure 17. Co-occurrence Network-UK (First 25 Bigrams)**

**A screenshot of a cell phone

Description automatically generated**

**Figure 18. Collocate Clouds-USA**

A picture containing different, photo, various, table

Description automatically generated

**Figure 19. Co-occurrence Network-USA (First 25 Bigrams)**

**About Omdena**

**Building AI through global collaboration**

**Omdena** is a global platform where changemakers build ethical and inclusive AI solutions to real-world problems through collaboration.

[Learn more about the power of Collaborative AI](https://omdena.com/).

## **References:**

[1] B. Dahal, S.A.P. Kumar, Z. Li, Topic modeling and sentiment analysis of global climate change tweets, Social Network Analysis and Mining, 9:24, 2019, <https://doi.org/10.1007/s13278-019-0568-8>.

[2] Gallagher, Ryan J., Kyle Reing, David Kale, and Greg Ver Steeg. "[Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge](https://www.transacl.org/ojs/index.php/tacl/article/view/1244)." Transactions of the Association for Computational Linguistics (TACL), 2017.

[3] <http://www.win-vector.com/blog/2016/05/pcr_part3_pickk/>.

[4] <https://bioconductor.org/packages/devel/bioc/vignettes/ClusterSignificance/inst/doc/ClusterSignificance-vignette.html#score-permutation>