Twitter Analysis on Sustainability

Anis Bouhamadouche
Department of Computer Science
University of Exeter, Exeter, UK
ab1429@exeter.ac.uk
710003602

***Abstract*—The aim of this project is to serve as contribution to the change that is happening in the area of sustainability and the application of NLP. I have used several NLP techniques to analyse tweets that speak about sustainability in the hope of gaining insight, and provide an efficient way to build the very first sustainability lexicon.**

***Index Terms*—NLP, ESG, sustainability, topic modelling, co-occurrence, semantics, lexicon**

I. Introduction

Sustainability has been a hot topic during the past few years. Governments, companies, and organizations are constantly investing in research in order to come up with solutions for various sustainability issues. On the other hand, ESG has been a center of interest for investors as it is, nowadays, used as a metric to assess investment risks based on Environment, Social, and Governance factors. Very often, the result of the ESG risk analysis is referred to as ESG ratings, which scores various companies regarding their performance in each of the ESG areas such as carbon emissions, waste management, corruption, tax, and employee satisfaction (the list is non-exhaustive).

Every organization puts its own assumptions while performing the analysis, using different taxonomies, which induced the absence of a baseline and ground truth scoring, making ESG ratings less reliable [1]. According to various discussions that I have had with some ESG experts and analysts both in the UK and USA, the main reason for this incoherence is the absence of regulations in this area which left a margin of creativity for organizations in reporting those ratings. Another issue that arises from ESG scores that are currently produced is the lack of correlation, that is due to different methodologies and assumptions made by different analysts.

This piece of work will be the first milestone towards building an open source sustainability lexicon. Building the lexicon will encourage research in NLP and sustainability in the hope of developing a standardized ESG taxonomy. The resulting taxonomy will facilitates the work of data scientists and researchers in NLP and sustainability. To build the sustainability lexicon, I have conducted an analysis on 2.6 million tweets (this work) pulled from Twitter API using sustainability and governance keywords in order to get relevant tweets. The aim here is to understand twitter content that is related to ESG, identify some keywords, analyse how words are related to each other, and take advantage of the semantic space through top2vec to categorize tweets for the ease of analysis and keywords extraction.

The first section in this report covers the data extraction details, dataset specifications, and data pre-processing. The section that follows reveals the details of the used techniques, explaining how different algorithms were implemented. Finally, the last section presents the obtained results and interpretation.

II. Data Acquisition and Preprocessing

Usually, the ESG score is derived from news articles as well as Twitter content. Considering the access to the data, Twitter provides quick and very easy access to large amount of data that we can leverage for this purpose. I have used Twitter API V2 to pull 2.6 million tweets relating to sustainability and governance posted starting the *January 2022* until the *1st of May same year*. I have used the Tweet Lookup, querying using more than 40 keywords and expressions derived from a proposed generic ESG taxonomy that looks like the following figure.

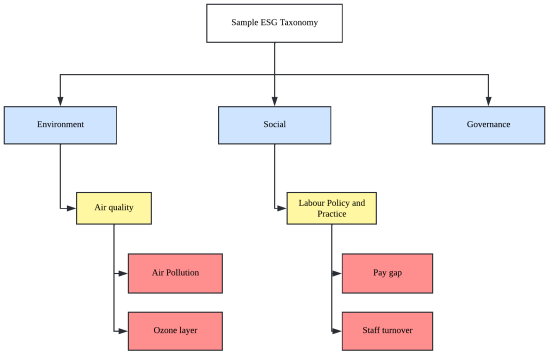

Fig. 1: Example ESG Taxonomy

The blue terms are called pillars, the yellow terms are themes, and the red terms are attributes. All of these terms have been used to query from Twitter API. The number of tweets that I have pulled is 2,628,954 tweets having the attributes shown in figure 2. My main interest here is in the content of the tweets that is contained under the column `text`. I have made sure to exclude retweets, remove duplicates, remove rows with null values of `text`, and remove the column `Unnamed: 0`.

I ended up with 2,433,073 tweets with 61 languages most of them in English (2,005,818). A brief exploratory data analysis shows that the users with the highest number of tweets relating

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2628954 entries, 0 to 487
Data columns (total 14 columns):
 #   Column                  Dtype
---  ------                  -----
 0   Unnamed: 0              object
 1   author_id               object
 2   username                object
 3   text                    object
 4   created_at              object
 5   retweets                float64
 6   replies                 float64
 7   likes                   float64
 8   quote_count             object
 9   in_reply_to_user_id     object
 10  geo                     object
 11  referenced_tweets_id    object
 12  referenced_tweets_type  object
 13  lang                    object
dtypes: float64(3), object(11)
memory usage: 300.9+ MB
```

Fig. 2: Tweets attributes

sustainability are mainly political figures, news companies, and CEOs of big tech companies. After that, I proceeded with NLP stops to start my analysis. I first converted tweets to lowercase, then, I have removed any links that are present in the tweet - in my case, most of the tweets contained links to other websites or tweets. Next, I have used `word_tokenize()` povided by `nltk` to segment tweets into individual words, removed stopwords, and finally, lemmatized them through `WordNetLemmatizer`. The size of the dictionary after preprocessing became 26,286,383, the reason why the size of the dictionary is this big is because of user mentions containing lots of unique words. The resulting frequency distance plot for the 40 most frequent words is shown in the following figure.

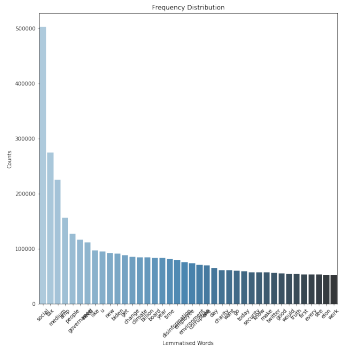

Fig. 3: Frequency distance of the 40 most frequent words

We can see that we already have some ESG related keywords that appear very frequently, as an example: corruption, social, tax, security, charity, and misinformation. This is also a good way to confirm that the gathered data is what we expected - related to sustainability.

III. Methodology and Discussion

The first thing that I wanted to explore is words associations and how words occur with each other. To achieve this task, co-occurrence networks are the best. Before building the networks, I have followed the same steps as mentioned in the previous section to process the data. However, I have added an additional step that makes the analysis more refined; I have used *Part of Speech (POS) tagging* also provided by `nltk` to identify the names in the tweets as we are interested in terms that relate to sustainability. I have only selected words with the tags `{'NN', 'NNS', 'NNP', 'NNPS'}`. Next, I proceeded to build co-occurrence networks using a *context window* of 2 words [2], in other words I have considered *bi-grams*. The algorithm words as follows:

Algorithm 1 Building co-occurrence data structure

Require: List of tokenized tweets
Ensure: List of tuples (token, adjacent token, 1/frequency)
  $edgelist \leftarrow []$
  $nodes \leftarrow []$
  for tweet in tokenized tweets list **do**
    for token in tweet **do**
      add the tuple (token, following token) to edgelist
  Create dictionary using edgelist elements and set values to 0
  for edge in edgelist **do**
    increment value of dictionary whenever edge observed
  keep only bi-grams observed > 2
  for tokens in dictionary **do**
    add to nodes (token, token neighbour, 1/frequency)
  return $nodes$

As indicated in the pseudo-code, the distance between each node (word) in the co-occurrence graph is determined by the inverse of the frequency; the number of times the two words appeared together - meaning the more frequent they appear together, the closer the nodes are. Before drawing the networks, I proceeded with computing degree and closeness centrality for the resulting co-occurrence network. The following figure shows the results for centralities:

	Degree	Closeness
tax	0.078404	0.331215
medium	0.061348	0.330274
amp	0.048693	0.326491
people	0.039890	0.320024
year	0.022008	0.307110
corruption	0.025585	0.301532
money	0.012655	0.294766
time	0.015681	0.291748
change	0.016781	0.290235
day	0.013480	0.289179

Fig. 4: Centralities sorted by closeness centrality for the 10 words with highest values (only 40000 tweets used here)

In our context, degree centrality indicates the semantic richness of the concepts and shows how related a specific term is to others connected to it. On the other hand, closeness centrality here shows that a term has been observed more frequently with other terms with also higher closeness centrality. In the table we can definitely see some relevant concepts such as tax and corruption, both have high closeness centralities. Both degree and closeness centrality have a correlation value of 0.256 which makes sense because they both rely on direct

links. I have then plotted few co-occurrence networks for some sustainability related keywords by selecting those words and their neighbours, I also used closeness centrality for the color and the size of the nodes. Note that the centrality was computed for the entire co-occurrence network. The following figures are examples of such networks.

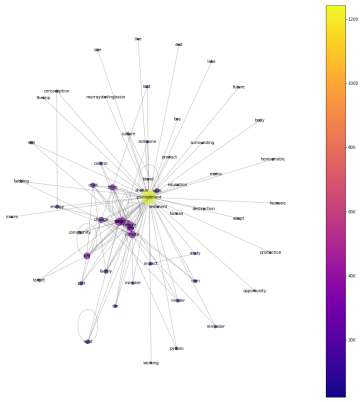

Fig. 5: Environment co-occurrence subnetwork

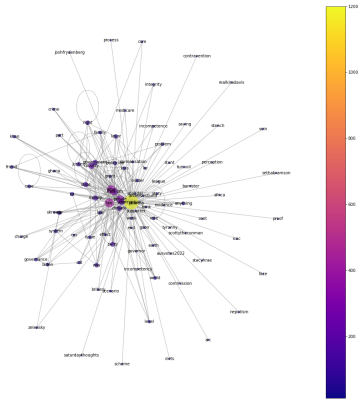

Fig. 6: Corruption co-occurrence subnetwork

We can see from both networks that most of the words that are closer to the center topic have also high values of closeness centrality compared to the ones that are far from the center. Reading those words also indicates that they actually relate to the center words. For the corruption network, we find the words `tax`, `government`, and `scandal` close to the word `corruption`. This provides an easier way to come up with a lexicon that evolves around corruption which falls under the pillar `Governance`. Another approach that could be used is through `word2vec` [3], which converts words into vectors in the semantic space allowing similar words to fall within the same region of the space [3]. To have a wider analysis, I have applied the same concept but converting tweets into the semantic space, allowing tweets having the same topic fall within the same region, that is achieved through `top2vec` [4]. The way this model works is by making joint document/word embeddings in a way that similar words express the same topic [4]. To get the vector representation for the topic, `top2vec` simply computes the centroid of the words evolving around that topic. Then, every new document (tweet) introduced to the model is represented by an embedding vector and the similarity is checked with the topic vector to be classified. Because of run-time restrictions on Collab, I have only used 500,000 tweets to train `top2vec`. I have obtained 3929 results during that process, some are relevant to sustainability and others are not. This is a good approach to also reduce the size of the corpus before extracting relevant words to be added to the lexicon. One of the issues for our use case is that top2vec produces fine-grained topic clusters, some topics can be combined under the same ESG theme or pillar. An example of the topic related to the word `environment` is provided in the code where I have got at least 10 topics related to environment. To solve this issue, we can use hierarchical topic modelling provided by `top2vec` thanks to `HDBSCAN` [4]. Finally, since each topic is represented by 50 most similar words by default (ref top2vec), we can use those keywords to build up our lexicon and taxonomy.

IV. Conclusion

In this work I have presented a potential way to approach the issues that exists with ESG data. I have proposed few steps to start building a sustainability lexicon that will help sustainability specialists and data scientists to perform research and analysis on sustainability issues to come up with better solution. I have started with a acquisition of tweets that evolve around that topic, then proceeded with a data processing pipeline that included converting text to lower case, removing links and stopwords, and lemmatization while keeping only nouns. I have build a co-occurrence network for words that occur with each other more than twice, plotted subnetworks having a sustainability keyword and its neighbour where I found out that closer words with also high closeness centrality are also relevant for sustainability and might also fall under the same ESG theme. Finally, I have performed topic modelling on the tweets where I obtained several topics that could be grouped as one where the representative keywords can be used to enrich the sustainability lexicon. Other techniques are still applicable and might come up with other insights such as building networks of users and see the interactions to detect communities based on their interest, and from there we can expect the tweets to be around a certain number of topics which allows us to know under which theme we would classify those words.

REFERENCES

[1] Abhayawansa, S. and Tyagi, S., 2022. Sustainable Investing: The Black Box of Environmental, Social, and Governance (ESG) Ratings.

[2] K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29.

[3] Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2022. Efficient Estimation of Word Representations in Vector Space. [online] arXiv.org. Available at: ¡https://arxiv.org/abs/1301.3781¿ [Accessed 23 May 2022].

[4] D. Angelov, "Top2vec: Distributed representations of topics," CoRR, vol. abs/2008.09470, 2020. [Online]. Available: https://arxiv.org/abs/2008.09470