A Comparison Evaluation Of Topic Modeling Methods For Climate Change Tweets

Orit Shahnovsky

ABSTRACT

Twitter and other social media platforms are increasingly used as data sources for analyzing public opinion on various topics, including climate change. This study explores a dataset of climate change-related tweets using two topic modeling techniques: Latent Dirichlet Allocation (LDA) and Top2Vec. The aim is to identify the main topics that emerge within the broader climate change discourse. The analysis reveals that discussions on climate change cover diverse subjects such as political activism, calls for action, and the impacts of global warming. Both LDA and Top2Vec are effective for this purpose, though careful application of these models is necessary to yield meaningful insights. This comparison underscores the strengths and limitations of each approach when applied to tweet data related to climate change.

ACM Reference Format:
Orit Shahnovsky. 2024. A Comparison Evaluation Of Topic Modeling Methods For Climate Change Tweets. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Climate change is one of the most urgent issues of our time, with significant environmental, social, and economic implications. Social media platforms, particularly Twitter, have become essential spaces for public discourse, where individuals, organizations, and policymakers discuss and debate climate change-related topics. This paper aims to apply two topic modeling algorithms—Latent Dirichlet Allocation (LDA) and Top2Vec—to identify and analyze topics within the broader theme of climate change. The insights gained from this analysis can contribute to raising awareness, promoting advocacy, and enhancing communication strategies around climate change. Furthermore, this study provides valuable guidance for future research seeking to understand public discourse on climate change through social media data. Twitter presents a particularly compelling case study due to its unique characteristics. While its 280-character limit poses challenges [6] for analysis, the platform's hashtag system effectively organizes conversations around specific themes and interests, making it a rich source for exploring the dynamics of climate change discussions.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

Early work in the field of topic modeling introduced LDA [3]. LDA provides a probabilistic framework that models each document as a mixture of topics, and each topic as a distribution over words. While LDA has been widely used for topic modeling, a newer method, Top2Vec, introduced by [1], presents a different approach. Top2Vec leverages vector embeddings from word and document representations, such as Word2Vec, to create a topic model without requiring the specification of the number of topics in advance. This makes it particularly useful for datasets where the underlying topic structure is not clearly defined. By clustering these embeddings in a semantic space, Top2Vec can generate high-quality topics in an unsupervised manner. LDA and Top2vec have been applied by [5]. In their work they to evaluated the performance of four topic modeling techniques, icluding BERTopic and NMF, on Covid Twitter posts, and assessed the performance of different algorithms concerning their strengths and weaknesses in a social science context. Based on certain details during the analytical procedures and on quality issues, this research sheds light on the efficacy of using BERTopic and NMF.

In [4] work, a large dataset of geotagged tweets containing certain keywords relating to climate change is analyzed using topic modeling and sentiment analysis techniques. Topic modeling shows that the diferent topics of discussion on climate change are diverse, but some topics are more prevalent than others. In particular, the discussion of climate change in the USA is less focused on policy-related topics than other countries.

3 DATASET

The analysis mas performed on Twitter Climate Change Sentiment Dataset[1]. This dataset comprises 43,943 tweets related to climate change, collected over the period from April 27, 2015, to February 21, 2018. After removing duplications, 41033 tweets were left. Table 1 presents a selection of tweets from the dataset. Since this work does not focus on sentiment analysis, the 'sentiment' column has been omitted from consideration.

The dataset requires cleaning before analysis. This process begins with manually examining a small sample to identify redundant information. Social media platforms, particularly Twitter, are often defined by the topics they discuss, frequently represented through hashtags. In this work, hashtags are of particular importance, as they capture the essence of the user's intent. To preserve this context, I used the TextPreProcessor from the ekphrasis[2] library to perform word segmentation on multi-word hashtags. This segmentation breaks down hashtags into their individual words, maintaining the original message's meaning. For instance, consider the sentence: *"Watch #BeforeTheFlood right here."* If the hashtag

[1] https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset?resource=download
[2] https://github.com/cbaziotis/ekphrasis

Table 1: Sample of tweets in the dataset

Sentiment	Message	tweetId
-1	"@tiniebeany climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom"	792927353886371840
-1	"RT @NatGeoChannel: Watch #BeforeTheFlood right here, as @LeoDiCaprio travels the world to tackle climate change https://t.co/LkDehj3tNn httÃ¢â‚¬Â¦"	793124211518832641
1	"Fabulous! Leonardo #DiCaprio's film on #climate change is brilliant!!! Do watch. https://t.co/7rV6BrmxjW via @youtube"	793124402388832256

were simply removed, the sentence would lose its intended focus. Additionally, the ekphrasis library was used to unpack contractions (e.g., don't → do not) and correct elongated words (e.g., coooool → cool).

Next, I employed the tweet-preprocessor library[3] to remove various elements from the tweets, including URLs, mentions, reserved words (e.g., RT, FAV), emojis, smilies, and numbers. This library also handled the removal of non-ASCII characters, such as Ã, Â, and â, which appeared in the original data due to encoding issues during the tweet hydration process.

Finally, using the nltk library, I removed stopwords and lemmatized the remaining text to further standardize and clean the dataset.

4 APPROACH

4.1 Model 1: Latent Dirichlet Allocation

LDA, is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities [3]. The generative process of LDA can be described as follows: given the M number of documents, N number of words, and prior K number of topics, he model learns to output two key parameters:

- Beta: The distribution of words for each topic K (the higher the beta, the more words the topic is composed of).
- Alpha: The distribution of topics for each document i (the higher the alpha, the more diverse the topics are across documents).

It is important to note that LDA is build from term frequency-inverse document frequency (TF-IDF). TF-IDF is a modification of the bag of words (BoW) feature extraction technique, and it tries to address the issues resulting from common yet semantically irrelevant words by accounting for each word's prevalence throughout every document in a text corpus.

In this research, the three hyperparameters was chosen as followed: a grid search was performed for the number of topics (K). The search for an optimal number of topics started with a range from 3 to 10, with a step of one. I did not use a larger number of topics since the results of 20 topics were already inferior and more topics would not be manageable. Beta and alpha were set to defualt.

Finally, to facilitate a clear interpretation of the extracted information from a fitted LDA topic model, pyLDAvis library was used to generate an inter-topic distance map. A screenshot of the statistical proximity of the topics can be seen in Figure 2. An interactive visualization is available at .

[3]https://pypi.org/project/tweet-preprocessor/

4.2 Model 2: Top2Vec

Top2Vec [1] uses word embeddings: vectorization of text data makes it possible to locate semantically similar words, sentences, or documents within spatial proximity. For example, words like "mom" and "dad" should be closer than words like "mom" and "apple." In this study, a pretrained embedding models, the Universal Sentence Encoder, was used to create word and document embeddings and it takes into account the context for each occurrence of a word. This model is suggested for smaller data sets. Since word vectors that emerge closest to the document vectors seem to best describe the topic of the document, the number of documents that can be grouped together represents the number of topics. However, since the vector space usually tends to be sparse (including mostly zero values), a dimension reduction was performed before density clustering. By using uniform manifold approximation and projection (UMAP), the dimensions were reduced to the extent that hierarchical density-based spatial clustering of applications with noise (HDBSCAN) could be used to identify dense regions in the documents. Finally, the centroid of the document vectors in the original dimension was calculated for each dense area, corresponding to the topic vector.

Notably, because words that appear in multiple documents cannot be assigned to one single document, they were recognized by HDBSCAN as noise. Therefore, Top2Vec does not require any preprocessing (e.g., stopwords removal), or stemming and lemmatization. To conclude this model, Top2Vec automatically provided information on the number of topics, topic size, and words representing the topics.

5 RESULTS

Although topic models bring in statistical analysis and can advance social science research, each of the algorithms has its own uniqueness and relies on different assumptions. Quantitative methods are limited in their ability to provide in-depth contextual understanding, and the results cannot be compared with any single value. Thus, the interpretation of the results still relies heavily on human judgment.

5.1 Model 1: LDA

Table 2 provides an overview of the 4 identified topics in the LDA model. Names were given based on the terms that contributed the most to a topic in reference to their TF-IDF weights.

Figure 1: Word Clouds based on Top2Vec Model

Figure 2 displays screenshot of the statistical proximity of the topics. The left-hand space each circle represents a topic, and its size reflects the proportion of tweets in the dataset assigned to that topic. Topic 1 has the largest share in the dataset, suggesting it dominates the overall theme. The proximity of the circles represents topic similarity. Topics closer together share more words or themes, while topics farther apart are more distinct. In our dataset topics appear relatively distinct, implying that the dataset is well-divided into unique themes.

The right-hand bar graph displays the most frequent terms in our model; Blue Bars represents the overall occurrence of the terms in the dataset, while the red Bars represents the estimated frequency of a term in its most relevant topic. Figure 3 is an example of the top-30 most relevant terms in topic 1. Since there are no blue bars, the terms do not have significant overall frequency in the dataset, meaning they are not common outside their associated topic.

The full visualization of the model results can be fount in https://bit.ly/49edRv0. Moving the mouse away from any of the topics, the bar graph returns to showing only the most frequent terms across the model overall.

5.2 Model 2: Top2vec

Top2Vec identifies topics as clusters of semantically similar documents. In this study, Top2Vec identified 224 topics in the dataset, which appears unusually high. Table 3 illustrates the descending distribution of tweet quantities per topic. Tweets often contain noise, such as slang, hashtags, and typos, which can cause the algorithm to identify "topics" that may lack meaningful coherence. Unlike LDA, Top2Vec does not require preprocessing, which might further contribute to this outcome. Furthermore, the brevity of tweets, typically reflecting narrowly focused ideas, can result in more fragmented topic assignments.

In order to acquire an overview of the importance of each term, a word cloud has been produced for better visualization for the first 4 topics (see Figure 1). Word clouds were generated with the font size of each word corresponding to the cosine similarity score, which measures how close the word vector was to the topic vector in semantic space. Thus, words with a high score were located at a small distance away from the topic vector in semantic space and were semantically closely related. A major observation is that many words are common between the topics between the topics, e.g *"deniers, algore, skeptic, cooling"*, unlike in LDA. The left section in Table 4 presents the keywords associated with the first four classified topics and provides names for these topics based on their thematic content. Since no pre processing has been made to the data, many words repeat itself in different versions (climatec change, climatechange) what makes it harder to identify the main keywords. Additionally, naming the fourth topic was challenging because the associated keywords were too general.

An advantage of the topic vectors is that the number of topics identified by Top2Vec can be hierarchically reduced to any desired number below the initial count. I reduced 224 topics to 4 and 5 categories and selected the result that was easiest to interpret with the help of ChatGPT. The results are listed in the ledft section in Table 4.

This is done by iteratively merging the smallest topic into its most semantically similar topic until the desired number of topics are reached. This is done by taking a weighted arithmetic mean of the topic vector of the smallest topic and its nearest topic vector, each weighted by their topic size. After each merge, the topic sizes are recalculated for each topic. This hierarchical topic reduction has the advantage of finding the topics which are most representative of the corpus, as it biases topics with greater size.

The categories in order of the number of documents were “Climate Sciences” with 41.0%, “Political Discourse” with 23.4%, “Polution” with 19.8% and “Global Impacts” with 15.8%.

Hierarchically reducing the number of topics can also be a disadvantage, as certain topics may "disappear" during the process. For instance, the initial analysis identified a topic labeled "melting glaciers," but this topic vanished after the topics were re-grouped.

6 EVALUATION

Researchers employ both qualitative and quantitative methods to evaluate topic models. Qualitative evaluation, often applied in real-world scenarios, involves manually inspecting the top keywords of each topic to assess their interpretability, a process that requires substantial domain expertise. Quantitative evaluation, on the other

Table 2: Topics identified by LDA

No.	Topic/content	Keywords
1	Climate Change Activism by Politicians	climate, change, trump, amp, say, said, could, fight, donald, hoax, want, effect, news, deal, would
2	Call to fight	global, warming, u, real, right, stand, still, stop, way, may, due, never, tell, itqs, problem
3	Debate about the impact	debate, new, one, believe, question, world, talk, think, know, science, president, good, planet, impact, also
4	A call for action	via, http, like, people, need, obama, make, issue, scientist, get, time, action, un, much, big

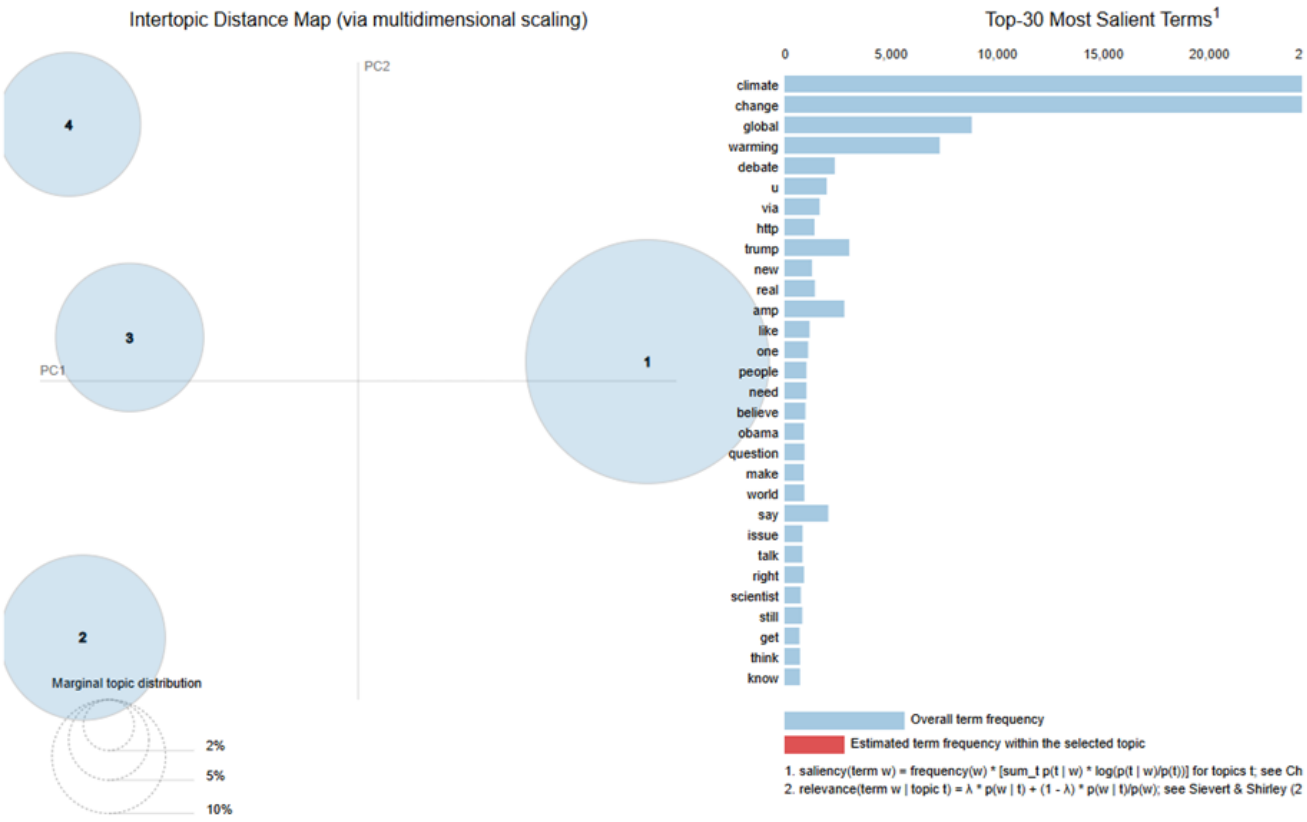

Figure 2: Visual inspection of LDA

hand, relies on metrics such as coherence scores [2], which measure the statistical probability and semantic cohesion of the topics generated by the model.

Topic coherence aims to assess the interpretability of a topic. It measures how coherent a topic is based on the semantic similarity of its top keywords. The coherence score ranges from 0 to 1, with 1 indicating perfect coherence and 0 indicating no coherence. While topic coherence is a popular metric, it is not an absolute measure of topic quality. However, it serves as a valuable tool for evaluating and comparing model performance.

Topic coherence methods generally sort each topic's key terms from highest to lowest term weights. It then selects the first n terms in each resepctive topic and measures the degeree of similarity of these terms within each topic. The Cv method passes over our entire coprus, enumerating term frequnecy and co-occurrence for the top n number of terms within each topic. It then uses these values to calculate normalized pointwise mutual information (NPMI) between every top word across the topics. In brief, NPMI is a statistical concept used to predict the probability that two independent events co-occur. NPMI prodcues a set of word vectors for each of the top key words considered. Cv then calculates distnace between between these vectors using cosine similarity. The final output coherence score is teh mean of these similarities.

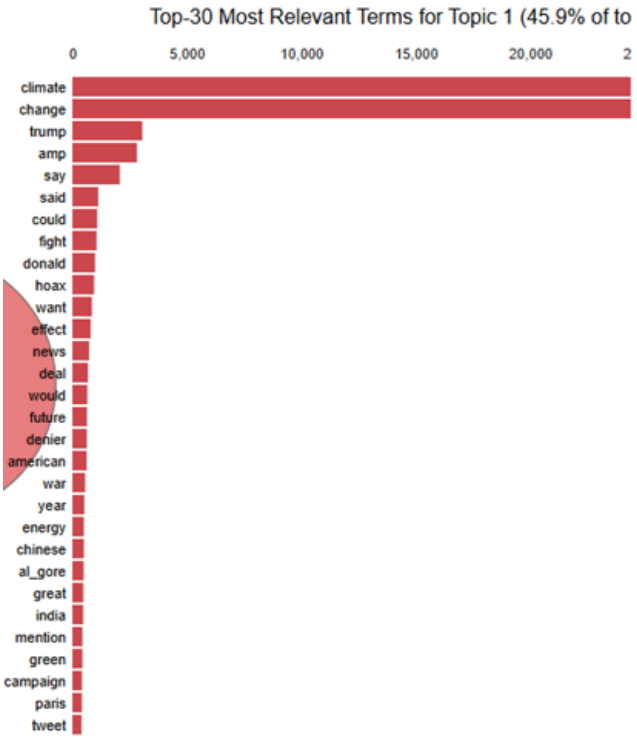

Figure 3: Top-30 most relevant terms in topic 1 in Top2vec model

Table 3: Quantity of tweets per topic

Topic No.	Quantity of tweets per topic
0	3419
1	2199
2	2111
3	1667
4	925
5	748
6	744
7	723
8	628
..	..
106	99
..	..
224	18

To evaluate topic coherence I used the top 10 words of each topic for LDA and Top2Vec.

Table 5 shows that Top2Vec outperforms LDA, demonstrating that it produces more coherent topics. However, for both algorithms, the underlying meanings of the topics are still subject to human interpretation.

7 DISCUSSION

In this study, I explored LDA and Top2Vec, two widely applied topic modeling algorithms on data climate tweets. I wanted to investigate what are the topics that are hidden within the tweets. It can be concluded that the tweets primarily discuss politicians' activism, calls to action, and the impacts of global warming on out planet.

Each model has its advantages and limitations. Top2Vec offers the benefit of automating the topic modeling process without the need to predefine hyperparameters such as the number of topics. However, a challenge arises from the hierarchical nature of topic discovery in Top2Vec, as the large number of topics generated can make it difficult to interpret and label them effectively. This requires either expert knowledge to name the topics accurately or the application of hierarchical clustering to group related topics.

Despite LDA's popularity, it has several weaknesses. To achieve optimal results, it often requires the number of topics to be predefined, as well as excessive data preprocessing, including removing stop words, lemmatization, and tokenization. Additionally, LDA relies on the bag-of-words representation of documents, which ignores the ordering and semantics of words. However, one advantage of LDA is the ease with which its outputs can be understood, and the visualizations facilitate easier adaptation of hyperparameters.

8 LIMITATIONS AND FURTURE WORK

There are several limitations in our research. First, LDA requires extensive data preprocessing, such as removing stop words, tokenization, and lemmatization, to achieve optimal performance. The necessity for manual preprocessing means that the quality of the model can be significantly influenced by the choices made during this step. To improve the model's performance, it may be useful to implement additional data preprocessing strategies. For example, replacing slang and abbreviations like "gr8" could help, as suggested by [7]. On the other hand, Top2Vec does not require such preprocessing, but it might still struggle with noisy or irrelevant data such as slang or misspellings, which are prevalent in tweets. Additionally, removing low-value tokens based on TF-IDF scores could help reduce the occurrence of "noise" words, further refining the model's relevance.

Moreover, Tweets, as short texts, inherently contain limited context, making topic extraction challenging. Both LDA and Top2Vec may fail to capture complex topics or subtle nuances in short content, leading to fragmented or overlapping topics.

In addition, Top2Vec relies on pretrained embeddings (Universal Sentence Encoder) to generate semantic representations of words and documents. The performance of these embeddings can be highly dependent on the domain and data they were trained on [1].

As a future research topic, one may incorporate geolocation data from tweets, allowing for an analysis of topics based on their geographic origin. Additionally, combining topic modeling with sentiment analysis could provide insights into how sentiment aligns with specific topics. This approach could reveal how certain words, such as "hoax" in negative tweets or "hope" in positive ones, are associated with different topics based on sentiment. This could lead to more nuanced topic differentiation that accounts for both word choice and emotional tone.

Table 4: Topics identified by Top2vec

First Topics identified by LDA			Hierarchical Topic Reduction		
No.	Topic/content	Keywords	No.	Topic/content	Keywords
1	Politicians climate change activism	denier, algore, globalwarming, skeptics, greenpeace, scientist, consensus, foxnews, hoax, tillerson, greenhouse, environmenta, putin, renewables, gore, macron	1	Climate Science	deniers, algore, globalwarming, greenpeace, greenhouse,hoax, scientist, foxnews, consensus, renewables, environmental, natgeo, putin, exxonmobil, exxon
2	Melting Glaciers	algore, greenhouse, arctic, antartica, snowing, hoax, temperatures, winter, melting, deforestation, cooling, extintion, scientisct, wildfires, freezing	2	Political Discourse	algore, deniers, trump, tillerson, epa, globalwarming, hillary, macron, warming, gop, putin, skeptic, potus, liberals, greenpeace, foxnews, irma, hoax, exxonmobil, conservatives, consensus, scientist, sanders, republicans, bernie
3	American climate change activism	trump, algore, tillerson, denier, globalwarming, climate, hillary, potus, macron, irma, putin gop, greenpeace, maga , hoax	3	Polution	deniers, algore, denier, warming, climate, globalwarming, skeptics, greenhouse, greenpeace, hoax, antarctica, arctic, scientist, hurricanes, temperatures, renewables, pollution, environmental, extinction, earth, gore, exxonmobil, natgeo, global
4	General Climate change	algore, deniers, globalwarming, greenpeace, greenhouse, climate, skeptics, hoax, snowing, cooling, arctic, temperatures, deforestation, environmental, foxnews, putin	4	Global Impacts	deniers, denier, climate, algore, globalwarming, warming, greenpeace, skeptics, hurricanes, skeptic, irma, deforestation, greenhouse, renewables, arctic, natgeo, reefs, epa, scientists, scientist, environmental, weather, hoax, climatechange, antarctica, extinction, catastrophic, tillerson, temperatures, gore, macron, consensus, putin

Table 5: Topic coherence scores

Model	C_v
LDA	0.224
top2vec	0.379

REFERENCES

[1] Dimitar P Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[2] Dimo Angelov and Diana Inkpen. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, 2024.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9: 1–20, 2019.

[5] Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498, 2022.

[6] Maciel M Queiroz. A framework based on twitter and big data analytics to enhance sustainability performance. *Environmental Quality Management*, 28(1):95–100, 2018.

[7] Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110:298–310, 2018.