Application of Traditional and Deep Learning Algorithms in Sentiment Analysis of Global Warming Tweets

Dragana Nikolova(✉), Georgina Mircheva, and Eftim Zdravevski

Faculty of Computer Science and Engineering, Ss Cyril and Methodius University, Skopje, Macedonia
dragana.nikolova.1@students.finki.ukim.mk,
{georgina.mircheva,eftim.zdravevski}@finki.ukim.mk

Abstract. The Earth's surface is continuously warming, changing our planet's average balance of nature. While we live and experience the impacts of global warming, people debate whether global warming is a threat to our planet or a hoax. This paper uses relevant global warming tweets to analyze sentiment and show how people's opinions change over time concerning global warming. This analysis can contribute to understanding public perception, identify misinformation, and support climate advocacy. This paper proposes a data processing pipeline encompassing traditional and deep learning based methods, including VADER, TextBlob, Doc2Vec, Word2Vec, LSTMs, to name a few. The extensive testing shows that the combination of document embeddings and neural networks yields the best results of up to 97% AUC ROC and 93% accuracy. The findings enable the comprehension of human attitudes and actions related to this worldwide issue in production environments.

Keywords: natural language processing · sentiment analysis · global warming · machine learning · deep learning

1 Introduction

Increased temperatures, severe storms, drought, rising oceans, loss of species, and health risks are threats imposed by global warming. Our planet is at risk, and we must fight for the health of our planet. But to fight, people must understand why and how global warming is happening and what we can do to slow down the process of it. The emission of greenhouse gases is one of the biggest drivers of global warming, of which more than 90% is the emission of carbon dioxide [17]. In the emission of CO_2, we humans influence with burning fossil fuels for energy consumption, transportation, deforestation, manufacturing [3]. Knowing the causes of global warming helps us understand climate change and how we can all contribute to avoiding worse harm.

Classifying tweets based on sentiment analysis in the context of global warming is an interesting and beneficial activity with potential impacts on climate

P. J. Coelho et al. (Eds.): GOODTECHS 2023, LNICST 556, pp. 49–61, 2024.
https://doi.org/10.1007/978-3-031-52524-7_4

communication, policy development, and public awareness. The results obtained from such an analysis can provide valuable insights for addressing one of the most pressing challenges of our time.

In this paper, we conduct sentiment analysis on tweets related to global warming. Twitter serves as a platform where individuals express their opinions in raw and informal texts. Analyzing tweets has been a prevalent practice over the years. Notably, a relevant study [14] addresses a similar challenge using tweets, focusing on abusive language detection. Another study [15] employs distant supervision techniques and word embeddings on tweets with additional unlabeled data to enhance stance classification. Sentiment analysis on tweets is used even in health care, as explained in a paper [6] that investigates public sentiments surrounding COVID-19 vaccination using Twitter data. By employing natural language processing, the authors identify different sentiment categories and analyze the sentiment trends over time and in response to vaccination-related events.

Our approach starts with preprocessing the tweets and extracting features from the same. Starting with traditional supervised models, we compare the results of Naïve Bayes, Decision Trees, and Random Forest models. Then we train a neural network, comparing the results of using word embeddings and document embeddings. Additionally, we perform an unsupervised clustering model.

The paper's organization is as follows. We begin by reviewing the relevant literature on sentiment analysis for global warming in Sect. 2. Next, in Sect. 3, we describe the dataset used for our analysis. Section 4 provides a detailed overview of the preprocessing steps undertaken. The models trained and the methodology employed are presented in Sect. 5. In Sect. 6, we present the outcomes of our sentiment analysis for global warming-related tweets, followed by a discussion of the findings in Sect. 7. Finally, in Sect. 8, we present our conclusions and summarize the key takeaways from this study.

2 Related Work

Sentiment analysis allows us to gain insights into specific topics and therefore it is broadly used, especially in social media monitoring. With sentiment analysis on global warming tweets, we gain an understanding of the public's perception of global warming and in which hands is our planet.

In [19], seven lexicon-based approaches were used for sentiment analysis on climate change. Namely, SentiWordNet, TextBlob, VADER, SentiStrength, Hu and Liu, MPWA, and WKWSCI were used in combination with classifiers such as Support Vector Machine, Naïve Bayes, and Logistic Regression. They have reached the best accuracy using hybrid TextBlob and Logistic Regression. Additionally, they discovered that using lemmatization improved the accuracy by 1.6%.

In 2017, a paper for sentiment analysis on global warming was published, where participants proved that positive tweets are increasing after 2014 [16]. They have used global warming tweets worth ten years and applied Naïve Bayes,

Support Vector Machines, and Linear Support Vector classification (SVC). They reached the best accuracy using Linear SVC with unigram and bigram combinations.

The same year a paper was published for real-time Twitter sentiment analysis using an unsupervised method [5]. They have used a variety of dictionaries to calculate the polarity and intensity of opinion in a tweet. With great focus on the preprocessing part, they established slang correction, acronyms replacement, POS tagging, phonetic inconsistencies correction, and noun standardization. With the unsupervised approach, they developed a system for visualizing opinions on tweets in real-time. In our paper, we also focus on the preprocessing part which is known to have a huge effect on the results, as explained in a paper [10] where they go in-depth about how the selection of appropriate preprocessing methods can enhance the accuracy of sentiment classification.

In [18] the authors explore topic modeling and sentiment analysis of global warming tweets using big data analysis. This paper analyses the discussion of global warming on Twitter over a span of 18 months and concludes that there are seven main topics. The sentiment analysis shows that most people express positive emotions about global warming, though the most evoked emotion found across the data is fear, followed by trust. Another recent study [8] deals with the topic of sentiment analysis on global warming tweets using naïve Bayes and RNN.

3 Dataset

Our analysis will involve two types of data. First, we will preprocess the text data, which will be used to train an unsupervised model through clustering. Next, we will perform supervised learning using neural networks. Once we have identified the highest accuracy model, we will use it to analyze how people's understanding of global warming is changing over time.

The first dataset contains labeled global warming tweets. This dataset was downloaded from Kaggle [1]. These tweets are between April 27, 2015 and February 21, 2018. In total, 43,943 tweets are available. For each tweet, the identifier, text, and sentiment are available.

The second dataset contains 308,371 unlabeled tweets published between September 21, 2017 and May 17, 2019 from the Twitter API, which is publicly available [11]. Since only tweet identifiers are available from this dataset, we used tweepy python library to retrieve text and publication date for each tweet.

The main goal is to perform classification, where we have negative sentiment or class 0, and positive sentiment or class 1. We will use the processed texts as input to the machine learning models and the dates will serve to analyze the results and see if people's opinion about global warming is gradually changing to positive or if people see it as a hoax. To present a more comprehensive picture of public sentiment, future studies could incorporate neutral sentiment analysis alongside positive and negative sentiment.

4 Data Preprocessing

The initial stage in our analysis involves preprocessing textual data. The aim here is to prepare the text in a manner that can be utilized as input for a machine learning model. We have outlined the steps involved in the process in Fig. 1. By following this ordered sequence of preprocessing steps we ensure that the text data is optimally prepared for analysis.

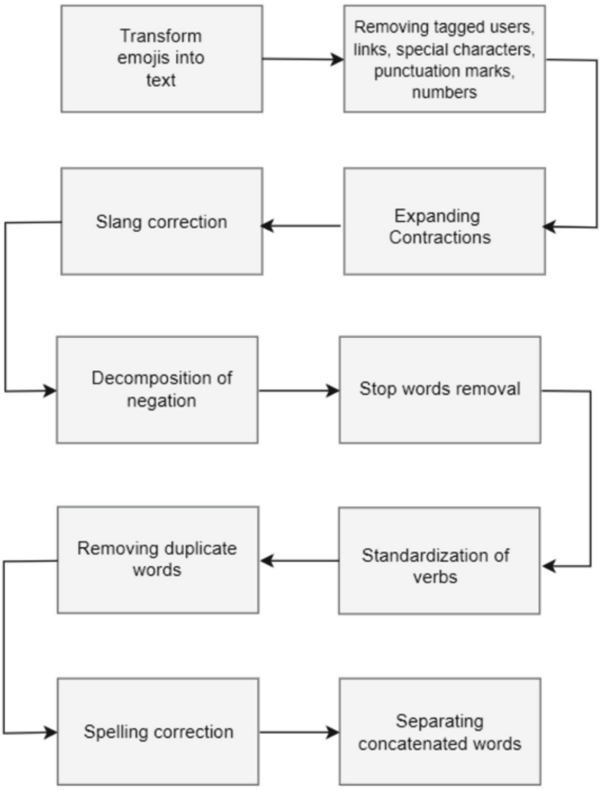

Fig. 1. Text Processing Steps.

4.1 Transforming Emojis in Text

The first step in preprocessing is the detection of emojis and their transformation into appropriate text. Emojis can have a significant impact on sentiment analysis for several reasons as adding context, or introducing subjectivity as explained in a study [20]. We iterate through each sentence and each word and see if the characters match any of the existing emojis in the python library emoji, which are then transformed into words.

4.2 Text Cleaning

The next step in our analysis is cleaning the text data. This involves removing tagged users, links, special characters, punctuation marks, and numbers. Furthermore, we ensure that each word in the dataset is represented in lowercase letters to avoid any discrepancies caused by variations in capitalization. By performing this cleaning process, we can improve the quality and consistency of our data, thereby enhancing the accuracy of our subsequent analyses.

4.3 Expanding Contractions

Shortened form of a group of words forms a contraction. When a contraction is written in English, letters are replaced by an apostrophe. Expansion of these words is achieved by using the python library contractions. By using the fix function from this library, the words are simplified and an example of that is given in Table 1, where on the left side we have contractions and on the right side we have the extended words. The examples are from our dataset.

4.4 Slang Correction

In the dataset we have spotted slang abbreviations which consist of a couple of letters. For figuring out which words are contained in the abbreviations, we needed to translate each slang in its long form. For translating the slangs, we have first extracted all existing English slangs from a web page [4]. Using the Python library BeautifulSoup we iterate through the web page HTML and store each slang and the translations in a dictionary. Then we compare each word in our dataset with the stored slang and replace each with their correct meaning.

4.5 Decomposition of Negation

To deal with negation, antonym replacement using WordNet was used to replace the word that comes after the words “not” and “never”. An example of such a transformation is given in Table 2 where the sentences have the same meaning before and after the transformation, but words on the right side have negation words obtained as antonyms of those on the right side. The purpose of this step is to give more meaning to negative words. This step positively influenced our results by enhancing the accuracy of sentiment classification and reducing ambiguity.

4.6 Stop Words Removal

English stop words are removed using the python library nltk. By ignoring these words, we ignore giving meaning to words that are used often in English sentences, for example conjunctions and pronouns, such as “I”, “me”, “myself”, “we”, “you”, “because”, etc.

4.7 Verb Standardization

We perform verb standardization to represent all verbs in the future or past tense in their lemma. In morphology and lexicography, a lemma is a canonical form or a form used in dictionaries. The en_core_web_sm [2] module from spacy python library is used for this purpose.

4.8 Spelling Corrections

The dataset is downloaded from twitter in the form of tweets where people express their opinion, in our case it is the opinion on global warming. Because people have the absolute freedom to write their opinions, there may be spelling mistakes. Such mistakes in the words contribute to the fact that the words themselves do not exist in the dictionary of the English language, and thus do not have a role and meaning in the sentiment analysis. To deal with this, we introduce automatic word spelling correction using the spell function from the python autocorrect library. An example of corrected words from the data set is given in Table 3.

4.9 Separating Concatenated Words

During the text preprocessing, we encountered instances where multiple words had been concatenated into a single word, which does not exist in the English language and thus lack a clear meaning. To address this issue, we developed a method in which we iterated through each concatenated word letter by letter and checked if the words exist using the check function from the enchant python library. An example of how we separated concatenated words in our dataset can be found in Table 4. By implementing this step, we enhanced the tokenization process, ensuring that each word in a concatenated sequence is treated as a separate entity. This, in turn, led to more accurate and meaningful text analysis.

Table 1. Expanding Contractions

Contractions	Extended words
Here's my harsh reality	Here is my harsh reality
Today I'm much more worried	Today I am much more worried
It's a lesson	It is a lesson
We wouldn't have forest fires	We would not have forest fires

5 Machine Learning Models

We will elaborate on three different strategies for solving sentiment analysis using the labeled data. In each strategy we are using set of lexicon-based features

Table 2. Decomposition of Negation

Before decomposition	After decomposition
Has not earned any votes	Unearned any votes
He not accept the evidence	He refuse the evidence
Not prepared for global warming	Unprepared for global warming

Table 3. Spelling Correction

Wrong spelling	Correct Spelling
possition	position
individuen	individual
beweging	begging
earthi	earth
kmart	smart
healthcaren	healthcare
societys	society
crite	write

Table 4. Separating Concatenated Words

Concatenated words	Separated words
urbanplanning	urban plan
resillienceforall	resilience for all
greatbarrierreed	great barriers reef
climatechangeisreal	climate change is real
didyouknow	did you know
savethereef	save the reef
natureseedle	nature seed
recordbreak	record break
scientistgobhi	scientist gob hi

and we are performing word and sentence embeddings. Using the best accuracy model, we will predict the unlabeled data and give insights in the results. Additionally, we will perform clustering on the unlabeled data as an unsupervised learning. For evaluation we used accuracy, macro average F1-score calculated using precision and recall, and AUC ROC (Area Under the ROC curve).

5.1 Classification

VADER + TextBlob + Traditional Models. First, for each of the tweets, we find VADER (Valence Aware Dictionary and Sentiment Reasoner) features. VADER is a lexicon and rule-based sentiment analysis tool specifically attuned for sentiment expressed in social media [7]. It uses a combination of list of lexical features that are labeled according to their semantic orientation as positive or negative. For each tweet we get how positive it is, how negative it is, and compound metric that calculates the sum of all ratings that are normalized between –1 (extremely negative) and +1 (extremely positive). Then for each tweet we extract the polarity and subjectivity using TextBlob. Polarity is the output from TextBlob that lies between –1 and 1, where –1 refers to a negative feeling and +1 refers to a positive feeling. Subjectivity is the output that lies between 0 and 1 and refers to personal opinions and judgements [13]. With VADER and TextBlob we have 5 features in total.

Only 9.08% tweets of the whole dataset are labeled as negative tweets, which makes our dataset unbalanced. To solve that, the data set was balanced using oversampling with the SMOTE python library. With oversampling we duplicate the data from the minority class, which in our case is the class with tweets labeled as negative.

Using VADER and TextBlob metrics we trained classifiers whose accuracy metrics are given in Table 5. We trained the classifiers on the labeled data, of which 20 used for testing. The highest accuracy was obtained with Random Forest classifies.

VADER + TextBlob + Doc2Vec + Neural Network. To represent each tweet numerically, we employed the use of Doc2Vec. By doing so, we were able to map each tweet to a vector of size 100. We then supplemented these vectors with additional features from VADER and TextBlob, which increased the vector size to 105 for each tweet. These enhanced vectors were then used as inputs for a sequential neural network. This approach demonstrates the importance of combining diverse techniques in order to achieve best possible results. 60% of the dataset was used for training, 20% for validation and 20% for testing. With this model we have reached maximum accuracy of 92.9%, as indicated in Table 6.

VADER + TextBlob + Word2Vec + Neural Network. We constructed a neural network that included an LSTM layer and Word2Vec vectors as the embedding layer. Each word was represented by a vector of size 200. In addition to the embedding vectors, we also utilized VADER and TextBlob features as additional input to the neural network. To train and evaluate the model, 60% of the data was reserved for training, 20% for validation, and 20% for testing. After testing, we found out that with this model we achieved the minimum accuracy for all evaluation metrics.

5.2 Clustering with K-Means

We conducted clustering on the 308,371 unlabeled tweets in our dataset. For each tweet, we extracted VADER and TextBlob features, and added a TF-IDF vector to each tweet. TF-IDF estimates how relevant a word is to a document in a collection of documents. If a word appears frequently in one document but not in others, it is likely to be highly relevant to that document. To further enhance our analysis, we also represented each tweet as a vector of size 100 using Doc2Vec. By combining these features (VADER, TextBlob, TF-IDF, Doc2Vec), we trained a clustering model (kmeans) with k = 2, which resulted in two categories: positive and negative tweets. Through this approach, we were able to classify 28,919 negative tweets out of 308,371.

Table 5. Accuracy Scores For Traditional Models

Model	Accuracy	Macro Average F1-score	AUCROC
Bernoulli Naive Bayes	54.58%	54%	54%
Complement Naive Bayes	53.59%	53%	53.57%
Multinomial Naive Bayes	53.63%	53%	53%
Kneighbors	71.51%	71%	71%
Decision Tree	77.37%	77%	77.3%
Random Forest	78.61%	79%	78.6%
Logistic Regression	54.11%	54%	54%
Multi-layer Perceptron	59%	59%	58.98%
Ada Boost	62.20%	62%	62.18%

6 Results

Figure 2 presents the percentage of negative tweets from the total number of tweets, grouped by year. To generate Fig. 2, we used labeled data and supplemented it with our own classification of unlabeled data. Our analysis reveals that the number of negative tweets in 2017, 2018, and 2019 is considerably lower than in 2015 and 2016, as shown in the graph. Therefore, it can be inferred that people's attitudes towards global warming have become more positive in recent years, potentially indicating a shift towards more proactive measures to address the issue.

Table 6. Final Accuracy Scores

Model	Accuracy	Macro Average F1-score	AUC ROC
VADER + TextBlob + Doc2Vec + Neural Network	92.9%	93%	97.2%
VADER + TextBlob + Random Forest	78.61%	79%	78.6%
VADER + TextBlob + Word2Vec + Neural Network	59.04%	59%	58.69%

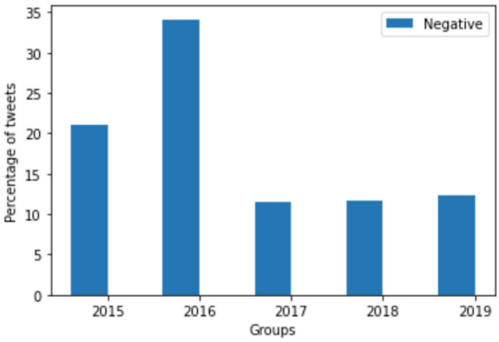

Fig. 2. Percentage of negative tweets.

It's important to note that a reduction in negative tweets may not solely signify a positive shift in public perception of global warming. Several factors could contribute to this trend, which requires further investigation. These factors may include overexposure and decreased attention.

7 Discussion

One of the limitations of the paper is that only two classes are considered, namely positive and negative tweets, while neutral could also be equally important. In future research, it is essential to explore the incorporation of a much broader semantic representation of language. This can be achieved by leveraging advanced approaches like deep learning architectures and state-of-the-art language models.

Recent advancements in natural language processing, such as the XLNet [9] model, have shown promising results in training high-performing sentiment classifiers. XLNet, a transformer-based model, addresses the limitation of BERT's unidirectional context by employing permutation-based training. By considering

all possible permutations of input tokens, each token can effectively attend to any other token in the sequence, capturing a more comprehensive context for sentiment analysis [21].

Another model that holds potential for sentiment analysis is RoBERTa, a variant of BERT. RoBERTa fine-tunes the pretraining process to enhance performance. This involves utilizing more data, employing larger batch sizes, and eliminating the next-sentence prediction task present in BERT. These improvements contribute to RoBERTa's ability to achieve better results on various NLP tasks, including sentiment analysis [12].

However, it's important to acknowledge that models like XLNet and RoBERTa often require large datasets for effective training, which can be a challenge in certain domains, such as our case of analyzing social media data related to global warming. As the discussions surrounding global warming have gained momentum in recent years, it has become a pertinent topic on social media. However, due to the relatively recent surge in these discussions, the availability of labeled data remains limited, hindering the development of sentiment analysis models for this specific context.

To address this scarcity of labeled data, our study focused on carefully preprocessing the available data. Additionally, we adopted a combination of multiple approaches and incorporated techniques tailored for analyzing social media content, such as VADER.

In conclusion, while the paper highlights some crucial limitations, such as the exclusion of neutral sentiment and the challenges of limited data availability, it provides a foundation for future research to explore more advanced language models and innovative strategies to improve sentiment analysis on social media data related to global warming. By leveraging the power of transformer-based models like XLNet and RoBERTa, and by adapting to the unique characteristics of social media language through techniques like VADER, we can enhance sentiment analysis and gain valuable insights into public perceptions and attitudes towards global warming.

8 Conclusion

The world is currently experiencing the effects of global warming, which is partially caused by human activities that emit greenhouse gases, such as carbon dioxide. As a result, the number of tweets about global warming has been on the rise, and there is a sharp divide between those who believe in its existence and those who deny it. To better understand this phenomenon, we developed machine learning models that classify global warming related tweets using both labeled and unlabeled data. After testing various methods, we found that the best results were achieved using document embeddings and neural networks. By harnessing the power of machine learning, we can better understand the patterns of human behavior and opinions surrounding this global concern.

Acknowledgements. This article is based upon work from COST Action CA19121 (Good Brother: Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living), supported by COST (European Cooperation in Science and Technology). COST is a funding agency for research and innovation networks. COST Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career, and innovation. More information at https://www.cost.eu.

References

1. Climate change. https://www.kaggle.com/datasets/edqian/twitter-climate-change sentiment-dataset
2. English pipeline. https://spacy.io/models/en
3. Global emissions. https://www.c2es.org/content/international-emissions
4. Internet slang. https://www.internetslang.com/
5. Azzouza, N., Akli-Astouati, K., Oussalah, A., Bachir, S.A.: A real-time twitter sentiment analysis using an unsupervised method. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, pp. 1–10 (2017)
6. Dandekar, A., Narawade, V.: Twitter sentiment analysis of public opinion on COVID-19 vaccines. In: Bansal, J.C., Engelbrecht, A., Shukla, P.K. (eds.) Computer Vision and Robotics. Algorithms for Intelligent Systems, pp. 131–139. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-8225-4_10
7. Hutto, C., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
8. Joy, D.T., Thada, V., Srivastava, U.: Sentiment analysis on global warming tweets using Naïve Bayes and RNN. In: Nanda, P., Verma, V.K., Srivastava, S., Gupta, R.K., Mazumdar, A.P. (eds.) Data Engineering for Smart Systems. LNNS, vol. 238, pp. 225–234. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2641-8_21
9. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimed. Tools Appl. **82**(3), 3713–3744 (2023)
10. Krouska, A., Troussas, C., Virvou, M.: The effect of preprocessing techniques on twitter sentiment analysis. In: 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1–5 (2016). https://doi.org/10.1109/IISA.2016.7785373
11. Littman, J., Wrubel, L.: Climate change tweets ids. In: GWU Libraries Dataverse. Harvard Dataverse (2019). https://doi.org/10.7910/DVN/5QCCUU
12. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
13. Loria, S., et al.: Textblob documentation. Release 0.15 **2**(8), 269 (2018)
14. Markoski, F., Zdravevski, E., Ljubešić, N., Gievska, S.: Evaluation of recurrent neural network architectures for abusive language detection in cyberbullying contexts. In: Proceedings of the 17th International Conference on Informatics and Information Technologies. Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science (2020)
15. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Trans. Internet Technol. (TOIT) **17**(3), 1–23 (2017)

16. Mucha, N.: Sentiment analysis of global warming using twitter data. In: Computer Science Masters Papers. North Dakota State University (2018)
17. Olivier, J.G., Schure, K., Peters, J., et al.: Trends in global co2 and total greenhouse gas emissions. PBL Net. Environ. Assess. Agency **5**, 1–11 (2017)
18. Qiao, F., Williams, J.: Topic modelling and sentiment analysis of global warming tweets: evidence from big data analysis. J. Organ. End User Comput. (JOEUC) **34**(3), 1–18 (2022)
19. Sham, N.M., Mohamed, A.: Climate change sentiment analysis using lexicon, machine learning and hybrid approaches. Sustainability **14**(8), 4723 (2022)
20. Shiha, M., Ayvaz, S.: The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng. (IJCEE) **9**(1), 360–369 (2017)
21. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. **32** (2019)