Unsupervised Semantic Sentiment Analysis of Tweets on NetZero and Carbon Offsets

Danila Rozhevskii, Prof. Jay S Golden[1,✉]

[1]Carbon Finance Team, Dynamic Sustainability Lab (DSL), Syracuse University, Syracuse, United States

The project's goals are to analyze unlabeled tweets in English that can help understand public opinion on blockchain and its role in the transition to a net-zero economy around the world and find possible correlations between the geo-political and demographical background of those Tweets. The results could help reflect on the existing approaches to carbon offsets and net-zero strategies. First, the team collects data on tweets about either current general perception or blockchain's role in the transition to a net-zero carbon economy. Next, the project uses unsupervised learning (word2vec with KMeans) to split preprocessed data into a desirable number of sentiments by considering their context, keywords, and hashtags. Finally, the project uses Sentiment Intensity Analyzer (VADER) and LLM models such as BERT to compare and adjust the sentiment polarity of every tweet to the common average, which could help avoid contextual misunderstandings and supplement each other.

Sustainability | Unsupervised Sentiment Analysis | NLP | LDA

Introduction

At Dynamic Sustainability Lab, the team is trying to understand the public response to recent innovations in carbon credits technologies, such as the use of blockchain. If successful in figuring out what affects people's trust and perception over the last two years, the lab could create high-quality recommendations for businesses in various fields.
As one of the most active platforms for quick discussions, Twitter was chosen to be the source of the data. The size of tweets enables researchers to quickly analyze the change of opinion across different periods of time and various regions. The challenge is in clustering these unlabeled tweets into positive, negative, and neutral categories since the dataset has a very specific context to it, the Net Zero strategies, and carbon offsets discussions.
Keyword-based algorithms for sentiment detection, such as VADER, are able to approximate the sentiment of a tweet based on a proportion of keywords, which ignores the relationship between those words and their contextual meaning of them. Also, VADER suffers from not understanding sarcasm and misspelled words. Large Language Models (LLMs) such as BERT are pre-trained on large corpora of words to effectively analyze complex sentences and determine their sentiment in accordance with the relationship between individual sentences and words. However, this approach also suffers from contextual misalignment; when the tweet only references certain events, organizations, or people, there is no way for pre-trained models to approximate the sentiment without being contextually aware. Lastly, the K-means algorithm is a popular approach when dealing with clustering problems. The project assumes that the heavily contextualized data would have tweeted about certain topics or certain mentions being clustered in groups that can be determined. These centroids would not have as accurate sentiment approximation as pre-trained models, however, they could tune those on.

Previous Works

In the past, there was a lot of research done in the area of sentiment analysis on various media. In their paper about sentiment analysis of microblogging data Agashini. V. Kuma and K. N. Meera use the K-Means clustering algorithm with the goal of researching the conditions and parameters that affect the accuracy of the algorithm in the prediction of the positive, negative, or neutral sentiment of the given data. The authors used Elbow Analysis to find the optimal number of clusters, which turned out to be 4 instead of the predicted 3. The paper exposes the clash of the K-Means method with intended results and describes the commonality of unsupervised sentiment analysis.(3) One of other works was done by Sharon Susan Jacob and Dr. R. Vijayakumar. In their paper, the authors implement various algorithms Bisecting K-Means, and modifying form Gaussian mixture model to compare their sentiment analysis performance on big data [1]. Based on the results of the research, the most effective method was the Gaussian mixture model-based Mean-shift clustering approach, which yielded 87.3% accuracy in comparison to 57.7% by Bisecting K-Means. These results could be used to improve the results of this paper in the future. (1) Finally, Anil Kumar in his paper on Mining Reasons for Sentiment Variation on Twitter using Cosine Similarity Measurement proposes the COsine SImilarity MesuremeNT (COSINT) method for LDA to extract Reason from Reason Candidates more precisely and thus helps explain the variation of the sentiment. In my research, I also use the LDA model in order to find the context behind the sentiment change of tweets about carbon credits over the course of 2021-2022 years. However, the proposed method improves the precision of the mined reasons from the data and is easier to understand. The results of this research could be used in the future to improve the topic modeling of tweets on carbon credits (2).

Approach

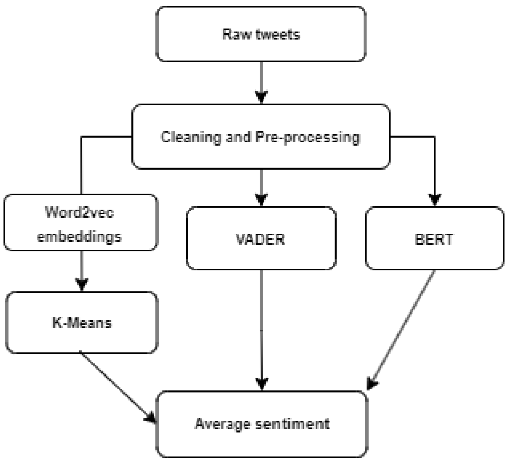

Fig. 1. Project workflow schema.

In this paper, the project collects and analyzes the contents of around 225,098 English tweets that discussed either the current general perception of carbon credits or the blockchain role in the transition to a net-zero carbon economy in the past 2 years period (1/1/2021 – 10/31/2022).

After pre-processing and vectorizing the contents of the tweets using the Word2vec model, the data is passed through 3 algorithms: custom-trained K-Means clustering and pre-trained VADER and BERT, which calculate the sentiment polarity of the tweet: positive, negative, or neutral. As a result, the combination of 3 different techniques gets the average result from them as the final sentiment. Tested on arbitrarily marked data, the approach had an accuracy of 60%.

I also investigated the common topics among the collected tweets over the years using the Gensim Linear Discriminant Analysis(LDA). The best model for each year exceeded the 0.40 coherence score, which shows relatively poor interpretability of the results. So as the last step, I used ChatGPT to help generate complete topic sentences from LDA output and improve the final results significantly.

Datasets

A. Training. The project's training (target) data was collected from Twitter, one of the biggest social media platforms with opinions posted daily on almost anything. The short nature of each tweet allows researchers to quickly analyze and track the dynamic of opinions of millions of people on the platform. Around 400,000 tweets were scraped from the front page that had hashtags #carboncredits, #carbonoffets, or #netzero. After data cleaning and English language filtering, there were 225,098 tweets left for analysis.

B. Testing. To test the accuracy and the effectiveness of the proposed method, I test it on the arbitrary data of 929,544 random tweets from Kaggle scrapped using tweepy which is used to access the tweets using Twitter API. The tweets have been annotated with 4 different categories(positive, negative, uncertainty, and litigious) and they can be used to detect sentiment. However, to match the training data, I only left positive, negative, and neutral (uncertainty) tweets for testing.

Methods

C. Sentiment Analysis: K-Means. First, I wanted to explore one of the most popular techniques for unsupervised sentiment analysis, K-Means clustering. The size of the dataset made it possible to create a large enough dictionary of words for the Word2vec model. I decided to work with 3 clusters: positive, negative, and neutral. The neutral cluster is supposed to collect possible spam tweets or tweets with not enough information for humans to determine the sentiment.

Fig. 2. K-Means sentiment results for 2021-2022.

According to the results of the K-means model in Figure 2, there is a very sharp rise in positively identified tweets in October 2021 and so on. In 2022, the positive tweets count did not drop much and was dominating throughout the year up to next October. Of course, the K-means results can not be as accurate as those of Large Language Models and key-word based algorithms like VADER. However, K-Means is a good method for topic clustering. I can argue that it is possible to cluster tweets with a certain number of topics as positive or associate topics with a certain sentiment such as negative or positive. It is not always correct, but I think it can capture the trend as well as the context of the tweets.

D. Sentiment Analysis: VADER (Valence Aware Dictionary and sEntiment Reasoner). Second, I imported and applied the VADER algorithm on the same per-processed text as that given to K-Means. VADER is a key-based algorithm for sentiment analysis, which means it has its own dictionary of words for sentiment classification (5).

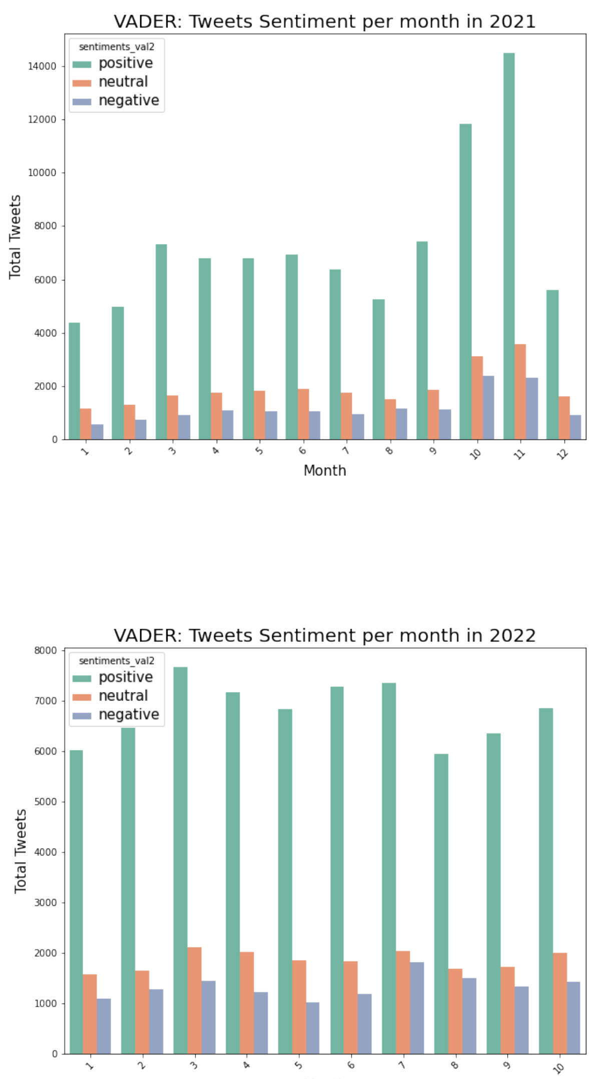

Fig. 3. VADER sentiment results for 2021-2022.

From Figure 3, it is clear that the tweets have a similar trend as the one K-Means algorithm produced, which adds more credibility to the latter. There was also an increase in positive tweets with a sharp rise around October 2021 and staying the same. VADER is correct in identifying key positive or negative words and evaluating each tweet based on the compound score of all the words in it. However, VADER can not learn and associate certain topics as positive or negative, since it only applies a pre-trained model.

E. Sentiment Analysis: BERT (Bidirectional Encoder Representations for Transformers). Finally, I decided to run BERT, which is a model with pre-trained language representations that has an internal library for sentiment analysis (6). BERT is able to identify sentiment based on common keywords, sentence structure, as well as the context of each tweet based on the generated embeddings. By design, BERT is able to identify either positive or negative tweets. It is one of the most advanced unsupervised methods for sentiment analysis yet and I wanted to see how similar its results are to the K-Means model.

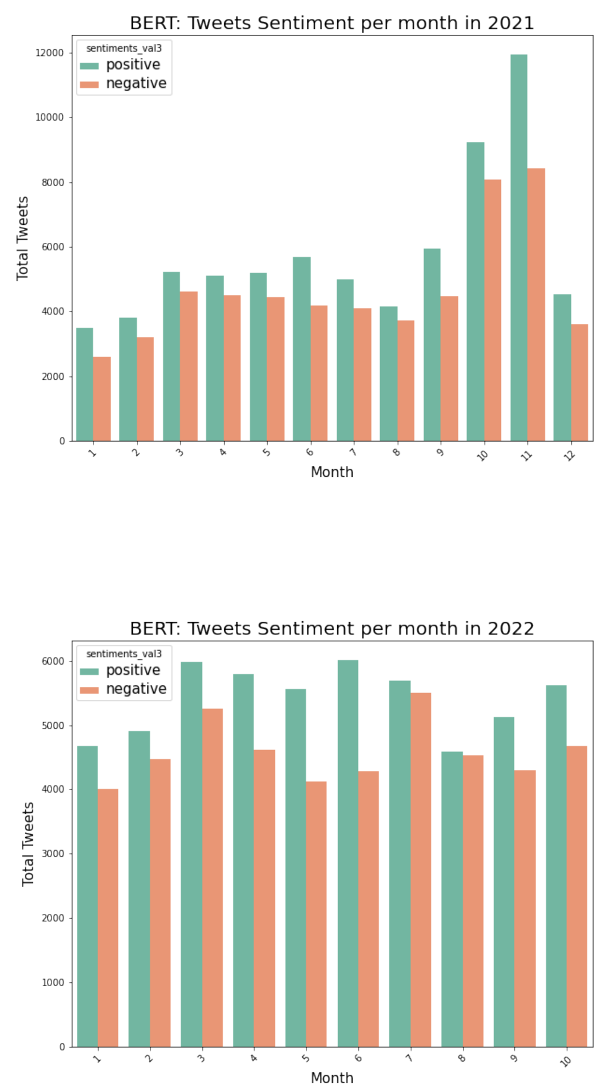

Fig. 4. BERT sentiment results for 2021-2022.

From Figure 4, it is clear that it repeats the same sentiment trend as VADER and also provides even more credibility to the K-Means method. All 3 algorithms show a sharp increase in positive tweets. Since BERT only segregates positives and negatives, the neutral tweets are combined with negatives. This is why, it looks like BERT identified more negatives.

F. Sentiment Analysis: Average of 3. I also combined the results of previous methods and calculated the average sentiment identified by them. It shows the same trend, but it is able to boost the number of negative tweets that were misidentified or simply missed by K-Means based on the results from VADER and BERT.

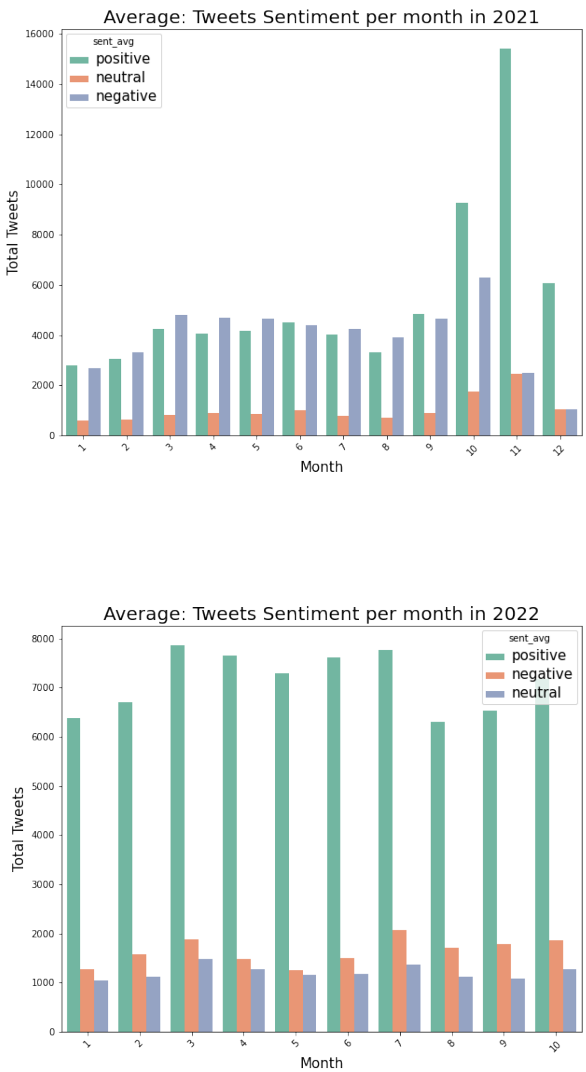

Fig. 5. Average sentiment results for 2021-2022.

Figure 5 shows the dynamics of the average sentiment of tweets based on the results of 3 algorithms. I wanted to test this approach against individual methods on a random set of 300,000 tweets. The testing set does include true sentiment labels, by performing the tests I could evaluate the actual accuracy of the results.

G. Accuracy Analysis. Based on the testing data in Table 1, the approach of taking an average reaches 58% accuracy.

Method	Precision	Support
K-Means	0.44	299,926
VADER	0.57	299,926
BERT	0.77	299,926
Average	0.58	299,926

Table 1. Accuracy table for all the methods on testing data.

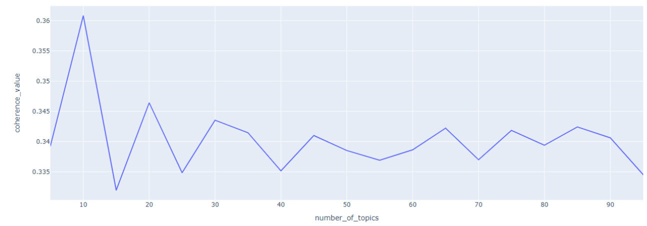

Fig. 6. Optimal number of topics for the LDA model.

Even though it is less accurate than BERT, it might work around the context of the tweets better. Since the carbon credits topic is quite narrow, there could be certain people mentioned most often in a positive or negative context. That is what, in theory, K-Means can capture.
The overall results show an increase in positive opinions about carbon credits and the Net Zero policy since October 2021. Throughout the 2022 year, positive tweets were dominating over negative and neutral. I also wanted to investigate what kind of particular topics people were talking about during this period.

Topic Modeling

H. LDA: Gensim Linear Discriminant Analysis. The second goal of the research was to examine the topics people were talking about when they mentioned carbon credits or Net Zero. I used Linear Discriminant Analysis on the cleaned training data. Latent Dirichlet Allocation (LDA) is a Bayesian network that explains a set of observations through unobserved groups, and each group explains why some parts of the data are similar (4).

I. LDA: Hypertuning. To hyper-tune the model, I used several components or topics and the learning rate of the model. Those parameters control how many meaningful labels the model tries to learn from the data. The learning rate affects the convergence of the loss function and so influences the quality of discovered topics. To measure the performance of the model, I used perplexity and coherence scores. Perplexity metric is widely used for language model evaluation and is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance. The coherence score reflects the interpretability of the resulting topics or themes. The lower the score, the hard it is to interpret or understand topics.
Table 2 and Figure 6 show the results of the cross-validation with learning decay and a number of topics as tuning param-

eters. The most effective learning decay rate was 0.7 and the most optimal number of topics was 20 with the final coherence score of 43.5%.

Topics	Learning Decay	Perplexity	Coherence
10	0.7	-7.559	0.425
15	0.7	-7.613	0.434
20	0.7	-7.724	0.435
25	0.7	-7.789	0.420
30	0.7	-7.861	0.410
35	0.7	-7.943	0.405
40	0.7	-8.018	0.421

Table 2. Parameters hyper tuning for LDA model.

J. LDA: Improving the interpretability of topics with ChatGPT. In this section, I tried improving the coherence score of the final model. I chose the top 3 raw topics out of 20 that the LDA model generated. They do not sound coherent enough for business leaders or stakeholders to evaluate, so I decided to enhance their interpretability with the help of ChatGPT.

Number	Year	LDA Topic
1	2021	carbon energy help emission amp achieve zero reduce business industry.
2	2021	zero emission net 2050 target achieve 2030 reach commit plan.
3	2021	amp transition work read uk support role new energy climate.
1	2022	carbon use emission reduce offset credit help amp solution capture.
2	2022	zero net business help find journey sustainability carbon build support.
3	2022	climate change action amp transition business future work make an impact.

Table 3. LDA raw top 3 topics output for each year.

After getting raw topics, I used those raw topics in ChatGPT with the following prompt: "Paraphrase this sentence to increase its interpretability in the given domain". I collected the improved topics in the following table.

Number	Year	ChatGPT Topic
1	2021	Carbon-free energy sources can help reduce emissions and enable businesses and industries to achieve zero carbon goals.
2	2021	In order to achieve the net-zero emissions target by 2050, it is essential to commit to a plan that aims to reach the 2030 milestone.
3	2021	The transition towards new energy sources and the efforts to mitigate climate change require the UK's support and active participation.
1	2022	To achieve net-zero carbon emissions, it is essential to transition towards new, zero-carbon fuels and technologies, while also offsetting any remaining carbon emissions from fossil fuels.
2	2022	The journey towards sustainability and achieving net-zero carbon emissions requires businesses to find ways to build and support initiatives that help reach these goals.
3	2022	Businesses have the power to make a significant impact on the future by acting on climate change and working towards a smooth transition to a sustainable economy.

Table 4. Improved topics with ChatGPT for each year.

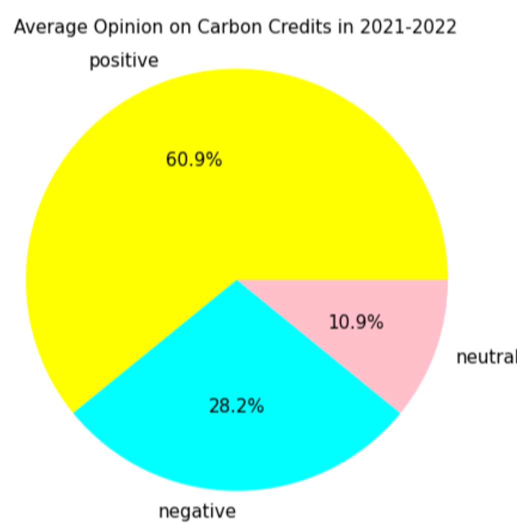

Fig. 7. Average sentiment results for 2021-2022.

Conclusions

The final results show a clear increase in positive feedback on carbon credits and net zero on Twitter from October 2021 and staying stable for most of 2022. According to Vienne's diagram in Figure, the biggest portion of total tweets from the 2021-2022 period was identified as positive. The method of the average result across various algorithms gives a believable output of unsupervised sentiment analysis and could be further improved with related works mentioned before.

The recommendations that I could make based on the results of the research and the next steps are the following:

1. Due to the new Twitter policy changes, I will have to modify the method for future tweets extraction. Alternatively, I planned to switch to another popular discussion board such as Reddit or Quora.

2. From the analysis, global events such as COP could trigger a lot of attention to the discussion of sustainable topics and NetZero strategy on the internet. DSL needs to work on promoting that kind of event more.

3. Considering the high activity of AMP Clean Energy in the UK, it would make sense to look for opportunities for DSL collaboration with them. I'm sure, we can find common ground.

Bibliography

1. Sharon Susan Jacob, Dr. R. Vijayakumar, "Sentiment Analysis Over Twitter Big Data Using Modified Meanshift Clustering Algorithm" *JOURNAL OF CRITICAL REVIEWS*, VOL 7, ISSUE 15, 2020.
2. Savitha Mathapati, Anil D, Tanuja R, S H Manjula, and Venugopal K R, "COSINT: Mining Reasons for SentimentVariation on Twitter using Cosine SimilarityMeasurement", *10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2018.
3. Agashini. V. Kumar, K. N. Meera, "Sentiment Analysis Using K Means Clustering on Microblogging Data Focused on Only the Important Sentiments " *10th IEEE International Conference on Emerging Trends in Engineering Technology Signal and Information Processing (ICETET-SIP-22)*, 2022.
4. David M. Blei, Andrew Y. Ng, Michael I. Jordan., "Latent Dirichlet Allocation" *Journal of Machine Learning Research 3 (2003) 993-1022*, 2003.
5. C.J. Hutto, Eric Gilbert., "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text" *Eighth International AAAI Conference on Weblogs and Social Media*, Vol. 8 No. 1 (2014).
6. CJacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" *Google AI Language*, 2019.