Alberto Scinetti (13115316)
Data Processing - Final Project

Project Description

This project will focus on learning about Vader sentiment analysis with its polarity scores and various algorithms for topic modelling with Python.

The project aims at answering the following question: *"How is the public opinion oriented towards climate related posts share by NASA on Facebook?"*.

With its own subquestions:

- *What is the public sentiment towards NASA's climate-related posts on Facebook?*
- What are the main topics being discussed within the analysed public?

1. Sentiment Analysis with Vader

The first part of the analysis focused on learning about sentiment analysis using Vader and its implementation on climate-related Facebook comments on posts shared by NASA from 2020 to 2023 found at https://www.kaggle.com/datasets/kanchana1990/public-opinion-on-nasas-climate-posts-fb-data.

Sentiment analysis is a natural language processing techniques employed to understand and identify sentiment of a written text; as a result a text can be marked as positive, negative or neutral (Dhaoui, Webster, and Tan 2017). Sentiment analysis is widely employed to analyse various type of user generated content such as news article, social media content or customer reviews. Many insights can be drawn by understanding the sentiment of a written text, from identifying customer preference to understand opinion on public matters (Dhaoui, Webster, and Tan 2017).

Vader, which stands for Valence Aware Dictionary and sEntiment Reasoner, is a relatively simple sentiment analysis tool designed to examine social media texts (Hutto and Gilbert 2014). Vader is a pre trained model which provides different polarity scores for neutrality, positivity and negativity of a given text. In Vader, each word already has a predefined score; words scores put together are able to identify the sentiment of a given text (Hutto and Gilbert 2014). Vader also takes into account the intensity of the sentiment, which can be determined by punctuation and capitalisations styles.

However, it is important to keep in mind, that Vader doesn't take into account the relation that words in the text have with each other (Hutto and Gilbert 2014).

Within this project Vader was employed to identify public sentiment within NASA's climate-related posts shared on Facebook between 2020 and 2023. The Vader pre trained model was imported on a Jupyter Notebook employed for the analysis. Each comment on NASA's post were collected in a data-frame, cleaned and marked with the Vader polarity tags of negative, positive and neutral with its related scores. Various graph were plotted to understand the distribution of negative, positive and neutral comments.

By looking at the graph on the Notebook a conclusion was drawn: the sentiment scores from the conducted Vader sentiment analysis indicate that most comments in the dataset are perceived as neutral or only slightly positive/negative. The comments with negative or positive polarity score indicators are only slightly impactful, contributing to a generally more neutral sentiment (strongly positive and strongly negative comments are relatively rare). Overall, the public sentiment towards NASA's climate-related posts on Facebook can be considered as neutral.

2. Topic Modelling

In the second part of the analysis, topic modelling techniques (Latent Dirichlet Allocation and Top2Vec) were adopted in order to identify the main topics being discussed within the analyzed space of public opinion regarding NASA's climate-related posts shared on Facebook between 2020 and 2023.

Topic Modelling is a a machine learning technique used to identify topics within a large amount of texts. It serves as a useful tool in analysis of textual data; its main aims are uncovering information, content recommendation, text summarisation and document organisation (Saxton 2018).

LDA (Latent Dirichlet Allocation) is one of many topic modelling algorithm which assigns topic to texts by analysing word frequencies. The LDA algorithms allows the user the number of topics that wants to be identified. Within the algorithm each corpus of text is assumed to be a mix of topics and

each topic is composed of different words. The algorithm then assigns words per topic with a score which identifies their relevance to the assigned topic (Seth 2021). The user needs to interpret the topic words and their scores to correctly tag each topic.

For this project the LDA algorithm was asked to identify the three main topics within comments on climate-related posts shared on Facebook between 2020 and 2023. The algorithm identified the following 3 topics: “Long-Term Planet Climate Change”, “Global Climate Change Implications” and “Environmental Impact of Carbon Emissions” (words related to each topic and their scores can be found in the notebook).

Top2Vec is another topic modelling algorithm which is easier to implement compared to LDA given that it does not need as much preprocessing of the data. The algorithm doesn’t need the user to specify the number of topics, it will identify this number itself. The algorithm works similar to LDA focusing more on clustering of words and automation of topics identifications (Lande 2022).

For this project Top2Vec identified the following 2 topics : “Climate Change Global Discussion” and “Climate change Implications and Impact of Carbon Emissions” (words related to each topic and their scores can be found in the notebook)

By looking at the topic identified, topic words and word scores conclusions can be made:

- The main topics identified for analysis by comparing and compiling the results were two. The first topic concerns general discussions about global climate change, focusing on temperature changes around the planet and their impact on global warming. The second one focuses on the implications of climate change, mainly concentrating on carbon emission levels and their effects.
- The main topics discussed within the analyzed comments are highly similar to each other and only slightly vary from each other, given the limitations of the analyzed data. A more extensive set of comments would be needed to closely identify topics within the scrutinised comment sections. This corpus of texts might not have been the best for conducting a topic modelling analysis given its limitations; however, it was still very insightful to learn about the algorithms and their processes.

Bibliography

Dhaoui, Chedia, Cynthia M. Webster, and Lay Peng Tan. 2017. “Social Media Sentiment Analysis: Lexicon versus Machine Learning.” *Journal of Consumer Marketing* 34 (6): 480–88. https://doi.org/10.1108/JCM-03-2017-2141.

Hutto, C., and Eric Gilbert. 2014. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.” *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 216–25. https://doi.org/10.1609/icwsm.v8i1.14550.

Lande, Janhavi. 2022. “Understanding Topic Modeling with Top2Vec.” *Medium* (blog). June 29, 2022. https://medium.com/@janhavil1202/understanding-topic-modeling-with-top2vec-cdf58bcd6c09.

Saxton, Micah D. 2018. “A Gentle Introduction to Topic Modeling Using Python.” *Theological Librarianship* 11 (1): 18–27. https://doi.org/10.31046/tl.v11i1.506.

Seth, Neha. 2021. “Topic Modeling and Latent Dirichlet Allocation (LDA) Using Gensim.” 2021. https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/.