*Umadevi Areekkara Kizhakkemana*

*Student ID: 10592856*

7135CEM - Modelling and Optimization Under Uncertainty

Coursework

Part 1

Topic Modelling and Sentiment Analysis on Climate Tweets

Abstract:

This paper focuses on comparing multiple topic modelling techniques on climate related tweets. Topic modelling techniques like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Non-negative Matrix Factorization (NMF) are explored here. Since the dataset contains sentiment score from trusted reviewers, general sentiment analysis also carried out as part of this study to analyze what the world is thinking about the climate change. The main idea of this work is to apply different topic modeling algorithms on the same corpus and compare the behavioral difference between each of them.

Introduction:

As per United Nations, climate change can be considered as the shift in weather and temperature patterns. This shift can be natural or man-made [1]. From last few decades the effect of human activities on the earth impacts its climate balance and resulted in sudden climate changes. Climate change has become one of the most concerning global challenge for the society [2]. The consequences include water scarcity, draughts, forest fires, polar ice melts, raising sea level and may more. Scientists and ordinary people have different opinion on the climate change issues. Many think that human should avoid or reduce the activities like burning fuels, plastic usage etc. They pointed out the reduced pollution rate in the pandemic time as the main evidence for that since during the pandemic time, vehicle usage was less. Still, many others think that even though human reduce these kind of activities, climate change is an inevitable phenomena that will happen as per the long historic trends [3]. Many think that a volcanic eruption may contribute more adverse to the global warming than any human activities. Thus, people around the world have different viewpoints on climate changes. The main aim of this coursework is to understand ‘what people are thinking about climate change’ by analyzing recent climate related tweets. Upsurge in the usage of social media enabled a new opportunity to collect public opinion without directly do a survey. Tweeter is one of the most influencing and effective source in that kind.

Much research has been done so far on the basis of the tweeter data in public domain to analyze climate related issues. Few of them incorporated efficient topic modelling techniques also as part of their research. One of the predominate and recent one is the study of *Gaytan et al* in 2021. They analyzed the tweeter data collected using recurrent neural networks (RNN) and unsupervised learning Latent Dirichlet Allocation (LDA) topic modelling, to understand the actions to be taken to tackle climate change related issues [2]. Another recent research conducted by *Yasin et al* in 2020 was about thesentiment analysis of tweets on global warming. In their research paper, it mentioned the linear and probabilistic classifier methods for efficient sentiment analysis on text data [4]. ‘Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis’ by *Sidi et al* is another vital study on the topic model application on tweeter data. In this study, they explored text mining using LDA topic modelling and Sentiment analysis. It helped them to give insight views in business and scientific fields in Canada [5]. A much important study in this field is conducted by *Salvatore et al* in 2020. Their paper explores the nature of topics posted by public on Twitter during world Lupus Day to extract latent topics that occur in tweets and to identify what information is most discussed among them [6]. They succeeded in identifying the hidden information about the healthcare system by revelling the disagreement between the topics discussed by influencers on the same issue [6].

In this coursework, exploring the possibilities of LDA, LSA and NMF on recent climate tweets.

Dataset:

Recent climate tweets collected from a public data source (Kaggle). This data set contains nearly 50,000 tweets pertaining to climate change collected between 2018 and 2021. Each tweet is labelled as one of the following classes:

2 (News): The tweet links to factual news about climate change

1 (Pro): The tweet supports the belief of man-made climate change

0 (Neutral): The tweet neither supports nor refutes the belief of man-made climate change

-1 (Anti): The tweet does not believe in man-made climate change

The problem here is to identify the important topics that public broadcast about the climate change related issues. The results give a general awareness among the public regarding the discussions. Similarly, it will help the government authorities to gather public opinion and act accordingly.

Methodology:

Three different topic modelling techniques are applied in this course work. Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Non-negative Matrix Factorization (NMF). Sentiment analysis based on the sentiment score available also tried out.

Latent Semantic Analysis (LSA):

Topic modelling techniques are built on the assumption that each document is a mixture of topics, and each topic is a collection of words. The idea here is to identify the latent (hidden) topics in each document.

Latent Dirichlet Allocation (LDA):

References:

1. United Nations, Climate Action, an article about climate change in UN website, <https://www.un.org/en/climatechange/what-is-climate-change>
2. Gaytan Camarillo M, Ferguson E, Ljevar V, Spence A. Big Changes Start With Small Talk: Twitter and Climate Change in Times of Coronavirus Pandemic. *Front Psychol*. 2021;12:661395. Published 2021 Jun 15. doi:10.3389/fpsyg.2021.661395
3. Mann M.E. (2009) Climate Variability and Change, Last 1,000 Years. In: Gornitz V. (eds) Encyclopedia of Paleoclimatology and Ancient Environments. Encyclopedia of Earth Sciences Series. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4411-3\_42
4. Yasin Kirelli, Seher Arslankaya, "Sentiment Analysis of Shared Tweets on Global Warming on Twitter with Data Mining Methods: A Case Study on Turkish Language", Computational Intelligence and Neuroscience, vol. 2020, Article ID 1904172, 9 pages, 2020. <https://doi.org/10.1155/2020/1904172>
5. Sidi Yang, Haiyi Zhang, Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis, World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:12, No:7, 2018
6. Salvatore Pirri, Valentina Lorenzoni , Gianni Andreozzi, Marta Mosca and Giuseppe Turchetti, Topic Modeling and User Network Analysis on Twitter during World Lupus Awareness Day, International Journal of Enviornmental Research and Public Health, Jul 2020