

## Task 2

## Title: Topic Modelling Starting with Manually Collected Dataset of Unstructured Text

**Submitted**

**by**

**TOMISIN SAMUEL ISEDOWO**

**(@00749815)**

**In fulfillment of**

**Natural Language Processing**

**For**

**MASTER OF SCIENCE IN ARTIFICIAL INTELLIGENCE**

# FACULTY OF SCIENCE, ENGINEERING AND ENVIRONMENT

**April 2024.**

# INTRODUCTION

A component of natural language processing is Topic modeling, which can be used in mining a bulk sum of text or data where a collection of documents is taken by a topic model as an input, and without supervision attempts to uncover the underlying topics in this collection. (Jordan, 2003)

During this task, Latent Dirichlet Allocation (LDA) and BERTopic model were used to get the topics from the dataset. The data was on a manually collected unstructured dataset and was scrapped from Scopus using the keywords Weather.

The dataset will be examined and cleaned before used to train the model which will generate reasonable topics from the scraped dataset.

Weather is said to be the atmospheric condition at a particular place with time. It is measured in terms of variables including wind speed and direction, air temperature, humidity, atmospheric pressure, cloudiness, and precipitation. Weather can change at any time.

This task utilizes topic modeling techniques LDA and BERTopic for the analysis of extensive unstructured text data. The aim of this task is to detect key topics from the chosen dataset, thereby significantly reducing the time required for manual reading of dataset.

# LITERATURE REVIEW

Topic modeling is unsupervised learning which is a very powerful tool in Natural Language. To get a good result, a large amount of text dataset needs to be analyzed to get topics.

Weather refers to the atmospheric conditions of a specific location. It comprises of various elements such as temperature, humidity, precipitation, wind speed and direction, atmospheric pressure, cloud cover, and visibility (Uttar et al, 2016).

Weather is a key determinant to the development of an environment. This cut across agricultural produce, infrastructure, health etc. Forecasting the weather gives important information about the coming weather. Weather forecasting uses a variety of methods, from simple sky observation to extremely complex computerized mathematical programming. (Tektas, 2010).

For this task, two models are trained with the dataset which determine the relevant topics from the dataset.

# DATASET

For this task, two dataset was examined which was scraped from New York Times and Scopus. After examining the dataset, the Dataset scraped from Scopus was used. The dataset was generated with the keyword ‘Weather’ which means the atmospheric conditions present in a specific location at a particular time. and series of articles are generated for the search. The articles were scrapped to get the dataset which was used to train the model to get topics relating to the keyword.

# EXPLANATION AND PREPARATION OF THE DATASET

For this task, the dataset was scraped from Scopus. Before the dataset was scraped from the Scopus, The Article API was adopted to search for the Keyword, Weather. Each page of the searched article was exported to Excel named Weather.csv.

The dataset was uploaded to google colab and a little cleaning was initiated to get the dataset ready for the training. I decided to scrape the dataset from Scopus because of the structure of the dataset. I previously created a developer account and to get the API key for New York time but after scraping the data from NYT, I noticed the content of the dataset was repeated a lot.

# METHODOLOGY

**IMPLEMENTATION WAS CARRIED ON THE LOCAL ENVIRONMENT (LINUX) AND CLOUD (GOOGLE COLAB)**

Linux which is an open source operating system stands as a cornerstone of the computing world, offering a powerful and versatile platform for users and developers. Linux has become a mainstay in servers, desktops, mobile devices, and even embedded systems. The terminal helps to install the needed packages for the implementation of the task. Before installing the packages, a virtual environment was created for the purpose of this assessment which helps to isolate python packages and dependencies for the project, the virtual environment was created and activated . Now the packages are installed using this “**pip install pyLDAvis==3.4.1 gensim==4.3.1 spacy==3.5.2 bertopic==0.14.1 flair==0.12.2 numpy==1.23.5 scikit-learn==1.2.2 pandas==2.0.1 nltk==3.8.1”.** The kernel was installed and named “python3 -m ipykernel install --user -- name=Assessment2\_env”, this is done to run the code on the working environment “jupyter notebook”. A computer screen with text

Description automatically generated

# Google Colab denotes Google Collaboratory, which is a cloud-based Jupiter notebook environment. It is an adaptable platform for data analysis, machine learning, and Python programming. in a browser-based interface (Cloud), with no setup required.

# A screenshot of a computer Description automatically generated

# LIBRARIES USED

The below libraries are used:

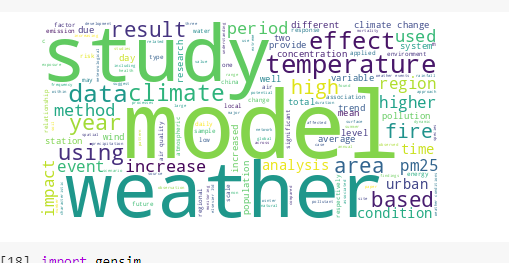
* **pandas:** pandas is one of the basic Python library which help in manipulation of data and analysis.
* **os:** The os helps in performing tasks like file and directory manipulation, environment variable access, etc.
* **nltk:** This is a platform for building a python program when working with large high level language data. The interface is very easy to use. This nltk provide a lot of text processing libraries.
* **BERTopic:** BERTopic is a Python library for topic modeling. It is based on clustering and then extracting topics from the clusters.
* **UMAP:** UMAP (Uniform Manifold Approximation and Projection) helps to reduce the dimensions that is used for representing high-dimensional data in lower dimensions. This helps to preserve the local architecture of the data.
* **HDBSCAN**: HDBSCAN denotes Hierarchical Density-Based Spatial Clustering of Applications with Noise. This algorithm has the ability to finds all the clusters of varying densities in the presence of noise. It's particularly useful for datasets where clusters have irregular shapes and varying densities.
* **KMeans**: KMeans is a popular clustering algorithm used for partitioning a dataset into a set of k clusters.
* **CountVectorizer:** CountVectorizer is a method for converting a collection of text documents into a matrix of token counts. It's commonly used as a preprocessing step for text data before applying machine learning algorithms.
* **SentenceTransformer:** SentenceTransformer is a Python library for computing dense vector representations of sentences and paragraphs using transformer-based models like BERT.

# EXPERIMENTAL PROCEDURE, SETTING, AND OPTIMIZATION OF MODEL HYPERPARAMETERS.

**DATA PREPARATION**

The dataset used for this task was scraped from Scopus. While preparing the dataset for the training, some of the columns in the dataset were dropped because it does not contain data needed for the training. The dataset column was filtered to 2 columns.

During preprocessing, stop words are removed from each paragraph, punctuation is removed then lemmatizes the remaining words. While preprocessing the dataset, a simple word cloud was used, the wordcloud package was imported to assist in getting the most common words in the dataset. This gives the result in a visualized format. This is because it is very important to know the content of the dataset and the types of words included.



**Fig 1: Pictorial view of the common words in the dataset**

# MODEL SELECTION

During the implementation of this task, Latent Dirichlet Allocation (LDA) is employed in topic modeling. LDA presumes that every document in a corpus contains different topics, and each word in the document is attributable to a document's topics. LDA attempts to get all underlying topics in a corpus and the words associated with each topic by analyzing the co-occurrence patterns of words across documents. LDA is to extract different topics from text dataset.

BERT is a Natural Language Processing Model proposed by Google Researchers (Pan et al 2022). This topic modeling technique uses transformers (BERT embeddings) and class-based TF-IDF to generate dense clusters. This method concentrated on topic mining and previously trained external knowledge to capture more accurate contextual representations. ( Liakata, 2020).

A diagram of a document

Description automatically generated

**Fig 2: Topic Modelling Architecture**

# HYPERPARAMETER

Here is the list of the hyperparameters and their function:

* **Number of Topics (K):** The Number of Topics determines the topics that the model identify in the corpus. It's crucial to choose an appropriate value for K based on the characteristics of the dataset and the desired granularity of topics.
* **Alpha (α):** Alpha is a hyperparameter that controls the sparsity of the document-topic distributions. A lower value of alpha results in documents containing fewer topics, while a higher value leads to documents covering more topics. It's typically set as a scalar value or a vector, with a scalar value resulting in symmetric priors for all topics and a vector allowing different priors for each topic.
* **Beta (β):** Beta is a hyperparameter that controls the sparsity of the topic-word distributions. Like alpha, a lower value of beta results in topics containing fewer words, while a higher value leads to topics covering more words. It's also typically set as a scalar value or a vector, with a scalar value resulting in symmetric priors for all words and a vector allowing different priors for each word.
* **Number of Iterations:** This hyperparameter determines the number of iterations or sweeps through the entire dataset that the model performs during training. Increasing the number of iterations may lead to better convergence but also increases computational time.

# VISUALIZATION:

# A screenshot of a graph Description automatically generated

**Fig 3:** **Relevant terms for Topic 1**

A screenshot of a computer

Description automatically generated

**Fig 4: Relevant terms for Topic 2**

A screenshot of a computer

Description automatically generated

**Fig 5: Relevant terms for Topic 3**

A screenshot of a computer

Description automatically generated

**Fig 6: Relevant terms for Topic 4**

A screenshot of a computer

Description automatically generated

**Fig 7: Relevant terms for Topic 5**

A screen shot of a graph

Description automatically generated

**Fig 8: Term score per topic**

A screenshot of a computer

Description automatically generated

**Fig 9: Intertopic Distance Map**

A screenshot of a graph

Description automatically generated

**Fig 10: Bertopic Hierarchical Clustering**

A screenshot of a computer screen

Description automatically generated

**Fig 11: Similarity Matrix**

A screenshot of a computer

Description automatically generated

**Fig 12: Topic Word Score**

# RESULTS AND DISCUSSION

During this task, LDA and Bertopic models were trained and evaluated using a dataset that was created using keyword Weather. The dataset was used to train Latent Dirichlet Allocation (LDA) and gets topics, then trained by Bertopic model. The result shows the top 30 most relevant terms for each topic and 5 topics are predicted for both LDA and Bertopic.

## Meaningful Topic and Interpretation of Topic for LDA

## Topic 1: Weather impacts on Health

Weather exerts a significant influence on human health, encompassing a range of impacts from heat-related illnesses to respiratory issues and vector-borne diseases. Understanding these connections allows for targeted public health interventions and community resilience strategies. With climate change contributing to more frequent and intense weather events, proactive measures, including public education, healthcare preparedness, and policy initiatives, are essential to protect and promote health in the face of changing weather patterns.

## Topic 2: Effect of extreme weather event

The scope of this topic is centered on extreme weather events present significant challenges with wide-ranging impacts on communities and ecosystems. Addressing these challenges requires a multi-faceted approach, including improved disaster preparedness, sustainable land-use practices.

## Topic 3: Effect of extreme weather event

Urban events, from festivals to outdoor concerts, markets, and sports games, bring communities together and contribute to the vibrancy of cities. However, the success and safety of these events can be greatly influenced by weather conditions. Understanding and adapting to weather patterns is essential for event planners and city authorities to ensure both enjoyment and preparedness.

## Topic 4: Effect of extreme weather event

## Wind modeling plays a crucial role in various industries, from renewable energy planning to urban design and environmental assessments. However, the accuracy and reliability of wind models are heavily influenced by prevailing weather conditions. Understanding these effects is essential for obtaining precise wind predictions and making informed decisions.

## Topic 5: The study of weather data

The analysis of the weather data provides valuable insights into local climate conditions over the study period. Trends in temperature, precipitation patterns, wind speed variations, and relative humidity were observed.

# CONCLUSION

Learning through the complexity of weather offers a multi-faceted journey of discovery. From the scientific principles governing atmospheric dynamics to the cultural and societal impacts of weather events, each facet provides valuable insights. Exploring weather fosters curiosity, critical thinking, and a deeper appreciation for the natural world. As we navigate the complexities of climate change and its effects, this knowledge becomes not just a tool for understanding, but a call to action for environmental stewardship and resilience-building.

The report shows that the dataset which was scraped with key word “Weather” was used to train the two models, LDA and BERTopic to prove the effectiveness.

Substantial achievements to date offer promising opportunities to utilize advanced technology in converting extensive raw data into practical insights. This method is particularly beneficial for interdisciplinary analysis focused on enhancing our understanding of intricate subjects, like the advancement of weather detection techniques.

# REFERENCES

Allen, R.G., L.S. Pereira, D. Raes, and M. Smith. 1998. Crop evapotranspiration – guidelines for computing crop water requirements. U.N. Food and Agriculture Organization, Rome. FAO Irrigation and Drainage Paper FAO56. www.fao.org/docrep/x0490E (accessed Nov. 2005)

Bhawna Mallick2 Student, Department of Computer Science and Engineering, Galgotias Educational Institutions, Greater Noida, Uttar Pradesh, India

International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 12, December 2016 Copyright to IJARCCE DOI 10.17148/IJARCCE.2016.51237 177 A Review on Weather Forecasting Techniques Garima Jain1,

M.Tektas, “Weather Forecasting Using ANFIS and ARIMA (2010): A Case study of Istanbul,” Enviornment Research , Engineering and Management , vol. 1(51), pp.5-10.

M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

Ma T, Pan Q, Rong H, Qian Y, Tian Y, Al-Nabhan N. T-BERTSum: topic-aware text summarization based on BERT. IEEE Trans Comput Soc Syst. 2022;9(3):879–890. doi: 10.1109/TCSS.2021.3088506. [CrossRef] [Google Scholar]

Peinelt N, Nguyen D, Liakata M. tBERT: topic models and BERT joining forces for semantic similarity detection. BERT; 2020. pp. 7047–7055. [Google Scholar]