University of Padova

Department of Information Engineering

Master Thesis in ICT for Internet and Multimedia

(Cybersystems)

An Experimental Assessment of the Efficacy of BERTopic.

Supervisor

Prof. Tomaso Erseghe

University of Padova

Master Candidate

Farin Binta Zahir

Student ID

2041369

Academic Year

2022-2023

"Patience is a virtue, and I'm learning patience. It's a tough lesson."
— Elon Musk

Abstract

Topic modelling is an unsupervised machine-learning technique for finding abstract topics in a large collection of documents. It helps in organizing, understanding, and summarizing large collections of textual information while discovering the latent topics that vary among documents in a given corpus. Recently, newly developed algorithms for topic modelling, such as BERTopic have gained significant attention from researchers and continue to attract growing interest. This research not only sheds light on the efficacy of using these advanced algorithms but also emphasizes the importance of possessing certain technical skills for conducting meaningful investigations in this domain. Efficient, speedy, and scalable implementations of these algorithms are essential for handling vast corpora of text data. Additionally, to ensure the success of this study and meaningful comparisons among various topic modelling approaches, proficiency in technical skills such as data analysis and data visualization is imperative. Utilizing Python as the programming language of choice provides the flexibility and robustness required for algorithmic implementations, while a solid foundation in statistical modelling and mathematical skills is indispensable for accurate calculation and prediction. Specifically, the main contribution of the study is to introduce the NMI (Normalized Mutual Information) and modularity which are the two different evaluation metrics used to assess the quality of clusters or topics generated by clustering algorithms, including those used in BERTopic. In essence, this research not only explores the effectiveness of state-of-the-art topic modelling algorithms but also underscores the significance of technical expertise in data analysis, data visualization, Python programming, and statistical modelling to facilitate comprehensive comparisons within the field of topic modelling.

Contents

Listing of figures

Listing of acronyms

POS ..... Part of Speech

NLTK ..... Natural Language Toolkit

BERT ..... Bidirectional Encoder Representations from Transformers

UMAP ..... Uniform Manifold Approximation and Projection

HDBSCAN ..... Hierarchical Density-Based Spatial Clustering of Applications with Noise

NMI ..... Normalized Mutual Information

TF-IDF ..... Term Frequency and Inverse Document Frequency

1 Introduction

Natural Language Processing (NLP) and unsupervised learning methods are closely related by playing a crucial role in processing and understanding natural language text data. Unsupervised learning is a classification of machine learning, in which the algorithm is given data without explicit instructions on what to do with it, and its task is to find specific patterns, structures, links, or relationships within the data or documents.

In Natural Language Processing (NLP), topic modeling is a technique used to discover abstract topics or themes within a collection of large text documents [1]. It is a form of unsupervised machine learning method that does not require labelled data or prior knowledge of the topics within the corpus and helps in organizing, summarizing, and understanding the content of outsized textual datasets. Topic modeling identifies patterns in words and phrases used across documents and groups them into coherent and meaningful topics.

Including BERTopic, Several popular topic modeling techniques, such as PLSA, LDA, NMF, and LSA, are used to extract topics and themes from numerous documents. In Specific, BERTopic (Bidirectional Encoder Representations from Transformers) emerged in 2020 as a novel approach to topic modeling, leveraging the power of BERT's contextual embeddings to improve the accuracy and interpretability of topics extracted from text data. BERTopic typically requires less text preprocessing than traditional topic modeling methods. This simplifies the data preparation process and makes the data more accessible. Moreover, BERTopic libraries and implementations are available in widespread programming languages, such as Python.

Generally, Normalized Mutual Information (NMI) and modularity are two individual evaluation metrics used to assess the quality of clusters or topics generated by BERTopic or similar clustering algorithms. A higher NMI indicates a stronger relationship between words and classes or topics, implying a better classification or topic modeling performance. On the other hand, modularity refers to the degree to which a system's components or parts can be separated and recombined. There are various types of modularity found across different disciplines and

systems. In general, it has a scale or interval ranging from -1 to 1 in which values can vary.

The universal process of topic modeling involves the following steps:
• Data acquisition
• Data prepossessing.
• Topic Modeling
• Evaluation metrics
• Result analysis

Figure 1.1: Pipeline of the universal process of topic modeling.

This research focuses on the working efficiency of using this advanced algorithm and emphasises the importance of possessing certain technical skills for conducting meaningful investigations in this domain. Efficient, speedy, and scalable implementations of this algorithm are essential for handling vast corpora of text data. The dominant objective of the research is to introduce the NMI (Normalized Mutual Information), traditional modularity and TF-IDF-based modularity to investigate the excellence of the clusters or topics generated by BERTopic. To determine the quality of the obtained results, they are turned into graphs, by which qualitative explanations can be established.

In synthesis, we illustrate the results of the research focused on:

□ Implementation of the BERTopic Model
□ Implementation of fine-tuning of the BERTopic Model
□ Implementation of NMI with the BERTopic Model
□ Evaluation of Modularity with the BERTopic Model
□ Evaluation of TF-IDF-based Modularity with the BERTopic Model
□ Implementation of the Static graphs from the obtained results.
□ Implementation of the 3D graphs from the obtained results.
□ Implementation of the 3D animated graphs from the obtained results.

2

Dataset

A dataset, or data set, is a collection of data related to a particular topic, theme, or industry. Datasets include different types of information, such as numbers, text, images, videos, and audio, and can be stored in various formats, such as xlsx, CSV, JSON, or SQL. So, a dataset typically involves structured data for a specific purpose and is related to the same subject [2].

Sequential and partitioned datasets are two different ways to organize data for various purposes, such as analysis, modelling, or processing. The key difference between sequential and partitioned datasets is in their organization and use cases. Sequential datasets maintain a specific order of data points over time and are often used for time series analysis. Partitioned datasets, on the other hand, involve dividing data into subsets or partitions based on specific attributes or criteria, which allows for efficient processing and analysis of data grouped by relevant categories or segments.

In the initial phase of this experiment, a total of five distinct datasets (YELP, Bishop Topic Modelling Dataset, Topic Modeling for Research Articles 2.0, Twitter-metoo, and Twitter-Covid) were employed for the fine-tuning process. Notably, it is essential to acknowledge that while certain datasets within this repertoire were annotated, others remained unannotated, reflecting the diversity of data sources and content under investigation. Subsequently, a selection was made to utilize two specific datasets (Twittwer- metoo and Twitter- covid) exclusively for the remaining phases of the experiment.

Twitter is a popular social media platform and microblogging service that allows users to share short messages, called ”tweets,” with their followers. Tweets can contain up to 280 characters, although there have been occasional changes to the character limit over time. Users can post text-based tweets, as well as multimedia content like photos, videos, and links. Hashtags on Twitter are used to categorize tweets, make them discoverable, and engage in conversations on specific topics or events. So hashtags on Twitter play a significant role in organizing

conversations, tracking trends, and enhancing the discoverability of tweets. They are widely used for various purposes, including social interaction, news dissemination, marketing, and event promotion.

A Twitter dataset refers to a sophisticated collection of data from the Twitter social media platform. These datasets can include a wide range of information gathered from tweets, user profiles, and other Twitter-related data. Twitter datasets are commonly used for research, data analysis, machine learning, sentiment analysis, social network analysis, and various other applications. Twitter datasets can vary in size and scope, from small collections of tweets on a specific topic to extensive archives of Twitter data covering a significant time period. Researchers and data analysts often use Twitter APIs or third-party tools to collect and create such datasets for their specific research or analysis needs.

2.1 Twitter-metoo

This dataset was collected weekly from the Twitter API through Social Feed Manager using the POST statuses/filter method of the Twitter Stream API. Tweets range from June, 2017, to June, 2018. The following list of 76 terms includes the hashtags used to collect data for this dataset : #metoo, #timesup, #metoostem, #sciencetoo, #metoophd, #shittymediamen, #churchtoo, #ustoo, #metooMVMT, #ARmetoo, #TimesUpAR, #metooSociology, #metooSexScience, #timesupAcademia, #metooMedicine, #MyCampusToo, #howiwillchange, #iwill, #believewomen, #GoTeal, #BelieveChristine, #IStandWithDrFord, #IStandWithChristineBlaseyFord, #believesurvivors, #whyididntreport, #himtoo, #istandwithbrett, #confirmkavanaguhnow, #metooMcdonalds, #metoomovement, #muteRKelly, #WeBelieveDrFord, #WeBelieveSurvivors, #HandsOffPantsOn, #MeAt14, #HeToo, #MeTooLiars, #metoolynchings, #metoohucksters, #metoohustle, #ItWasMe, #Ihave, #TimesUpTech, #GoogleWalkout, #mosquemetoo, #faithandmetoo, #SilenceIsNotSpiritual, #HealMeToo, #TimesUpHarvard, #NoCarveOut, #TimesUpx2, #MeetingsToo, #metoonatsec, #healmetoo, #GamAni, #ShulToo, #harvardhearsyou, #metooarcheology, #TimesUpPayUp, #metooarcheology, #metooHBCU, #TimesUpHC,#aidtoo, #garmentmetoo, #mutemetoo, #mutetimesup, #metoopolisci, #copstoo, #TimesUpBiden, #MeTooNoMatterWho, #IBelieveTara, #BelieveAllWomen, #metoomilitary, #harvard38, #comaroff, and #harvardletter.

This dataset contains 3 Excel files and the size is only 3.15 MB. The dataset contains 3024 rows and 99 basic columns. Finding a relevant dataset with a great variety should be an ideal dataset for the data science community to jump and start their journey in NLP or Topic modelling.

df

	Source (A)	id	text	created_at	community	WC	Analytic	Clout	Authentic	Tone	...	Colon	SemiC	QMark	Exclam	Dash	Quote	Apostro	Parenth	OtherP	agency
0	0	1013189209488790016	#PrideMonth may be ending today, but we will c...	2018 Sat Jun 30 22:35:06 +0000	12	35	60.73	95.65	43.37	78.19	...	2.86	0.00	0.0	0.00	0.00	0.00	2.86	0.0	14.29	0.564734
1	1	1013143915308950016	‚ÁúThis disaster leaves us in extreme poverty....	2018 Sat Jun 30 19:35:07 +0000	11	36	98.14	95.20	11.80	2.09	...	2.78	2.78	0.0	0.00	5.56	2.78	2.78	0.0	13.89	0.052641
2	2	1013128815575469952	We want more women on ballots □üó±Ô∏è when we ...	2018 Sat Jun 30 18:35:07 +0000	13	37	86.66	94.74	21.26	75.85	...	0.00	0.00	0.0	2.70	5.41	0.00	0.00	0.0	8.11	0.558930
3	3	1013098625768000000	‚ÁúThe idea that sexism & misogyny in onli...	2018 Sat Jun 30 16:35:09 +0000	7	31	93.26	50.00	20.84	25.77	...	0.00	6.45	0.0	0.00	0.00	0.00	0.00	0.0	16.13	0.315561
4	4	1013075989685340032	When women are targeted online, the abuse is m...	2018 Sat Jun 30 15:05:12 +0000	0	41	95.24	76.78	14.43	1.00	...	0.00	4.88	0.0	2.44	0.00	0.00	0.00	0.0	12.20	0.702103
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3018	3018	870081565363232000	‚ÁúWe will hold our leaders accountable and ma...	2017 Thu Jun 01 00:56:25 +0000	15	26	54.88	97.24	1.40	99.00	...	0.00	0.00	0.0	0.00	3.85	0.00	0.00	0.0	15.38	1.102073
3019	3019	870078835689127040	‚Ãúlt‚Äôs a real treat to	2017 Thu Jun 01	12	27	97.84	35.54	15.12	99.00	...	0.00	0.00	0.0	0.00	3.70	0.00	0.00	0.0	18.52	0.188077

Figure 2.1: Twiiter-metoo dataset.

2.2 Twitter-Covid

The current dataset contains a total of 3 Excel files and 12.6MB, Tweet IDs for Twitter posts that mentioned "COVID" as a keyword or as part of a hashtag (e.g., COVID-19, COVID19) between February and April of 2020. The predominant dataset contains 9000 rows and 108 basic columns. This dataset was collected weekly from the Twitter API through Social Feed Manager using the POST statuses or filter method of the Twitter Stream API. This dataset, like most datasets collected via the Twitter Search API, is a sample of the available tweets on this topic and is not meant to be comprehensive. Some COVID-related tweets might not be included in the dataset either because the tweets were collected using a standardized but intermittent (hourly) sampling protocol or because tweets used hashtags or keywords other than COVID (e.g., Coronavirus or #nCoV). It can be used in a lot of Domains such as Topic Modelling, Content Identification, etc.

The "Twitter-MeToo" dataset is used to create partitioned datasets, specifically within the temporal scope of the year 2018. The partitioning process involves dividing the data into subsets or partitions based on relevant attributes, such as time or specific criteria, to enable more targeted and efficient data analysis and processing within the specified time frame.

In contrast, the "COVID-Twitter" dataset is utilized in its entirety, without the need for partitioning. The entire dataset, which comprises data spanning across various periods, is employed for comprehensive data analysis and research, without any temporal restrictions, in the context of the COVID-19 pandemic.

	Unnamed: 0	id	author_id	created_at	referenced_tweets	text	retweets	likes	replies	quotes	...	Colon	SemiC	QMark	Exclam	Dash	Quote	Apostro	Parenth	OtherP	agency
0	0	1227248333871173632	14499829	2020-02-11T15:09:38.000Z	[{'type': 'replied_to', 'id': '1227248085368865...	@DrTedros ?? BREAKING ??\n\n"We now have a nam...	8154	7798	1178	2027	...	6.06	0.00	0.0	0.0	21.21	6.06	3.03	0.0	18.18	0.087864
1	1	1227248587263407744	14499829	2020-02-11T15:10:39.000Z	[{'type': 'replied_to', 'id': '1227248333387117...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	416	920	73	107	...	0.00	1.82	0.0	0.0	1.82	3.64	0.00	0.0	30.91	0.299545
2	2	1227248866127159298	14499829	2020-02-11T15:10:57.000Z	[{'type': 'replied_to', 'id': '1227248587253406...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	282	736	58	52	...	0.00	0.00	0.0	0.0	2.13	4.26	0.00	0.0	34.04	0.276209
3	3	1227248956552878083	14499829	2020-02-11T15:11:43.000Z	[{'type': 'replied_to', 'id': '1227248866127159...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	485	621	37	56	...	0.00	1.64	0.0	0.0	3.28	3.28	0.00	0.0	29.51	0.016819
4	4	1227249112342396928	14499829	2020-02-11T15:12:44.000Z	[{'type': 'replied_to', 'id': '1227248956552878...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	55	110	4	1	...	0.00	0.00	0.0	0.0	5.13	5.13	0.00	0.0	48.72	1.223480
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8995	8995	1249099307166261255	2809959174	2020-04-11T22:17:36.000Z	NaN	It's adversity that gives us the opportunity t...	209	1630	150	32	...	0.00	0.00	0.0	0.0	0.00	0.00	6.00	0.0	6.00	0.956527
8996	8996	1249104459889692674	86626845	2020-04-11T22:38:05.000Z	NaN	The list of successful countries vs #COVID19 i...	560	1173	87	82	...	0.00	0.00	0.0	0.0	0.00	0.00	0.00	0.0	7.50	0.461828
8997	8997	1249106220750909440	1222205118466154496	2020-04-11T22:45:04.000Z	NaN	Supporting our partners in the fight against #...	1671	9717	216	179	...	0.00	0.00	0.0	0.0	0.00	0.00	0.00	0.0	13.04	0.658837
8998	8998	1249116241123594245	34747812	2020-04-11T23:24:53.000Z	NaN	To help battle #COVID19 across the country, we...	388	3036	209	105	...	2.08	2.08	0.0	0.0	0.00	0.00	4.17	0.0	12.50	0.796271
8999	8999	1249118174567596032	251686253	2020-04-11T23:32:34.000Z	[{'type': 'quoted', 'id': '1248924729710428160'}]	RIP #EddieJohnson & #BlancheJohnson	382	1907	23	11	...	0.00	6.25	0.0	0.0	0.00	0.00	0.00	0.0	37.50	0.375974

Figure 2.2: Twiiter-Covid dataset.

3

Preprocessing

A Proper dataset or multiple dataset collection, cleaning, and evaluation are hypercritical and crucial stages in the field of data analysis. Data is similar to oil. The oil obtained from natural sources is not as worthy or as usable in its original state as it is after refining. So Data also needs to be refined or preprocessed the data before passing it to the model for better understanding and performance. Otherwise, the entire process will be garbage in garbage out with dirty data.

Text preprocessing is a method to clean the text data and make it ready to feed data to the model. There are so many ways in the human language to express the same emotion with different attitudes or ways. This is why Text data always contains noise in various forms like emotions, punctuation, special characters and text in different cases. This is only the main problem we have to deal with because machines will not understand words, they need numbers so we need to convert text to numbers in an efficient manner [3]. So it can be stated that Data cleaning is a process of removal of incorrect, incomplete, and inaccurate data which also replaces the missing data. The quality of the input data directly impacts the effectiveness of the subsequent analysis, highlighting the importance of careful and thorough preprocessing [4].

The preprocessing for this study was structured into three distinct stages: superficial cleaning, deep cleaning, and empty row elimination. This segmentation was implemented due to the distinct data input requirements of various algorithms and the differing methods employed for data analysis. The empty row elimination process followed deep cleaning to ensure the integrity and quality of the dataset. Each of these stages was carefully chosen to align with the specific demands of the corresponding algorithm and the overall research objectives.

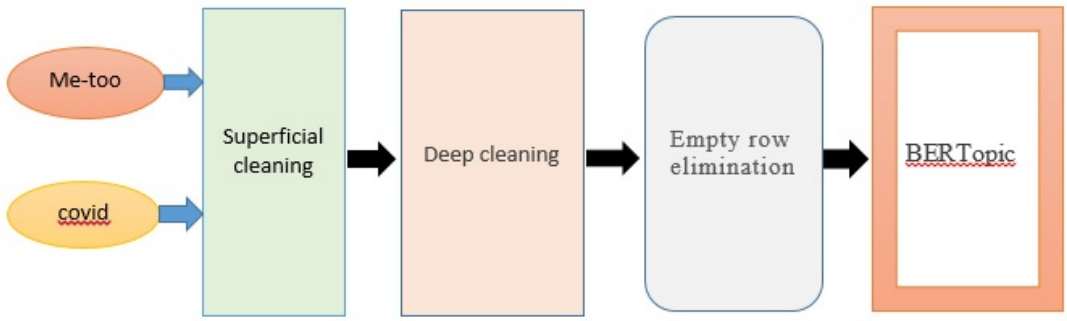


Figure 3.1: Structure of Preprocessing Steps of the system.

3.1 Superficial cleaning

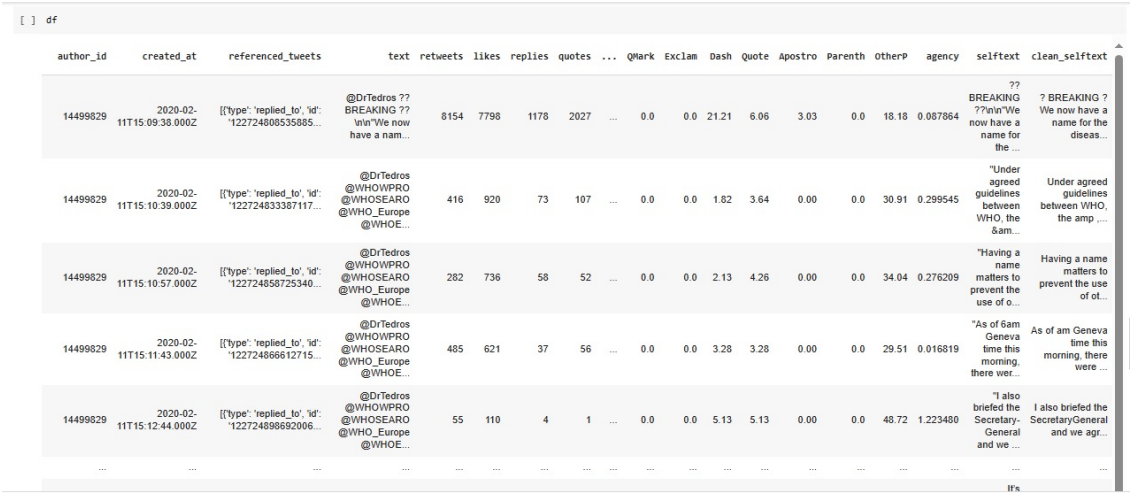
Twitter, in particular, restricts each post to a maximum of 280 characters [5]. Although these short and unstructured posts conform to social media practice, they increase the complexity for algorithms to make sense of digital interaction. Common challenges arise from using compound words, acronyms, and ungrammatical sentences [6]. Despite the productive and unexpressed nature of compound words they often complicate computational analysis [7]. Other difficulties emerge when data are meaningless (i.e., noisy data) or when there are many gaps present in the data (i.e., sparse data; [8]).

In contrast to scientific paper text data, the dataset under consideration exhibits a lack of cleanliness and is characterized by various forms of noise. This noise encompasses instances of double punctuation, duplicated letters, misspelt words, emoticons, special characters, randomly arranged letters, multiple usages of the same letters in a row to emphasise content, and content originating from automated bots and human moderators. Such noise levels introduce an element of ambiguity for the algorithms employed, as they may occasionally misconstrue non-textual elements as integral parts of the textual data. This is why superficial cleaning aims to prepare the data for further processing, making it more manageable and ensuring it meets the basic quality standards necessary for accurate analysis.

To determine which elements should be addressed in the superficial cleaning stage, a manual review of the dataset was undertaken. The key criterion for selecting these cleaning actions was to eliminate extraneous non-textual elements while preserving the core text. Consequently, this cleaning process did not include actions such as lemmatization or the removal of stopwords.

The following steps represent the superficial cleaning process.

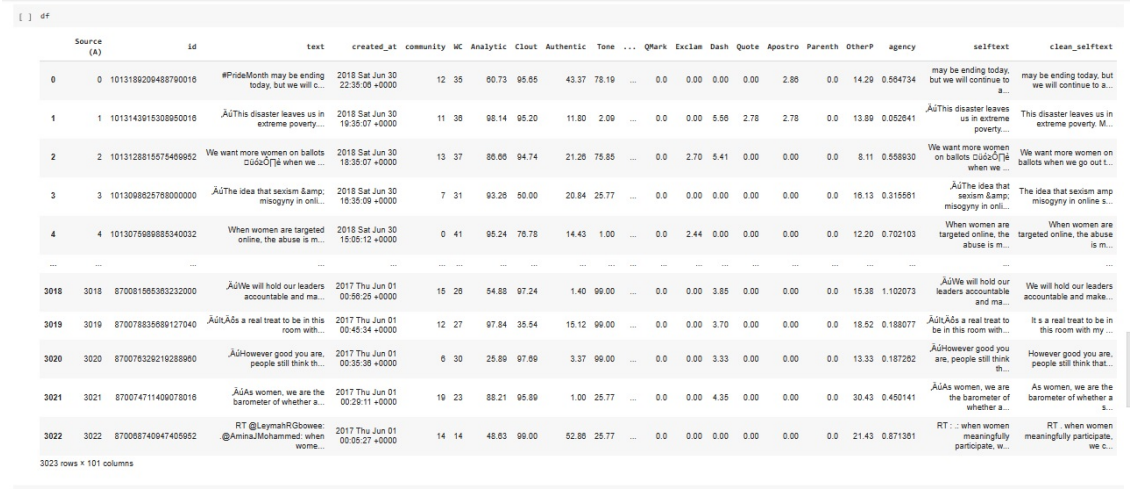
1. Fixing the contractions.
2. Removing website links.
3. Removing accented characters.
4. Removing the text inside square brackets which mentions the following link or a warning such as '[read this]'.
5. Removing emoji.
6. Removing moderator messages such as the warnings for the owner of collected posts or tweets.
7. Removing hashtags and mentions with punctuation and text such as #keyword or user-name.
8. Removing double spaces.
9. Removing non-text special words which are based on the text type such as '&#x200B' or '&amp'.
10. Removing extra used new lines.
11. Limiting all the repetitions to two characters and removing the extra characters.
12. Removing punctuation except for main sentence punctuation.
13. Removing the sentences which represent the rule of the community such as starting with'[**(Full Rules)**]'.
14. Removing numbers.



[ ] df

	author_id	created_at	referenced_tweets	text	retweets	likes	replies	quotes	...	QMark	Exclam	Dash	Quote	Apostro	Parenth	OtherP	agency	selftext	clean_selftext
	14499829	2020-02-11T15:09:38.000Z	[{'type': 'replied_to', 'id': '1227248085358855...	@DrTedros ?? BREAKING ?? \n\n"We now have a nam...	8154	7798	1178	2027	...	0.0	0.0	21.21	6.06	3.03	0.0	18.18	0.087864	?? BREAKING ??\n\n"We now have a name for the ...	? BREAKING ? We now have a name for the diseas...
	14499829	2020-02-11T15:10:39.000Z	[{'type': 'replied_to', 'id': '1227248333871177...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	416	920	73	107	...	0.0	0.0	1.82	3.64	0.00	0.0	30.91	0.299545	"Under agreed guidelines between WHO, the &am...	Under agreed guidelines between WHO, the amp ,...
	14499829	2020-02-11T15:10:57.000Z	[{'type': 'replied_to', 'id': '1227248587253401...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	282	736	58	52	...	0.0	0.0	2.13	4.26	0.00	0.0	34.04	0.276209	"Having a name matters to prevent the use of o...	Having a name matters to prevent the use of ot...
	14499829	2020-02-11T15:11:43.000Z	[{'type': 'replied_to', 'id': '1227248666612715...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	485	621	37	56	...	0.0	0.0	3.28	3.28	0.00	0.0	29.51	0.016819	"As of 6am Geneva time this morning, there wer...	As of am Geneva time this morning, there were ...
	14499829	2020-02-11T15:12:44.000Z	[{'type': 'replied_to', 'id': '1227248986920061...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	55	110	4	1	...	0.0	0.0	5.13	5.13	0.00	0.0	48.72	1.223480	"I also briefed the Secretary-General and we ...	I also briefed the SecretaryGeneral and we agr...
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3.2: Preprocessing of Twitter-Covid dataset after cleaning contractions, emoji, accented characters, mentions, hashtags, and punctuations.



[ ] df

	Source (A)	id	text	created_at	community	WC	Analytic	Clout	Authentic	Tone	...	QMark	Exclam	Dash	Quote	Apostro	Parenth	OtherP	agency	selftext	clean_selftext
0	0	1013189209488790016	#PrideMonth may be ending today, but we will c...	2018 Sat Jun 30 22:35:06 +0000	12	35	60.73	95.65	43.37	78.19	...	0.0	0.00	0.00	0.00	2.86	0.0	14.29	0.564734	may be ending today, but we will continue to a...	may be ending today, but we will continue to a...
1	1	1013143915308950016	ÃúThis disaster leaves us in extreme poverty....	2018 Sat Jun 30 19:35:07 +0000	11	36	98.14	95.20	11.80	2.09	...	0.0	0.00	5.56	2.78	2.78	0.0	13.89	0.052641	ÃúThis disaster leaves us in extreme poverty....	This disaster leaves us in extreme poverty. M...
2	2	1013128815575469952	We want more women on ballots □úózÕ∏è when we ...	2018 Sat Jun 30 18:35:07 +0000	13	37	86.66	94.74	21.26	75.85	...	0.0	2.70	5.41	0.00	0.00	0.0	8.11	0.558930	We want more women on ballots □úózÕ∏è when we ...	We want more women on ballots when we go out t...
3	3	1013098625768000000	ÃúThe idea that sexism & misogyny in onli...	2018 Sat Jun 30 16:35:09 +0000	7	31	93.26	50.00	20.84	25.77	...	0.0	0.00	0.00	0.00	0.00	0.0	16.13	0.315561	ÃúThe idea that sexism & misogyny in onli...	The idea that sexism amp misogyny in online s...
4	4	1013075989885340032	When women are targeted online, the abuse is m...	2018 Sat Jun 30 15:05:12 +0000	0	41	95.24	76.78	14.43	1.00	...	0.0	2.44	0.00	0.00	0.00	0.0	12.20	0.702103	When women are targeted online, the abuse is m...	When women are targeted online, the abuse is m...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3018	3018	870081565383232000	ÃúWe will hold our leaders accountable and ma...	2017 Thu Jun 01 00:56:25 +0000	15	26	54.88	97.24	1.40	99.00	...	0.0	0.00	3.85	0.00	0.00	0.0	15.38	1.102073	ÃúWe will hold our leaders accountable and ma...	We will hold our leaders accountable and make...
3019	3019	870078835689127040	ÃúItÂ ôs a real treat to be in this room with...	2017 Thu Jun 01 00:45:34 +0000	12	27	97.84	35.54	15.12	99.00	...	0.0	0.00	3.70	0.00	0.00	0.0	18.52	0.188077	ÃúItÂ ôs a real treat to be in this room with...	It s a real treat to be in this room with my ...
3020	3020	870076329219288960	ÃúHowever good you are, people still think th...	2017 Thu Jun 01 00:35:36 +0000	6	30	25.89	97.69	3.37	99.00	...	0.0	0.00	3.33	0.00	0.00	0.0	13.33	0.187262	ÃúHowever good you are, people still think th...	However good you are, people still think that...
3021	3021	870074711409078016	ÃúAs women, we are the barometer of whether a...	2017 Thu Jun 01 00:29:11 +0000	19	23	88.21	95.89	1.00	25.77	...	0.0	0.00	4.35	0.00	0.00	0.0	30.43	0.450141	ÃúAs women, we are the barometer of whether a...	As women, we are the barometer of whether a s...
3022	3022	870068740947406952	RT @LeymahRGbowee: .@AminaJMohammed: when wome...	2017 Thu Jun 01 00:05:27 +0000	14	14	48.63	99.00	52.88	25.77	...	0.0	0.00	0.00	0.00	0.00	0.0	21.43	0.871361	RT : : when women meaningfully participate, w...	RT . when women meaningfully participate, we c...

3023 rows × 101 columns

Figure 3.3: Preprocessing of Twitter-metoo dataset after cleaning contractions, emoji, accented characters, mentions, hashtags, and punctuations.

```
[ ] clean_df
```

	clean_selftext	clean_text
0	? BREAKING ? We now have a name for the diseas...	breaking we now have a name for the disease...
1	Under agreed guidelines between WHO, the amp ,...	under agreed guidelines between who the amp w...
2	Having a name matters to prevent the use of ot...	having a name matters to prevent the use of ot...
3	As of am Geneva time this morning, there were ...	as of am geneva time this morning there were ...
4	I also briefed the SecretaryGeneral and we agr...	i also briefed the secretarygeneral and we agr...
...	...	...
8995	It is adversity that gives us the opportunity ...	it is adversity that gives us the opportunity ...
8996	The list of successful countries vs is extendi...	the list of successful countries vs is extendi...
8997	Supporting our partners in the fight against	supporting our partners in the fight against ...
8998	To help battle across the country, we are send...	to help battle across the country we are sendi...
8999	RIP amp Atlanta. Your country owed you better.	rip amp atlanta your country owed you better

9000 rows × 2 columns

```
[ ] df.columns

Index(['Unnamed: 0.1', 'Unnamed: 0', 'id', 'author_id', 'created_at',
       'referenced_tweets', 'text', 'retweets', 'likes', 'replies',
       ...
       'QMark', 'Exclam', 'Dash', 'Quote', 'Apostro', 'Parenth', 'OtherP',
       'agency', 'selftext', 'clean_selftext'],
      dtype='object', length=112)
```

Figure 3.4: Twitter-Covid dataset after all processes of the Preprocessing.

```
[ ] clean_df
```

	clean_selftext	clean_text
0	may be ending today, but we will continue to a...	may be ending today but we will continue to ad...
1	This disaster leaves us in extreme poverty. M...	this disaster leaves us in extreme poverty ma...
2	We want more women on ballots when we go out t...	we want more women on ballots when we go out t...
3	The idea that sexism amp misogyny in online s...	the idea that sexism amp misogyny in online s...
4	When women are targeted online, the abuse is m...	when women are targeted online the abuse is mo...
...	...	...
3018	We will hold our leaders accountable and make...	we will hold our leaders accountable and make...
3019	It s a real treat to be in this room with my ...	it s a real treat to be in this room with my ...
3020	However good you are, people still think that...	however good you are people still think that ...
3021	As women, we are the barometer of whether a s...	as women we are the barometer of whether a so...
3022	RT . when women meaningfully participate, we c...	rt when women meaningfully participate we can...

3023 rows × 2 columns

```
[ ] df.columns

Index(['Unnamed: 0', 'Source (A)', 'id', 'text', 'created_at', 'community',
       'WC', 'Analytic', 'Clout', 'Authentic',
       ...
       'QMark', 'Exclam', 'Dash', 'Quote', 'Apostro', 'Parenth', 'OtherP',
       'agency', 'selftext', 'clean_selftext'],
      dtype='object', length=102)
```

Figure 3.5: Twitter-metoo dataset after all processes of the Preprocessing.

3.2 Deep cleaning

Deep cleaning is a widespread procedure that goes beyond surface-level text cleaning to eliminate various forms of noise in text data and convert it into a standardized, cleaned format. This process involves multiple steps, including converting all text to lowercase, correcting spelling errors, removing punctuation, breaking the text into individual tokens, determining the part of speech for each word, and lemmatizing the words. The primary goal of deep cleaning is to extract the most critical information from the text and create a consistent format that can be fed into the BERTopic algorithm. This preparation allows the BERTopic algorithm to efficiently identify and categorize related topics within the data.

The following steps represent the deep cleaning process one by one in this research:
1. Lowercasing.
2. Correcting spellings,
3. Word tokenization,
4. Pos tagging,
5. Stop words removing,
6. Lemmatization

spaCy is a high-performance natural language processing library for Python. It is designed for production use and offers fast, efficient, and accurate linguistic annotation and text processing. spaCy includes pre-trained models for various languages, which can be used for tasks such as tokenization, part-of-speech tagging, named entity recognition, and more. It's widely appreciated for its speed and ease of use, making it a preferred choice for developers and data scientists working on NLP applications.

In this research, Spacy was utilized for several steps of deep cleaning, including word tokenization, POS tagging, stopword removal, and lemmatization. Word tokenization is the process of splitting a sentence or text into individual words or tokens [9]. POS tagging is Past of Speech tagging and involves assigning a grammatical category to each word in a sentence [9]. It was decided to use the Universal Dependencies part of speech tag set with Spacy. The Universal Dependencies tag set is a standardized set of POS tags that are applicable to many languages. It provides a universal way to label and annotate grammatical categories across different languages, which enables easy interoperability and comparability across different NLP tasks and applications.

List of tags [10]:
- ADJ: Adjective
- ADV: Adverb
- AUX: Auxiliary verb
- CCONJ: Coordinating conjunction
- DET: Determiner
- INTJ: Interjection
- NOUN: Noun

• NUM: Numeral
• PART: Particle
• PRON: Pronoun
• PROPN: Proper noun
• PUNCT: Punctuation
• SCONJ: Subordinating conjunction
• SYM: Symbol
• VERB: Verb
• X: Other

df

	id	author_id	created_at	referenced_tweets	text	retweets	...	Apostro	Parenth	OtherP	agency	selftext	clean_selftext	clean_selftext_v1	token	token_pos	pos_tag
0	1227248333871173632	14499829	2020-02-11T15:09:38.000Z	[{'type': 'replied_to', 'id': '122724808535885...	@DrTedros ?? BREAKING ?? \n\n"We now have a nam...	8154	...	3.03	0.0	18.18	0.087864	?? BREAKING ??\n\n"We now have a name for the ...	? BREAKING ? We now have a name for the diseas...	breaking we now have a name for the disease...	['breaking', 'we', 'now', 'have', 'a', 'name',...	['VERB', 'PRON', 'ADV', 'VERB', 'DET', 'NOUN',...	[('breaking', 'VERB'), ('we', 'PRON'), ('now',...
1	1227248587253407744	14499829	2020-02-11T15:10:39.000Z	[{'type': 'replied_to', 'id': '122724833387117...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	416	...	0.00	0.0	30.91	0.299545	"Under agreed guidelines between WHO, the &am...	Under agreed guidelines between WHO, the amp ,...	under agreed guidelines between who the amp w...	['under', 'agreed', 'guidelines', 'between', '...	['ADP', 'VERB', 'NOUN', 'ADP', 'PRON', 'DET', ...	[('under', 'ADP'), ('agreed', 'VERB'), ('guide...
2	1227248666127159298	14499829	2020-02-11T15:10:57.000Z	[{'type': 'replied_to', 'id': '122724858725340...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	282	...	0.00	0.0	34.04	0.276209	"Having a name matters to prevent the use of o...	Having a name matters to prevent the use of ot...	having a name matters to prevent the use of ot...	['having', 'a', 'name', 'matters', 'to', 'prev...	['VERB', 'DET', 'NOUN', 'VERB', 'PART', 'VERB'...	[('having', 'VERB'), ('a', 'DET'), ('name', 'N...
3	1227248856552878083	14499829	2020-02-11T15:11:43.000Z	[{'type': 'replied_to', 'id': '122724866612715...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	485	...	0.00	0.0	29.51	0.016819	"As of 6am Geneva time this morning, there wer...	As of am Geneva time this morning, there were ...	as of am geneva time this morning there were ...	['as', 'of', 'am', 'geneva', 'time', 'this', '...	['ADP', 'ADP', 'NOUN', 'NOUN', 'NOUN', 'DET', ...	[('as', 'ADP'), ('of', 'ADP'), ('am', 'NOUN'),...
4	1227249112342396928	14499829	2020-02-11T15:12:44.000Z	[{'type': 'replied_to', 'id': '122724898692006...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	55	...	0.00	0.0	48.72	1.223480	"I also briefed the Secretary-General and we ...	I also briefed the SecretaryGeneral and we agr...	i also briefed the secretarygeneral and we agr...	['i', 'also', 'briefed', 'the', 'secretarygene...	['PRON', 'ADV', 'VERB', 'DET', 'ADJ', 'CCONJ',...	[('i', 'PRON'), ('also', 'ADV'), ('briefed', '...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3.6: Twitter-Covid dataset after POS tagging and tokenization of the Preprocessing.

df

	Unnamed: 0.2	Unnamed: 0.1	Unnamed: 0	Source (A)	id	text	created_at	community	WC	Analytic	...
0	0	0	0	0	1013189209488790016	#PrideMonth may be ending today, but we will c...	2018 Sat Jun 30 22:35:06 +0000	12	35	60.73	...
1	1	1	1	1	1013143915308960016	ÄúThis disaster leaves us in extreme poverty....	2018 Sat Jun 30 19:35:07 +0000	11	36	98.14	...
2	2	2	2	2	1013128815575469952	We want more women on ballots □ûô≥Ô∏è when we ...	2018 Sat Jun 30 18:35:07 +0000	13	37	86.66	...
3	3	3	3	3	1013098625768000000	ÄúThe idea that sexism & misogyny in onl...	2018 Sat Jun 30 16:35:09 +0000	7	31	93.26	...
4	4	4	4	4	1013075989885340032	When women are targeted online, the abuse is m...	2018 Sat Jun 30 15:05:12 +0000	0	41	95.24	...
...	...	...	...	...	...	...	...	...	...	...	...
3018	3018	3018	3018	3018	870081565363232000	ÄúWe will hold our leaders accountable and ma...	2017 Thu Jun 01 00:56:25 +0000	15	26	54.88	...
3019	3019	3019	3019	3019	870078835689127040	ÄúIt‚Äôs a real treat to be in this room with...	2017 Thu Jun 01 00:45:34 +0000	12	27	97.84	...
3020	3020	3020	3020	3020	870076329219288960	ÄúHowever good you are, people still think th...	2017 Thu Jun 01 00:35:36 +0000	6	30	25.89	...
3021	3021	3021	3021	3021	870074711409078016	ÄúAs women, we are the barometer of whether a...	2017 Thu Jun 01 00:29:11 +0000	19	23	88.21	...
3022	3022	3022	3022	3022	870068740947405952	RT @LeymahRGbowee: .@AminaJMohammed: when wome...	2017 Thu Jun 01 00:05:27 +0000	14	14	48.63	...

	Apostro	Parenth	OtherP	agency	selftext	clean_selftext	clean_selftext_v1	token	token_pos	pos_tag
0	2.86	0.0	14.29	0.564734	may be ending today, but we will continue to a...	may be ending today, but we will continue to a...	may be ending today but we will continue to ad...	['may', 'be', 'ending', 'today', 'but', 'we', ...	['AUX', 'AUX', 'VERB', 'NOUN', 'CCONJ', 'PRON'...	[('may', 'AUX'), ('be', 'AUX'), ('ending', 'VE...
1	2.78	0.0	13.89	0.052641	ÄúThis disaster leaves us in extreme poverty....	This disaster leaves us in extreme poverty. M...	this disaster leaves us in extreme poverty ma...	['this', 'disaster', 'leaves', 'us', 'in', 'ex...	['DET', 'NOUN', 'VERB', 'PRON', 'ADP', 'ADJ', ...	[('this', 'DET'), ('disaster', 'NOUN'), ('leav...
2	0.00	0.0	8.11	0.558930	We want more women on ballots □ûô≥Ô∏è when we ...	We want more women on ballots when we go out t...	we want more women on ballots when we go out t...	['we', 'want', 'more', 'women', 'on', 'ballots...	['PRON', 'VERB', 'ADJ', 'NOUN', 'ADP', 'NOUN',...	[('we', 'PRON'), ('want', 'VERB'), ('more', 'A...
3	0.00	0.0	16.13	0.315561	ÄúThe idea that sexism & misogyny in onl...	The idea that sexism amp misogyny in online s...	the idea that sexism amp misogyny in online s...	['the', 'idea', 'that', 'sexism', 'amp', 'miso...	['DET', 'NOUN', 'PRON', 'VERB', 'ADJ', 'NOUN',...	[('the', 'DET'), ('idea', 'NOUN'), ('that', 'P...
4	0.00	0.0	12.20	0.702103	When women are targeted online, the abuse is m...	When women are targeted online, the abuse is m...	when women are targeted online the abuse is mo...	['when', 'women', 'are', 'targeted', 'online',...	['ADV', 'NOUN', 'AUX', 'VERB', 'ADV', 'DET', '...	[('when', 'ADV'), ('women', 'NOUN'), ('are', '...
...	...	...	...	...	...	...	...	...	...	...
3018	0.00	0.0	15.38	1.102073	ÄúWe will hold our leaders accountable and ma...	We will hold our leaders accountable and make...	we will hold our leaders accountable and make...	['we', 'will', 'hold', 'our', 'leaders', 'acco...	['PRON', 'AUX', 'VERB', 'PRON', 'NOUN', 'ADJ',...	[('we', 'PRON'), ('will', 'AUX'), ('hold', 'VE...
3019	0.00	0.0	18.52	0.188077	ÄúIt‚Äôs a real treat to be in this room with...	It s a real treat to be in this room with my ...	it s a real treat to be in this room with my ...	['it', 's', 'a', 'real', 'treat', 'to', 'be', ...	['PRON', 'AUX', 'DET', 'ADJ', 'NOUN', 'PART', ...	[('it', 'PRON'), ('s', 'AUX'), ('a', 'DET'), (...
3020	0.00	0.0	13.33	0.187262	ÄúHowever good you are, people still think th...	However good you are, people still think that...	however good you are people still think that ...	['however', 'good', 'you', 'are', 'people', 's...	['ADV', 'ADJ', 'PRON', 'AUX', 'NOUN', 'ADV', '...	[('however', 'ADV'), ('good', 'ADJ'), ('you', ...
3021	0.00	0.0	30.43	0.450141	ÄúAs women, we are the barometer of whether a...	As women, we are the barometer of whether a s...	as women we are the barometer of whether a so...	['as', 'women', 'we', 'are', 'the', 'barometer...	['SCONJ', 'NOUN', 'PRON', 'AUX', 'DET', 'NOUN'...	[('as', 'SCONJ'), ('women', 'NOUN'), ('we', 'P...
3022	0.00	0.0	21.43	0.871361	RT : .: when women meaningfully participate, w...	RT . when women meaningfully participate, we c...	rt when women meaningfully participate we can...	['rt', 'when', 'women', 'meaningfully', 'parti...	['VERB', 'ADV', 'NOUN', 'ADV', 'VERB', 'PRON',...	[('rt', 'VERB'), ('when', 'ADV'), ('women', 'N...

3023 rows × 108 columns

Figure 3.7: Twitter-metoo dataset after POS tagging and tokenization of the Preprocessing.

Tokenization breaks down a text document, such as a tweet, into individual units, which are typically words or subwords (subword tokenization is common for models like BERT). In the case of Twitter data, tweets are often short and contain multiple words, hashtags, mentions, and emojis. Tokenization helps segment these tweets into

meaningful components for analysis. Tokenization ensures that each word or subword is represented consistently, regardless of its length or complexity. This uniform representation is crucial for applying NLP models like BERT, as they require fixed-length input sequences. Tokenization can handle out-of-vocabulary words by breaking them into subword units and providing context for their interpretation. So Tokenized sequences are typically more computationally efficient to work with than raw text, as they enable faster processing and storage.

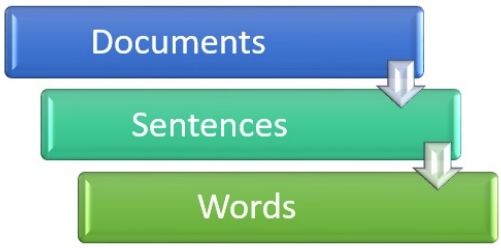


Figure 3.8: Tokenization procedure.

Stopwords removal is an essential preprocessing step in natural language processing (NLP) and text analysis tasks. Stopwords are common words in a language (e.g., ”the,” ”and,” ”is”) that are often removed from text data because they don’t carry significant meaning by themselves and can be noisy when performing natural language processing tasks like text analysis or information retrieval. In tasks like text classification, clustering, or topic modeling, the presence of stopwords can increase the dimensionality of the data, making it more challenging to extract meaningful patterns. Removing stopwords helps reduce the dimensionality and focuses the analysis on more significant terms. In topic modeling algorithms, stopwords can dominate topics and lead to less interpretable and coherent topics. Removing stopwords can lead to better, more interpretable topics. Performing tasks like text tokenization or part-of-speech tagging, and removing stopwords simplify the analysis process and make it more focused on content words.

In text processing, it’s often essential to employ normalization methods to transform words from their variations into their base form. This serves to reduce variability and align the words in the dataset with a predefined standard, enhancing computational efficiency by reducing the number of distinct features to manage. Lemmatization is a natural language processing (NLP) technique used to reduce words to their base or root form, known as the lemma. It is similar to stemming but more linguistically accurate because it considers the context and meaning of words. It helps in reducing the dimensionality of text data while preserving semantic integrity.

The main goal of lemmatization is to transform words into their dictionary or canonical form, making it easier to analyze and compare different word forms. For example, the lemma of the word ”running” would be ”run,” and the lemma of ”better” would be ”good.”

[ ]

id	author_id	created_at	referenced_tweets	text	retweets	...	OtherP	agency	selftext	clean_selftext	clean_selftext_v1	token	token_pos	pos_tag	stopword_removed	lemmatized_text
3871173632	14499829	2020-02-11T15:09:38.000Z	[{'type': 'replied_to', 'id': '122724808535885...	@DrTedros ?? BREAKING ?? \n\n"We now have a nam...	8154	...	18.18	0.087864	?? BREAKING ??\n\n"We now have a name for the ...	? BREAKING ? We now have a name for the diseas...	breaking we now have a name for the disease...	['breaking', 'we', 'now', 'have', 'a', 'name',...	['VERB', 'PRON', 'ADV', 'VERB', 'DET', 'NOUN',...	[('breaking', 'VERB'), ('we', 'PRON'), ('now',...	breaking disease covid spell covid hyphen covid	break disease covid spell covid hyphen covid
7253407744	14499829	2020-02-11T15:10:39.000Z	[{'type': 'replied_to', 'id': '122724833387117...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	416	...	30.91	0.299545	"Under agreed guidelines between WHO, the &am...	Under agreed guidelines between WHO, the amp ,...	under agreed guidelines between who the amp w...	['under', 'agreed', 'guidelines', 'between', '...	['ADP', 'VERB', 'NOUN', 'ADP', 'PRON', 'DET', ...	[('under', 'ADP'), ('agreed', 'VERB'), ('guide...	agreed guidelines amp find refer geographical ...	agree guideline amp find refer geographical lo...
6127159298	14499829	2020-02-11T15:10:57.000Z	[{'type': 'replied_to', 'id': '122724858725340...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	282	...	34.04	0.276209	"Having a name matters to prevent the use of o...	Having a name matters to prevent the use of ot...	having a name matters to prevent the use of ot...	['having', 'a', 'name', 'matters', 'to', 'prev...	['VERB', 'DET', 'NOUN', 'VERB', 'PART', 'VERB'...	[('having', 'VERB'), ('a', 'DET'), ('name', 'N...	having matters prevent use names inaccurate st...	have matter prevent use name inaccurate stigma...
6552878083	14499829	2020-02-11T15:11:43.000Z	[{'type': 'replied_to', 'id': '122724866612715...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	485	...	29.51	0.016819	"As of 6am Geneva time this morning, there wer...	As of am Geneva time this morning, there were ...	as of am geneva time this morning there were ...	['as', 'of', 'am', 'geneva', 'time', 'this', '...	['ADP', 'ADP', 'NOUN', 'NOUN', 'NOUN', 'DET', ...	[('as', 'ADP'), ('of', 'ADP'), ('am', 'NOUN'),...	geneva time morning confirmed cases reported a...	geneva time morning confirm case report amp tr...
2342396928	14499829	2020-02-11T15:12:44.000Z	[{'type': 'replied_to', 'id': '122724898692006...	@DrTedros @WHOWPRO @WHOSEARO @WHO_Europe @WHOE...	55	...	48.72	1.223480	"I also briefed the Secretary-General and we ...	I also briefed the SecretaryGeneral and we agr...	i also briefed the secretarygeneral and we agr...	['i', 'also', 'briefed', 'the', 'secretarygene...	['PRON', 'ADV', 'VERB', 'DET', 'ADJ', 'CCONJ',...	[('i', 'PRON'), ('also', 'ADV'), ('briefed', '...	briefed secretarygeneral agreed leverage power...	brief secretarygeneral agree leverage power en...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7166261255	2809959174	2020-04-11T22:17:36.000Z	NaN	It's adversity that gives us the opportunity t...	209	...	6.00	0.956527	It's adversity that gives us the opportunity t...	It is adversity that gives us the opportunity ...	it is adversity that gives us the opportunity ...	['it', 'is', 'adversity', 'that', 'gives', 'us...	['PRON', 'AUX', 'NOUN', 'PRON', 'VERB', 'PRON'...	[('it', 'PRON'), ('is', 'AUX'), ('adversity', ...	adversity gives opportunity work overcome cont...	adversity give opportunity work overcome conti...
9889692674	86626845	2020-04-11T22:38:05.000Z	NaN	The list of successful countries vs	560	...	7.50	0.461828	The list of successful countries	The list of successful countries vs is	the list of successful countries vs is	['the', 'list', 'of', 'successful',	['DET', 'NOUN', 'ADP', 'ADJ'	[('the', 'DET'), ('list', 'NOUN')	list successful countries vs	list successful country vs extend

Figure 3.9: Twitter-Covid dataset after eliminating stopwords and lemmatization of the Preprocessing.

[ ] df

	Unnamed: 0.2	Unnamed: 0.1	Unnamed: 0	Source (A)	id	text	created_at	community	WC	Analytic	...	OtherP	agency	selftext	clean_selftext	clean_selftext_v1	token	token_pos	pos_tag	stopword_removed	lemmatized_text
0	0	0	0	0	1013189209488790016	#PrideMonth may be ending today, but we will c...	2018 Sat Jun 30 22:35:06 +0000	12	35	60.73	...	14.29	0.564734	may be ending today, but we will continue to a...	may be ending today, but we will continue to a...	may be ending today but we will continue to ad...	['may', 'be', 'ending', 'today', 'but', 'we', ...	['AUX', 'AUX', 'VERB', 'NOUN', 'CCONJ', 'PRON'...	[('may', 'AUX'), ('be', 'AUX'), ('ending', 'VE...	ending today continue advocate right live true...	end today continue advocate right live true se...
1	1	1	1	1	1013143915308950016	‚ÄúThis disaster leaves us in extreme poverty....	2018 Sat Jun 30 19:35:07 +0000	11	36	98.14	...	13.89	0.052641	‚ÄúThis disaster leaves us in extreme poverty....	This disaster leaves us in extreme poverty. M...	this disaster leaves us in extreme poverty ma...	['this', 'disaster', 'leaves', 'us', 'in', 'ex...	['DET', 'NOUN', 'VERB', 'PRON', 'ADP', 'ADJ', ...	[('this', 'DET'), ('disaster', 'NOUN'), ('leav...	disaster leaves extreme poverty magdalena suta...	disaster leave extreme poverty magdalena sutam...
2	2	2	2	2	1013128815575469952	We want more women on ballots □üöżÖ∏è when we ...	2018 Sat Jun 30 18:35:07 +0000	13	37	86.66	...	8.11	0.558930	We want more women on ballots □üöżÖ∏è when we ...	We want more women on ballots when we go out t...	we want more women on ballots when we go out t...	['we', 'want', 'more', 'women', 'on', 'ballots...	['PRON', 'VERB', 'ADJ', 'NOUN', 'ADP', 'NOUN',...	[('we', 'PRON'), ('want', 'VERB'), ('more', 'A...	want women ballots vote lucia ran candidate lo...	want woman ballot vote lucia run candidate loc...
3	3	3	3	3	1013098625768000000	‚ÄúThe idea that sexism & misogyny in onli...	2018 Sat Jun 30 16:35:09 +0000	7	31	93.26	...	16.13	0.315561	‚ÄúThe idea that sexism & misogyny in onli...	The idea that sexism amp misogyny in online s...	the idea that sexism amp misogyny in online s...	['the', 'idea', 'that', 'sexism', 'amp', 'miso...	['DET', 'NOUN', 'PRON', 'VERB', 'ADJ', 'NOUN',...	[('the', 'DET'), ('idea', 'NOUN'), ('that', 'P...	idea sexism amp misogyny online spaces psychol...	idea sexism amp misogyny online space psycholo...
4	4	4	4	4	1013075989885340032	When women are targeted online, the abuse is m...	2018 Sat Jun 30 15:05:12 +0000	0	41	95.24	...	12.20	0.702103	When women are targeted online, the abuse is m...	When women are targeted online, the abuse is m...	when women are targeted online the abuse is mo...	['when', 'women', 'are', 'targeted', 'online',...	['ADV', 'NOUN', 'AUX', 'VERB', 'ADV', 'DET', '...	[('when', 'ADV'), ('women', 'NOUN'), ('are', '...	women targeted online abuse likely sustained s...	woman target online abuse likely sustain seria...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3018	3018	3018	3018	3018	870081565363232000	‚ÄúWe will hold our leaders accountable and ma...	2017 Thu Jun 01 00:56:25 +0000	15	26	54.88	...	15.38	1.102073	‚ÄúWe will hold our leaders accountable and ma...	We will hold our leaders accountable and make...	we will hold our leaders accountable and make...	['we', 'will', 'hold', 'our', 'leaders', 'acco...	['PRON', 'AUX', 'VERB', 'PRON', 'NOUN', 'ADJ',...	[('we', 'PRON'), ('will', 'AUX'), ('hold', 'VE...	hold leaders accountable sure deliver promises	hold leader accountable sure deliver promise
3019	3019	3019	3019	3019	870078835689127040	‚ÄúIt‚Äôs a real treat to be in this room with...	2017 Thu Jun 01 00:45:34 +0000	12	27	97.84	...	18.52	0.188077	‚ÄúIt‚Äôs a real treat to be in this room with...	It s a real treat to be in this room with my ...	it s a real treat to be in this room with my ...	['it', 's', 'a', 'real', 'treat', 'to', 'be', ...	['PRON', 'AUX', 'DET', 'ADJ', 'NOUN', 'PART', ...	[('it', 'PRON'), ('s', 'AUX'), ('a', 'DET'), (...	s real treat room heroes nobel peace laureate ...	s real treat room hero nobel peace laureate forum
3020	3020	3020	3020	3020	870076329219288960	‚ÄúHowever good you are, people still think th...	2017 Thu Jun 01 00:35:36 +0000	6	30	25.89	...	13.33	0.187262	‚ÄúHowever good you are, people still think th...	However good you are, people still think that...	however good you are people still think that ...	['however', 'good', 'you', 'are', 'people', 's...	['ADV', 'ADJ', 'PRON', 'AUX', 'NOUN', 'ADV', '...	[('however', 'ADV'), ('good', 'ADJ'), ('you', ...	good people think good men twice good	good people think good man twice good
3021	3021	3021	3021	3021	870074711409078016	‚ÄúAs women, we are the barometer of whether a...	2017 Thu Jun 01 00:29:11 +0000	19	23	88.21	...	30.43	0.450141	‚ÄúAs women, we are the barometer of whether a...	As women, we are the barometer of whether a s...	as women we are the barometer of whether a so...	['as', 'women', 'we', 'are', 'the', 'barometer...	['SCONJ', 'NOUN', 'PRON', 'AUX', 'DET', 'NOUN'...	[('as', 'SCONJ'), ('women', 'NOUN'), ('we', 'P...	women barometer society thriving cc	women barometer society thrive cc
3022	3022	3022	3022	3022	870068740947405952	RT @LeymahRGbowee: @AminaJMohammed: when wome...	2017 Thu Jun 01 00:05:27 +0000	14	14	48.63	...	21.43	0.871361	RT : .: when women meaningfully participate, w...	RT . when women meaningfully participate, we c...	rt when women meaningfully participate we can...	['rt', 'when', 'women', 'meaningfully', 'parti...	['VERB', 'ADV', 'NOUN', 'ADV', 'VERB', 'PRON',...	[('rt', 'VERB'), ('when', 'ADV'), ('women', 'N...	rt women meaningfully participate immediate re...	rt woman meaningfully participate immediate re...

3023 rows × 110 columns

[ ]

Figure 3.10: Twitter-metoo dataset after eliminating stopwords and lemmatization of the Preprocessing.

3.3 Empty row elimination

To ensure data quality, integrity, and the successful application of various data analysis and modeling techniques, empty row elimination is a vital step in data preprocessing. It helps ensure that the dataset is clean, complete, and ready for further analysis, leading to more accurate and reliable results.

In this study, a portion of the collected data was entirely deleted by manual screening. In some cases, only parts of the data were removed by applying the superficial cleaning and deep cleaning process. As a result, some rows were completely blank. Empty rows or rows with missing values can introduce noise into the dataset and affect the quality and reliability of any analysis or modeling conducted on the data. For this reason, we have to drop these rows to avoid computation problems because Empty rows can lead to errors, crashes, or unexpected behavior in such algorithms.

Listing 3.3: Empty row elimination

```
# Check and remove rows with empty 'lematised_text' column
df = df[df['lemmatized_text'].notna()]

# Save the cleaned data to a new CSV file
output_file = 'cleaned _output_file.csv'

df.to_csv(output_file, index=False)
print("done")
```

4

Tools and Techniques

Tools and techniques refer to the various instruments, methods, procedures, or strategies employed to achieve a particular goal or complete a task efficiently and effectively. These tools and techniques serve as aids or methodologies to streamline processes, solve problems, analyze data, communicate effectively, and achieve desired outcomes in various domains. Choosing the right tools and techniques for a specific task or goal is essential to optimize productivity and achieve success. In this section, we will delve into the tools and Techniques used to achieve the objectives of the research or project, shedding light on the practical aspects that underpin the work's validity and rigor.

The choice of algorithms was carefully made to align with the project's unique needs, aiming for the best possible results. The process of selecting these algorithms entailed a thorough examination of analogous projects and an extensive literature review to pinpoint the algorithms most suited to the project's objectives. So it serves as a roadmap for understanding how data was collected, analyzed, and interpreted, offering transparency and reproducibility.

4.1 Working environments

Google Colab, short for Google Colaboratory, is a cloud-based platform provided by Google for working with and running Jupyter notebooks. It can be considered a sophisticated working environment for various data science, machine learning, and research tasks. Google Colab eliminates the need for powerful local hardware and allows us to work from virtually anywhere. It also supports Jupyter notebooks, which are interactive documents that combine code, text, and visualizations. This makes it an excellent tool for data analysis, machine learning, and collaborative research. Google Colab comes with many popular data science and machine learning libraries pre-installed and provides free access to GPU (Graphics Processing Unit) and TPU (Tensor Processing Unit) resources, which can significantly accelerate deep learning tasks.

Using CUDA in this study can offer significant benefits in terms of performance and speed, particularly for computationally intensive tasks. CUDA enables high-performance parallel computing by allowing developers to write programs in C, C++, Python, or other languages and execute them on the GPU, taking advantage of the GPU's massively parallel architecture. This is especially useful for tasks that involve complex calculations, data processing, and simulations, as GPUs can perform these operations much faster than traditional central processing units (CPUs).

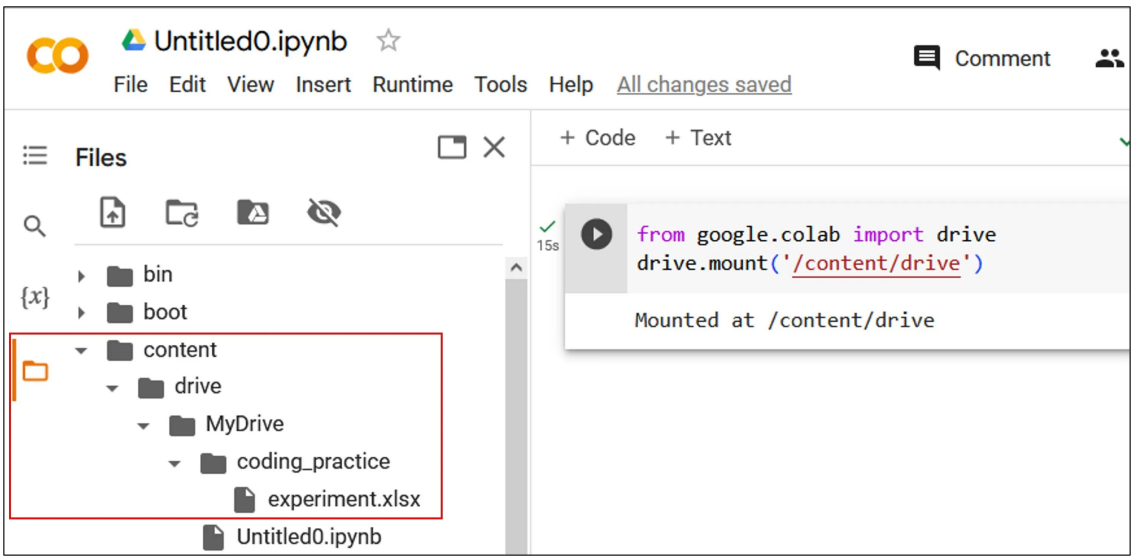


Figure 4.1: Working environment sample of the Google Colaboratory.

4.2 BERTopic

BERTopic is a technique that falls under the broader field of machine learning (ML) and, more specifically, natural language processing (NLP). It leverages deep learning models, specifically the BERT (Bidirectional Encoder Representations from Transformers) model, which is a neural network architecture, to extract embeddings and then combine these embeddings with traditional machine learning techniques such as clustering. Usually, BERTopic supports guided, supervised, semi-supervised, manual, long-document, hierarchical, class-based, dynamic, and online topic modeling [11].

Topic modeling is widely used in information retrieval, document organization, content recommendation, and text analysis. To do a task like the following it consist of several steps. In this study, Topic models were created by going through a series of steps, where different model parameters were carefully selected based on their appropriateness for the specific datasets. The subsequent list outlines the key stages of the topic analysis process, including the components of the topic model.

Overall, BERTopic generates topic representations through three steps. First, each document is converted to its embedding representation using a pre-trained language model. Then, before clustering these embeddings, the dimensionality of the resulting embeddings is reduced to optimize the clustering process. lastly, from the clusters of the documents, topic representations are extracted using the TF-IDF method [12]. we will delve into a detailed explanation of each step later on this section.

So Creating Topic Model by Defining the Following Items:
– Embedding Model
– UMAP Model
–HDBSCAN Model
–Topic Representation

4.2.1 Topic Model

Embedding Model

BERTopic specifically uses "all-MiniLM-L6-v2" from Hugging Face as its default sentence-transformer model, which can map sentences and paragraphs to vectors that are suitable for clustering or semantic search tasks and works well for English documents. "all-MiniLM-L6-v2" is used when the language is English otherwise paraphrase-multilingual-MiniLM-L12-v2 version is used when language is "multilingual".

In this study, we want to observe the efficacy of BERTopic keeping this embedding model as default. The same sentence transformer model was used for creating embeddings in the Twitter datasets as shown in the Code Listing 4.2.1.1: Embedding Model

```
from sentence_transformers import SentenceTransformer
embedding_model = NONE
```

UMAP Model

A crucial element in BERTopic is reducing the dimensionality of the input embeddings. High-dimensional embeddings can make clustering challenging due to the problems associated with high-dimensional spaces, often referred to as the "curse of dimensionality." To address this issue, a solution is to decrease the dimensionality of the embeddings, creating a more manageable dimensional space that allows clustering algorithms to operate effectively.

UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction technique commonly used in combination with BERTopic. UMAP is particularly effective for reducing the dimensionality of high-dimensional BERT embeddings while preserving the intrinsic structure and relationships within the data. UMAP is useful for data exploration, clustering, and classification tasks, and has been successfully applied to a variety of datasets, including text data [13].

The basic form of a UMAP model can be defined as below:
Listing 4.2.1.2: UMAP model

```
from umap import UMAP
umap_model = UMAP(n_neighbors=n_neighbors, n_components=n_components, min_dist=min_dist, random_state=random_state, metric=metric).fit(embeddings)
```

Here,

- n_neighbors: sets the number of neighbors to consider when constructing the low-dimensional representation.
- n_components: specifies the number of dimensions in the reduced space.
- min_dist: controls the minimum distance between points in the low-dimensional space.
- metric: This parameter sets the distance metric used for computing the pairwise distances between data points in the high-dimensional space.
- random_state: This parameter sets the random seed or state for the UMAP model. Setting a specific random state ensures that the UMAP model's results are reproducible.

HDBSCAN Model

Once we've reduced the dimensionality of our input embeddings, the next step is to group them into clusters based on their similarity. This clustering process is crucial because the quality of our clustering technique directly affects the accuracy of the topic representations we extract.HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm used for discovering clusters within data, particularly in cases where clusters have irregular shapes and varying densities. It builds a hierarchical representation of clusters, allowing it to automatically determine the number of clusters in the data. By default, BERTopic employs HDBSCAN for carrying out the clustering process.

Combining UMAP and HDBSCAN, these two methods enabled the discovery of significant groups of data points that share similarities, all the while maintaining the underlying topological structure of the data.

The basic form of an HDBSCAN model can be defined as below:

Listing 4.2.1.3: HDBSCAN model

```
Default HDBSCAN Model from hdbscan import HDBSCAN
hdbscan_model = HDBSCAN(min_cluster_size=min_cluster_size,metric=metric,
min_samples=min_samples,cluster_selection_method=cluster_selection_met
hod,)
```

Here,

• min_cluster_size: sets the minimum number of points required to form a cluster.

• min_samples: the minimum number of samples in a neighborhood for a point to be considered a core point.

• metric: This parameter specifies the distance metric used for clustering.

• prediction_data: This parameter allows the HDBSCAN model to predict the cluster labels for new or unseen data points based on the clustering structure it has learned from the training data. This is useful for making predictions or classifying new data points using the established clustering model.

min_topic_size, nr_topics and calculate_probabilities

• min_topic_size: This parameter sets a minimum size requirement for the topics. Only topics with a minimum number of data points equal to or greater than this value will be considered.

• nr_topics: This parameter specifies the desired number of topics you want to extract from your data.

• calculate_probabilities: When set to True, this parameter enables BERTopic to calculate topic probabilities for each document, indicating the likelihood of a document belonging to a particular topic.

The final version of a basic topic model is given below:

Listing 4.2.1.4: A basic form of a Topic Model.

```
Default BERTopic Model from BERTopic imp o r t BERTopic
t o p i c _ m o d e l = BERTopic ( em b e d di n g _mo d el = em b e d ding _mo d el , umap_model=umap_model
, h d b s c a n _mo d e l = h d b s c a n _mo d el , c t f i d f _ m o d e l = c t f i d f _ m o d e l , n r _ t o p i c s = n r _ t
o p i c s )
```

Topic Representation

At first, we created document embeddings using a pre-trained language model to obtain document-level information. secondly, we first reduce the dimensionality of document embeddings before creating semantically similar clusters of documents that each represent a distinct topic. These independent steps allow for a flexible topic model [12].

The key element in BERTopic is its use of Bag-of-Words representation along with TF-IDF weighting. TF-IDF is a numerical statistic that shows the relevance of keywords to some specific documents or it can be said that it provides those keywords, using which some specific documents can be identified or categorised [14]. So this approach is efficient and allows for the rapid generation of topic keywords, independent of the clustering process. Consequently, it enables easy and swift updates to topics after model training, eliminating the necessity for re-training.

Once the model is computed, we can output the most important topics or list of topics. Notably, Topic 0 with a count of -1 will always represent outliers and should not be considered any further [15]. The basic command form of topic representation of the model can be defined as below:
Listing 4.2.1.5: Topic Representation
topic_model.get_topic_info()

```
[ ]  1 topic_model.get_topic_info()
```

	Topic	Count	Name	Representation	Representative_Docs
0	0	218	0_et_le_la_je	[et, le, la, je, les, pas, pour, que, des, une]	[un avis de mais vue que mon amid na pas pay ...
1	1	4782	1_place_good_food_just	[place, good, food, just, like, great, time, r...	[this place used to be called the german corne...

Figure 4.2: A Sample of Topic representation.

4.2.2 Fine-tuning

While BERTopic provides good default performance, it is also an important task to do fine-tune its hyperparameters to align with the specific requirements of specific use cases because Hyperparameter tuning is an important optimization step for building a good topic model [16]. We can define any parameters in UMAP and HDBSCAN to optimize for the best performance.

As our predominant goal was to check the working effectiveness of the BERTopic, for this reason initially we chose four parameters for tuning to observe each step's efficiency of the BERTopic. Specifically, we choose two parameters from the UMAP model, one from the HDBSCAN model and the remaining one is nr_topic in the topic model.

To explore different combinations and evaluate the model's performance initially we have we made a range for each parameter such as for n_neighbors in the UMAP model we decided to start from the value of 5 and increased also 5 for each time within the range of 5 to 20. We did the same for n_components in the UMAP model and n_samples for the HDBSCAN model. Moreover, we had chosen a different range for nr_topic which is 110, 75, and 40. After running each combination we got [192 rows x 7 columns] where we had obtained n_components, n_neighbors, n_samples, nr_topics, execution_time_sec, topics, num_topics as a column in the data frame. We also save each model with a selected combination from the nested loop to do further evaluation.

On the other hand, to make observations more precise, we wanted to investigate more for n_samples and nr_topics because we desired to observe, what is the impact of the attribute of n_samples and nr_topics on the model. For that purpose, again we made different combinations where n_components and n_neighbors are fixed with the value of 5 and made a range for n_samples and nr_topics in specific. For n_samples we made a range from 5 to 17 with interval 2; whereas for nr_topics we made a range from 5 to 35 with interval 10. So the ultimate range for n_samples was 5, 7, 9, 11, 13, 15, 17 and for nr_topics was 5, 15, 25, 35. After running each combination we got [27 rows x 7 columns] where we had obtained n_components, n_neighbors, n_samples, nr_topics, execution_time_sec, topics, num_topics as a column in the data frame for the Twitter-metoo dataset. On the contrary, to compare the performance we run the shorter combinations for the Twitter-Covid dataset where n_components and n_neighbors are fixed with the value of 5; n_samples was 5, 7, 9, 11, 13, 15, 17 and nr_topics was 5, 15, 25, 35.

	n_components	n_neighbors	n_samples	nr_topics	execution_time_sec	\
0	5	5	5	110	42.259368	
1	5	5	5	75	31.607367	
2	5	5	5	40	28.733448	
3	5	5	10	110	28.329693	
4	5	5	10	75	25.589742	
..	...	...	...	...	...	
187	20	20	15	75	30.015115	
188	20	20	15	40	31.737725	
189	20	20	20	110	34.928932	
190	20	20	20	75	30.694806	
191	20	20	20	40	31.387585	

	topics	num_topics
0	[-1, 7, 87, 69, 69, -1, 79, 19, 79, 87, 79, 10...	102
1	[-1, 18, 1, 6, 6, -1, 4, 33, 4, 1, 4, 0, 4, 6,...	75
2	[-1, 1, 0, 2, 2, -1, 0, 14, 0, 0, 0, 0, 0, 2, ...	40
3	[38, -1, -1, 19, 19, 19, 31, 13, 31, -1, 31, 3...	43
4	[38, -1, -1, 19, 19, 19, 31, 13, 31, -1, 31, 3...	43
..	...	...
187	[-1, -1, 20, 14, 14, -1, -1, -1, -1, 20, -1, -...	26
188	[-1, -1, 20, 14, 14, -1, -1, -1, -1, 20, -1, -...	26
189	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	2
190	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	2
191	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	2

[192 rows x 7 columns]

Figure 4.3: Result of the fine-tuning with four parameters for the Twitter-metoo dataset of the BERTopic modeling.

	n_components	n_neighbors	n_samples	nr_topics	execution_time_sec	topics	num_topics
0	5	5	5	5	60.700305	[-1, 0, 0, 0, 0, -1, 0, 2, 0, 0, 0, 0, 0, 0...	5
1	5	5	5	15	32.052962	[-1, 0, 0, 0, 0, -1, 0, 2, 0, 0, 0, 0, 0, 0...	15
2	5	5	5	25	32.784346	[-1, 0, 0, 1, 1, -1, 0, 11, 0, 0, 0, 0, 0, 1, ...	25
3	5	5	5	35	35.748884	[-1, 1, 0, 2, 2, -1, 0, 13, 0, 0, 0, 0, 0, 2, ...	35
4	5	5	7	5	32.414424	[-1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...	5
5	5	5	7	15	31.873893	[-1, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0,...	15
6	5	5	7	25	31.058325	[-1, 3, 0, 1, 1, 1, 0, 8, 0, 0, 0, 0, 0, 1, 1,...	25
7	5	5	7	35	30.546888	[-1, 19, 0, 1, 1, 1, 0, 22, 0, 0, 0, 0, 0, 1, ...	35
8	5	5	9	5	30.555057	[-1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...	5
9	5	5	9	15	32.696713	[-1, -1, 0, 2, 2, 2, 0, 3, 0, 0, 0, 0, 0, 2, 0...	15
10	5	5	9	25	33.325244	[-1, -1, 0, 1, 1, 1, 0, 15, 0, 0, 0, 0, 0, 1, ...	25
11	5	5	9	35	30.776125	[-1, -1, 16, 0, 0, 0, 1, 22, 1, 16, 1, 4, 1, 0...	35
12	5	5	11	5	30.775454	[0, -1, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, ...	5
13	5	5	11	15	31.035789	[0, -1, -1, 0, 0, 0, 0, 2, 0, -1, 0, 0, 0, 0, ...	15
14	5	5	11	25	31.000767	[0, -1, -1, 1, 1, 1, 0, 20, 0, -1, 0, 0, 0, 1,...	25
15	5	5	11	35	31.702331	[0, -1, -1, 1, 1, 1, 0, 24, 0, -1, 0, 0, 0, 1,...	35
16	5	5	13	5	30.830236	[-1, -1, -1, 0, 0, 0, -1, 1, -1, -1, -1, 0, -1...	5
17	5	5	13	15	30.745393	[-1, -1, -1, 0, 0, 0, -1, 10, -1, -1, -1, 0, -...	15
18	5	5	13	25	30.496169	[-1, -1, -1, 1, 1, 1, -1, 19, -1, -1, -1, 0, -...	25
19	5	5	13	35	29.867522	[-1, -1, -1, 1, 1, 1, -1, 27, -1, -1, -1, 0, -...	35
20	5	5	15	5	29.117975	[0, -1, -1, 0, 0, 0, -1, 1, -1, -1, -1, 0, -1,...	5
21	5	5	15	15	30.035850	[0, -1, -1, 0, 0, 0, -1, 3, -1, -1, -1, 0, -1,...	15
22	5	5	15	25	34.462914	[0, -1, -1, 1, 1, 1, -1, 10, -1, -1, -1, 0, -1...	25
23	5	5	15	35	30.049447	[24, -1, -1, 12, 12, 12, -1, 13, -1, -1, -1, 2...	26
24	5	5	17	5	29.242348	[0, -1, -1, 0, 0, 0, -1, -1, -1, -1, -1, 0, -1...	5
25	5	5	17	15	29.655994	[0, -1, -1, 0, 0, 0, -1, -1, -1, -1, -1, 0, -1...	15
26	5	5	17	25	30.161395	[23, -1, -1, 13, 13, 13, -1, -1, -1, -1, -1, 2...	25
27	5	5	17	35	30.145634	[23, -1, -1, 13, 13, 13, -1, -1, -1, -1, -1, 2...	25

Figure 4.4: Result of the fine-tuning with two parameters for the Twitter-metoo dataset of the BERTopic modeling.

```
results_df
```

	n_components	n_neighbors	n_samples	nr_topics	execution_time_sec	topics	num_topics
0	5	5	5	5	139.792571	[0, 0, 0, 0, 0, 0, -1, -1, -1, 0, 0, 0, -1, -1,...	5
1	5	5	5	15	85.426112	[0, 0, 0, 0, 0, 0, -1, -1, -1, 0, 0, 0, -1, -1,...	15
2	5	5	5	25	89.661641	[0, 1, 0, 0, 1, 0, -1, -1, -1, 1, 0, 1, -1, -1,...	25
3	5	5	5	35	87.466829	[0, 1, 0, 0, 1, 0, -1, -1, -1, 1, 0, 1, -1, -1,...	35
4	5	5	7	5	54.948056	[0, 0, 0, 0, 0, 0, -1, -1, -1, 0, 0, 0, -1, -1,...	5
5	5	5	7	15	60.495787	[0, 0, 0, 0, 0, 0, -1, -1, -1, 0, 0, 0, -1, -1,...	15
6	5	5	7	25	57.897527	[0, 0, 0, 0, 0, 0, -1, -1, -1, 0, 0, 0, -1, -1,...	25
7	5	5	7	35	57.163404	[0, 5, 0, 0, 5, 0, -1, -1, -1, 0, 0, 1, -1, -1,...	35
8	5	5	9	5	39.773302	[0, 0, 0, 0, 0, 0, 0, 0, -1, -1, 0, 0, -1, -1, 0,...	5
9	5	5	9	15	42.779464	[0, 0, 0, 0, 0, 0, 0, 0, -1, -1, 0, 0, 0, -1, -1, 1,...	15
10	5	5	9	25	42.735903	[0, 2, 0, 0, 2, 0, 1, -1, -1, 1, 0, -1, -1, 3,...	25
11	5	5	9	35	42.297092	[0, 5, 0, 0, 5, 0, 2, -1, -1, 2, 0, -1, -1, 7,...	35
12	5	5	11	5	34.903953	[0, 0, 0, 0, 0, -1, 0, -1, -1, -1, 0, 0, -1, -1, ...	5
13	5	5	11	15	33.580067	[0, 0, 0, 0, 0, -1, 0, -1, -1, -1, 1, 0, -1, -1, ...	15
14	5	5	11	25	33.509704	[0, 3, 0, 0, 0, -1, 0, -1, -1, -1, 8, 0, 0, -1, -1, ...	25
15	5	5	11	35	34.245759	[0, 6, 0, 0, 0, -1, 0, -1, -1, -1, 13, 0, -1, -1,...	35
16	5	5	13	5	30.589635	[0, 0, 0, 0, -1, -1, 0, -1, -1, -1, 0, 0, 0, -1, -1,...	5
17	5	5	13	15	31.007399	[0, 3, 0, -1, -1, 0, -1, -1, -1, 1, 0, 0, -1, -1,...	15
18	5	5	13	25	31.497769	[0, 4, 0, -1, -1, 0, -1, -1, -1, 8, 0, 0, -1, -1,...	25
19	5	5	13	35	33.375579	[0, 5, 0, -1, -1, 0, -1, -1, -1, 10, 0, 0, -1, -1...	35
20	5	5	15	5	29.699422	[0, 0, 0, 0, -1, -1, 0, -1, -1, 0, 0, 0, -1, -1, ...	5
21	5	5	15	15	29.650414	[0, 3, 0, -1, -1, 0, -1, -1, 0, 1, 0, 0, -1, -1, ...	15
22	5	5	15	25	29.231499	[0, 5, 0, -1, -1, 0, -1, -1, 0, 11, 0, 0, -1, -1,...	25
23	5	5	15	35	29.559614	[0, 8, 0, -1, -1, 0, -1, -1, 0, 32, 0, 0, -1, -1,...	35
24	5	5	17	5	25.476224	[0, 0, 0, 0, -1, -1, 0, -1, -1, 0, 0, 0, -1, -1, ...	5
25	5	5	17	15	26.735473	[0, 3, 0, -1, -1, 0, -1, -1, 0, 1, 0, 0, -1, -1, ...	15
26	5	5	17	25	28.884964	[0, 6, 0, -1, -1, 0, -1, -1, 0, 7, 0, 0, -1, -1, ...	25
27	5	5	17	35	27.836486	[0, 7, 0, -1, -1, 0, -1, -1, 0, 9, 0, 0, -1, -1, ...	35

Figure 4.5: Result of the fine-tuning with two parameters for the Twitter-Covid dataset of the BERTopic modeling.

4.2.3 Reducing Outliers

Documents that do not fit into any of the defined topics are frequently encountered as outliers with the value of topic number -1. Nonetheless, it could be preferable in some applications to reduce the number of outlier records. It was chosen to use the outlier reduction procedure in order to prevent data loss throughout this study. After a BERTopic model has been trained, BERTopic supports multiple ways of minimizing the outliers or noise from the datasets. We have preferred the embeddings technique for both of the datasets because using the embeddings of these documents and calculating the cosine similarity between each outlier document's embedding and the embeddings of the selected themes is one method to limit the number of outliers. In doing so, the best topic embedding for every outlier is found. Pre-computed embeddings were used to expedite this process and avoid redundant computation [17].

Listing 4.2.2: Reduce outliers using the embeddings strategy.

```
# Reduce outliers using the ‘embeddings‘ strategy
new_topics = topic_model.reduce_outliers(docs, topics, strategy=”embeddings”)
```

4.2.4 Vocabulary Creation

For further analysis and to map topics to indices we needed to create a vocabulary from a list of unique topics. For that purpose, we started by creating a counter object from the 'collections' module. This object is used to count the frequency of each topic in the dataset. Then we had to extract the unique topics from the counter object by converting its keys into a list. This gave us a list of topics without duplicates. Next, we aimed to sort the unique topics alphabetically. After that, we created a vocabulary dictionary where each unique topic is paired with a numerical index. This dictionary allows us to represent topics with numerical values, making it easier to work with them further. Finally, we printed the vocabulary to confirm that it's been successfully created. In this experiment, we need to create this custom-made vocabulary many times for each model because later we need this individual topic vocabulary list to create a sparse matrix and dense.

```
# Print the vocabulary
print("vocabulary is ready:\n")
print(vocabulary)
```

```
vocabulary is ready:

{0: 0, 1: 1, 2: 2, 3: 3, 4: 4, 5: 5, 6: 6, 7: 7, 8: 8, 9: 9, 10: 10, 11: 11, 12: 12, 13: 13, 14: 14, 15: 15, 16: 16, 17: 17, 18: 18, 19: 19, 20: 20, 21: 21, 22: 22, 23: 23}
```

Figure 4.6: Sample creation of the vocabulary from one of the data sets.

4.3 Sparse Matrix

In the field of natural language processing, the ideas of “sparsity” and ”density” lead to the effective design and construction of these matrices for all high-dimensional data processing use-cases. Numerical representations of the texts are designed to build matrices that hold pertinent information from those texts, enabling the application of data analysis as well. Sparse matrix will be described in this section and dense matrix will be described after this section.

Usually, in mathematics, a matrix is any type of information or values that are presented in the form of rows and columns format and a sparse matrix is a matrix where the majority of elements belong to zero. Using a sparse matrix has the following two main advantages:

• Storage: Less memory can be required to store only those elements because there are fewer non-zero elements than zeros.

• Computing time: By logically creating a data structure that only traverses non-zero components, computing time can be reduced.

To convert the text data into a sparse matrix format, we need TF-IDF or Count Vectorization method, to represent the textual information sparsely. In this study, we used Scipy. Sparse matrix which is in the CSR format.

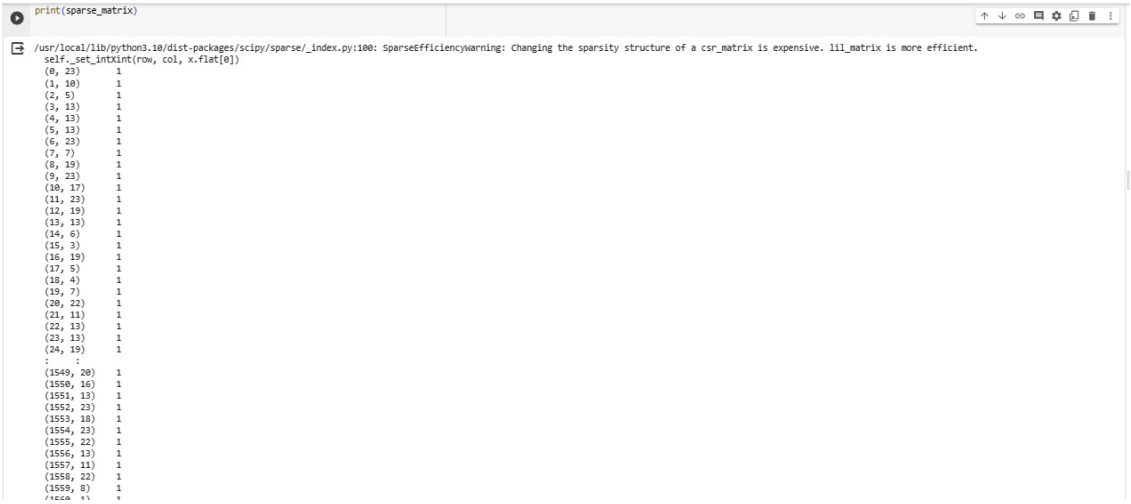


Figure 4.7: Sample creation of one of the sparse matrixes from the data set.

4.4 Dense Matrix

In contrast to the above matrix, a dense matrix comprises mostly non-zero elements. That means the dense matrix is the opposite of the sparse matrix. In this study, we have created two dense matrices, one refers to the document-topic matrix obtained after applying the BERTopic algorithm to your documents. The output of the Bertopic algorithm is a document-topic matrix, where each row corresponds to a document, and each column corresponds to a topic. The values in the matrix represent the strength of the association between each document and each topic.

On the other hand, we have used the CountVectorizer from scikit-learn to convert a collection of text documents to a matrix of token counts. We have essentially created a Document-Term Matrix (DTM), where each row corresponds to a document, and each column corresponds to a unique term (word). In this study, toarray() is used to convert the sparse matrix (dtm) into a dense matrix (dense_dtm), where each element of the matrix represents the count of a particular term in the corresponding document.

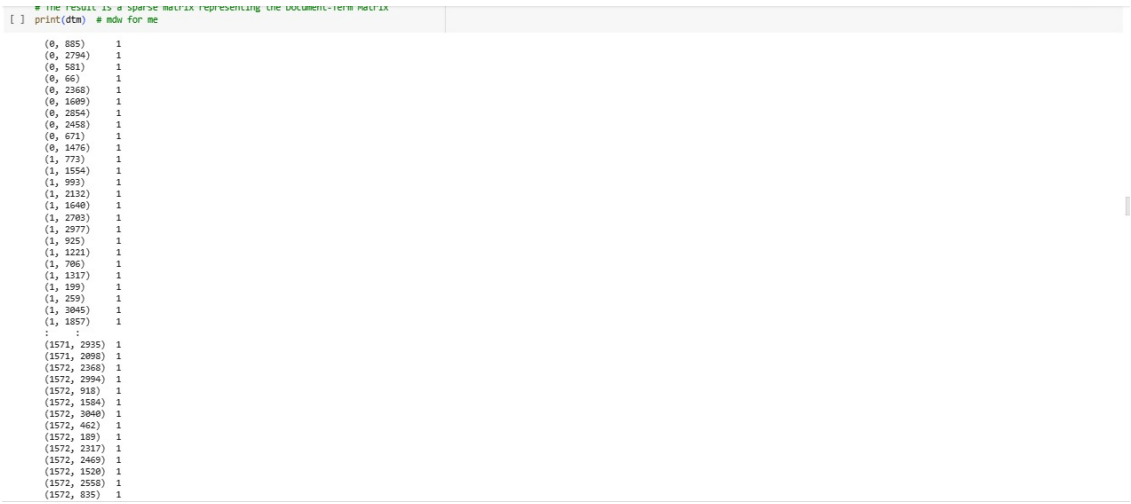


Figure 4.8: Sample creation of one of the dense matrixes from the data set.

5

Evaluation Methods

Evaluation methods refer to the various techniques and processes used to assess, measure, analyze, and judge the effectiveness, efficiency, performance, or value of something. The choice of evaluation method depends on the goals, resources, context, and nature of what is being evaluated. Considering various aspects such as similarity, network structure, and content relevance within communities like this experiment; choosing the appropriate evaluation method depends on the specific goals and characteristics of the dataset or network being analyzed.

In this section, we will describe about our strategies and methodologies which are NMI, traditional modularity and TF-IDF-based modularity because evaluating the performance of algorithms or models in the context of community detection or clustering is crucial to understand how well they organize and identify meaningful groups within a network or dataset. These evaluation measures allow researchers or practitioners to quantitatively assess and compare different algorithms or techniques for community detection in networks, aiding in the selection of the most appropriate approach for a specific context or dataset which is our one of the prime goals of this experiment.

5.1 NMI

In this study, we have used Normalized Mutual Information (NMI) measures to evaluate the results of the topic modeling techniques. NMI, or Normalized Mutual Information, serves as a metric to gauge the similarity between two sets of labels, commonly applied in the assessment of clustering or community detection algorithms. It is employed to evaluate the quality of clustering outcomes by comparing them to a set of known, ground truth labels. The NMI score varies from 0 to 1, with 1 denoting perfect agreement between the sets, and a score of 0 indicating that the two partitions offer no information about each other. This measure provides a normalized assessment, facilitating comparisons across different datasets or clustering algorithms.

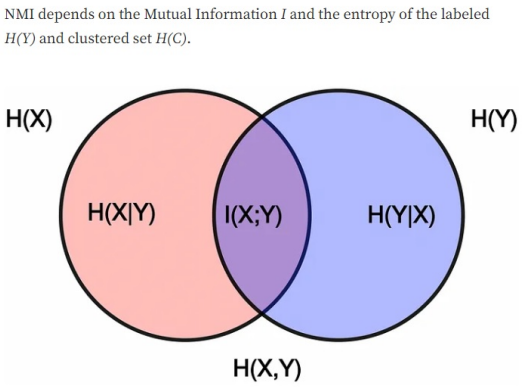


Figure 5.1: Venn diagram portraying the relation between different measures of entropy and Mutual Information . [18]

```
NMI = np.multiply(Pwc,logg(np.transpose(Pcgw)/pc)).sum()/Hc                    #(0 to 1)
print("Normalized Mutual Information:")
print(NMI)

Normalized Mutual Information:
0.4282964943303746
<ipython-input-306-36dd93b7378f>:2: RuntimeWarning: divide by zero encountered in log
  y = np.log(x)
```

Figure 5.2: Sample creation of NMI results from one of the models.

5.2 Modularity- QPcc

In this experiment, we potentially want to relate modularity to the output of a Bertopic-based clustering system. After using Bertopic to cluster the datasets, it can be assessed the quality of the resulting clusters the quality of the resulting clusters can be assessed to quantify the organizational structure of networks or graphs. It assesses how effectively a network can be divided into distinct modules, often referred to as groups, clusters, or communities. The score measures how similar an object is to its cluster compared to other clusters.It ranges from -1 to 1, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. However, the values are typically positive. We measure the traditional modularity as QPcc for each model like NMI.

When documents are grouped into communities based on topic modeling results, traditional modularity can assess the strength of the division of documents into topic-based clusters. Higher modularity values may indicate that documents within communities share more topics or are more thematically coherent within themselves than with documents in other communities.

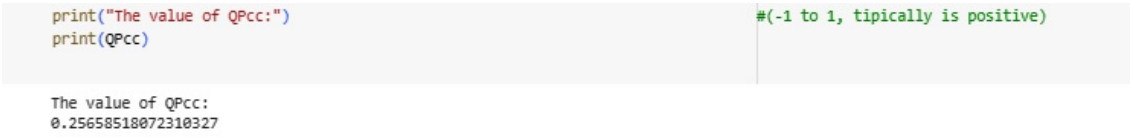


Figure 5.3: Sample creation of traditional modularity (QPcc) results from one of the models.

5.3 TF-IDF-based modularity- QTcc

TF-IDF is a combination of two different words i.e. Term Frequency and Inverse Document Frequency. First, the term "term frequency" will be discussed. TF is used to measure how many times a term is present in a document. Let's suppose, we have a document "T1" containing 5000 words and the word "Alpha" is present in the document exactly 10 times. It is a very well-known fact that the total length of documents can vary from very small to large, so it is a possibility that any term may occur more frequently in large documents in comparison to small documents. So, to rectify this issue, the occurrence of any term in a document is divided by the total terms present in that document, to find the term frequency. So, in this case, the term frequency of the word "Alpha" in the document "T1" will be [14]

TF = 10/5000 = 0.002

Now, inverse document frequency will be discussed. When the term frequency of a document is calculated, it can be observed that the algorithm treats all keywords equally, doesn't matter if it is a stop word like "of", which is incorrect. All keywords have different importance. Let's say, the stop word "of" is present in a document 2000 times but it is of no use or has very less significance, that is exactly what IDF is for. The inverse document frequency assigns lower weight to frequent words and assigns greater weight to the words that are infrequent. For example, we have 10 documents and the term "technology" is present in 5 of those documents, so the inverse document frequency can be calculated as [14]

IDF = log_e (10/5) = 0.3010

So the greater or higher occurrence of a word in documents will give higher term frequency and the lower occurrence of the word in documents will yield higher importance (IDF) for that keyword searched in the particular document. TF-IDF is nothing, but just the multiplication of term frequency (TF) and inverse document frequency (IDF). To calculate the TF-IDF we can do as [14]

TF-IDF = 0.002*0.3010 = 0.000602

In this study, we wanted to gather a score by the representation of documents based on their TF-IDF similarities and then apply modularity-based community detection algorithms to this graph. So it is a combined approach of TF-IDF and modularity. It ranges from -1 to 1; 0 for independence, 1 for complete association and -1 for no association. TF-IDF-based modularity is more sparse and can do better separation from others.

```
print("The value of QTcc:")                                    #(-1 to 1 , tipically positive)
print(QTcc)

The value of QTcc:
0.35402689300901474
```

Figure 5.4: Sample creation of TF-IDF-based modularity (QTcc) results from one of the models.

6

Data Visualization Process

The data visualization process is the representation of data in a graphical or pictorial format to help people understand the patterns, trends, and insights within the data. Bar charts, line charts, pie charts, scatter plots, and more are common types of visualizations used to represent different types of data. It is a crucial aspect of data analysis and communication, as visual representations often make complex information more accessible and comprehensible. When choosing visualization techniques for data analysis and communication, it's essential to consider factors like Data Characteristics (categorical, numerical, time-series, etc.), Message to Convey, Audience etc. By carefully selecting the appropriate visualization methods based on these considerations, effective communication with complex data insights in a clear and understandable manner to the intended audience can happen. In this section, we will narrate in detail what we have chosen for the data visualization process for this experiment.

6.1 Static Graphs

Visualizations make it easier to communicate complex data insights to a non-technical audience. Well-designed visualizations aid decision-making by providing a clear and quick understanding of data. Moreover, a static scatter plot is a type of data visualization that uses Cartesian coordinates to display individual data points on a two-dimensional plane. Each point represents the values of two variables, and the position of the point is determined by the values of these variables.

In this study, we have made 15 static scatter plots for individual representation of the parameters of n_components, n_neighbors, min_samples (n_samples), num_topics and nr_topics in contrast to NMI, QPcc and QTcc with inquiring four parameters. Additionally 36 static scatter plots for individual representation of the parameters of n_components, n_neighbors, min_samples (n_samples), num_topics and nr_topics in contrast to NMI, QPcc and QTcc with inquiring two parameters. A total of 42 static scatter plots were made for this experiment where all the blue dots are clearly visible. For this, we have used the Matplotlib library to create a scatter plot. By reading the stored data from a CSV file what we have achieved from each model, we have extracted the values for the n_component or n_neighbors or min_samples (n_samples), num_topics or nr_topics in the x-axis and NMI or QPcc or QTcc scores in the Y-axis, and then we have plotted these values. With this powerful tool, we want to create a visualizing relationships between two variables.

6.2 3D Graphs

Creating a 3D graph involves visualizing data in three-dimensional space that includes the x-axis, y-axis and z-axis. It can help convey more information compared to 2D graphs. 3D graphs are particularly useful when visualizing relationships involving three variables. Each axis represents one variable, and the position of points in the 3D space provides information about the interplay between these variables.

In this experiment, we have formed three 3D scatter plots for each dataset. Our aim is to interpret all data together from a single viewpoint. Choosing different color enhance the interpretability of the graph. Matplotlib, a popular Python plotting library, provides functionality for creating 3D plots which we have used in our cases. We also import the Axes3D class from the mpl_toolkits.mplot3d module to make an appropriate 3D graph as needed.

On the contrary, Rotating a 3D graph means changing its viewpoint or perspective by adjusting the angles from which it is viewed. In the context of 3D visualization, rotation provides a better understanding of the spatial relationships between data points.

Its orientation can be changed by rotating it along its axes. This dynamic manipulation allows us to explore the data from various angles, revealing hidden patterns or structures that may not be apparent from a single fixed viewpoint. By interactively rotating a 3D graph, we can gain a more nuanced understanding of the three-dimensional distribution of data points, which enhances our ability to interpret and analyze the presented information. For this reason, we have also created a 3D graph with the rotation ability for each dataset. For this purpose, we have used the Matplotlib library and also imported the Axes3D class from the mpl_toolkits.mplot3d module to create a graph.

7

Results Analysis

Result analysis refers to the process of examining, interpreting, and drawing some decision or conclusions from the outcomes or findings obtained from an experiment, study, investigation, or any other process aimed at gathering data. It involves a systematic evaluation of the collected information to derive meaningful insights, understand patterns, identify trends, and make informed decisions or recommendations based on the results. In this section, we will shed light on our graphs which are made by BERTopic and evaluation matrices from the two datasets and make a conclusive decision to keep in mind the performance evaluation of the BERTopic.

Comparing the Two Datasets

Time and speed are related to each other and monitoring speed and performance is our main objective, so first, we made two subsets, then started the experiment's work from the sub-datasets. Then we worked with the entire dataset in this experiment. In the beginning, instead of working with the entire dataset, we worked with subsets to get a proper direction to work with the entire dataset. Among our main objectives, this is also considered as an objective for this experiment. For this purpose, we started working with four parameters (n_components, n_neighbors, min_samples (n_samples), and nr_topics in the Twitter-metoo dataset for the year 2018 which we have portioned from the entire dataset. From now on, we have compared the direction with the Twitter-Covidr dataset.

Twitter-metoo with the 4 parameters:

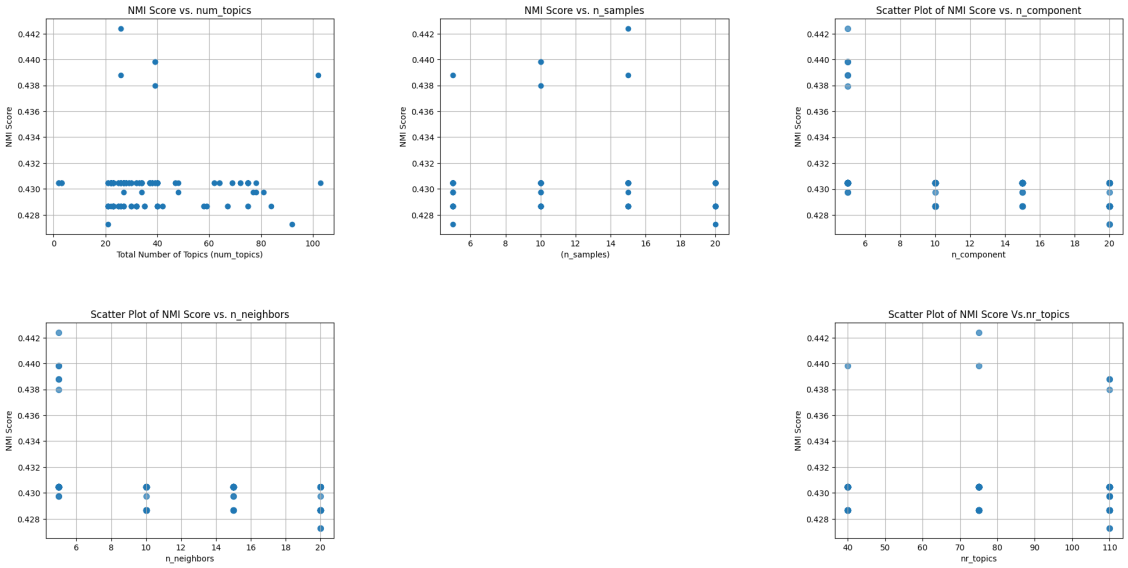


Figure 7.1: Graph representation among the NMI score, number of topics, min_samples (n_samples) and the nr_topics for the Subset of the Twitter-metoo with the four parameters.

The following graphs will provide a clearer understanding of n_components, n_neighbors, min_samples which is denoted as n_samples as a variable in the experiment, and nr_topics's performance of the BERTopic. From this experiment, we get assurance that a lower value of n_components and n_neighbors is better for better results. For this experimental dataset, n_components and n_neighbors gave the best outcome for the value of "5". These 2 parameters belong to the Umap which is a dimensionality reduction technique used for visualizing high-dimensional data in a lower-dimensional space. While working with different combinations of 4 parameters at the first stage, we got a clear idea of the performance of n_components and n_neighbors, but as the concept of min_samples which is denoted as n_samples as a variable in the experiment and nr_topics parameters was not so clear, we again decided that these 2 parameters needed to be observed in a larger scale; where min_samples(n_samples) belongs to HDBSCAN which is a density-based clustering system, extending them to provide a hierarchical clustering approach and nr_topic is an attribute of topic modeling.

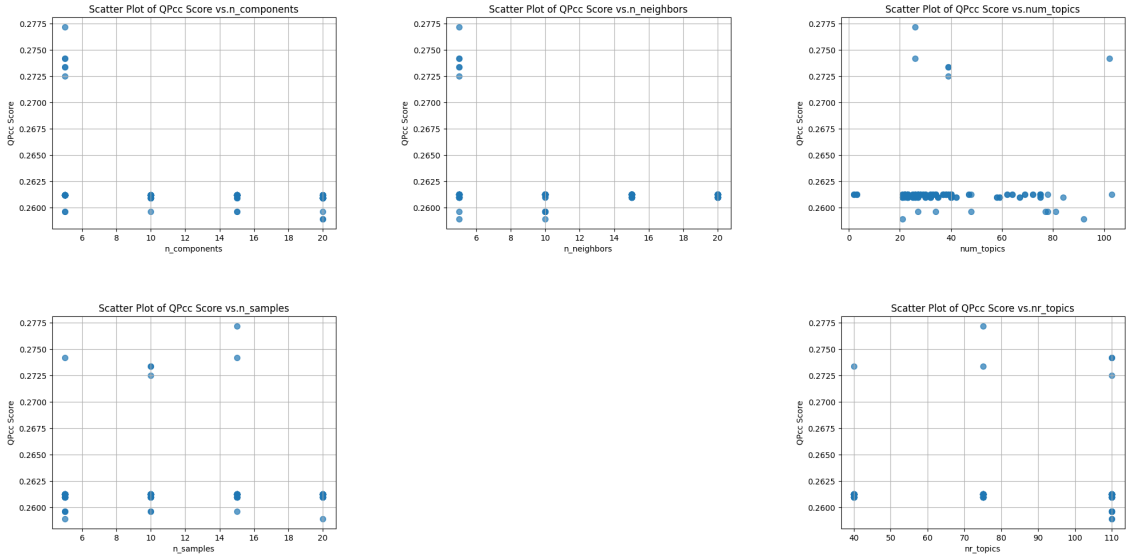


Figure 7.2: Graph representation among the QPcc score, number of topics, min_samples (n_samples) and the nr_topics for the Subset of the Twitter-metoo with the four parameters.

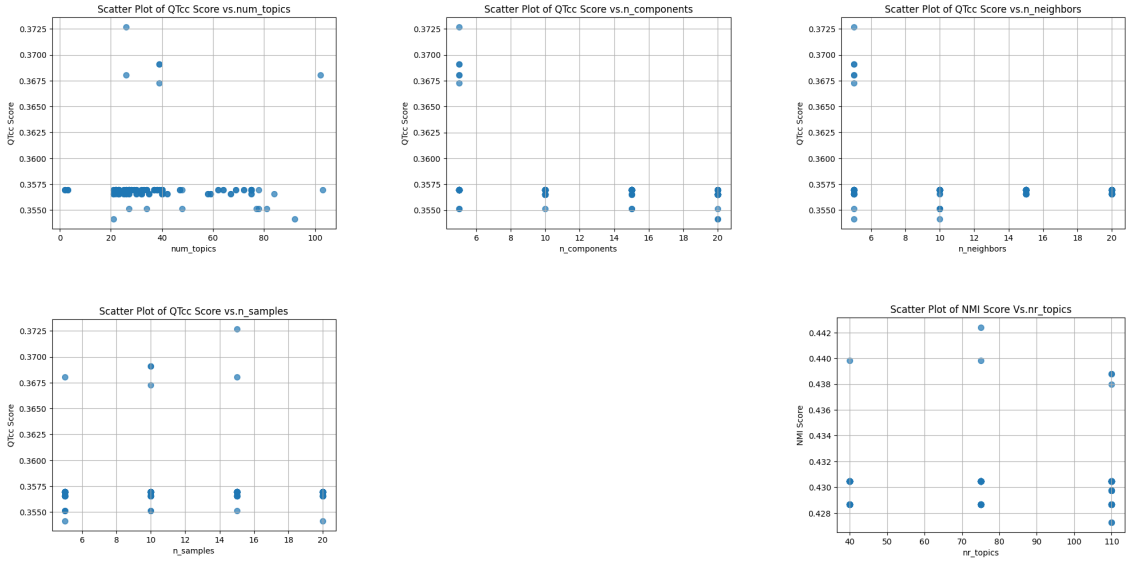


Figure 7.3: Graph representation among the QTcc score, number of topics, min_samples (n_samples) and the nr_topics for the Subset of the Twitter-metoo with the four parameters.

Therefore, considering time and performance, the value of n_components and n_neighbors are fixed with a low value, which is 5 for this experiment; we started the experiment again with different combinations for min_samples (n_samples) and nr_topic. After running this experiment, and observing the results, and all our uncertainties are removed. Considering the results, we realized that the results are not good for the small value of min_samples (n_samples). The best performance area for min_samples (n_samples) is the mid-range value. For example, in the case of this experiment, good results are coming for 7, 9, and 11.

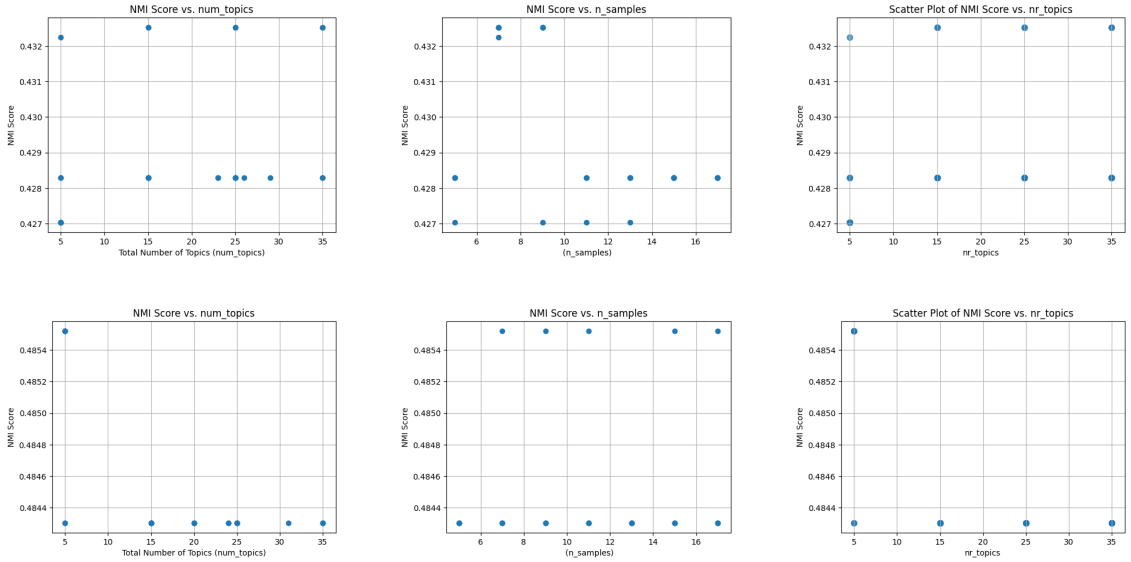


Figure 7.4: Graph representation among the NMI score, number of topics, min_samples (n_samples) and the nr_topics for the Subset of the Twitter-metoo and Twitter-Covid with the 2 parameters.

On the other hand, nr_topics also does not perform well at minimal values. The nr_topics outcome gives pleasant results in mid-range to high-range. It is imperative to mention that the values we set for nr_topic are not always the number of topics we get as output. It mostly depends on the number of min_samples (n_samples). For example, when we worked on the Twitter-metoo dataset in the 2nd stage, we set the value of nr_topic to 35 and the values of other combinations were n_components = 5, n_neighbors = 5, min_samples (n_samples) = 17; In this case, as a result, I got the number of total topics as 23. NMI, QPcc, and QTcc values are available based on the number of total topics. So it can be said that the number of topics is essentially impactful on the performance of BERTopic.

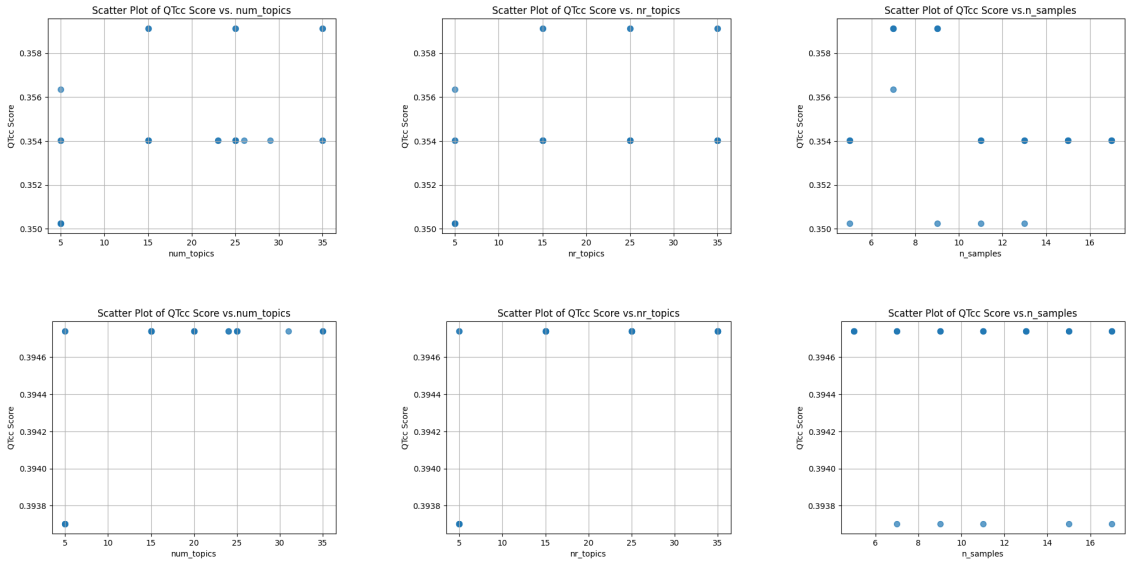


Figure 7.5: Graph representation among the QTcc score, number of topics, min_samples (n_samples) and the nr_topics for the Subset of the Twitter-metoo and Twitter-Covid with the 2 parameters.

For the subset of the Twitter-metoo, NMI values were obtained between 0.42 and 0.43 which is relatively high and close to 1. It suggests a good level of agreement between the true labels and the predicted labels obtained from the clustering algorithm. On the other hand, for the subset of the Twitter-Covid, the NMI values (0.484305 - 0.485520) suggest that the clustering algorithm has performed well in terms of capturing the underlying patterns in the data. The closeness of these values indicates a consistent level of agreement, and they are relatively high, suggesting that the clustering results are similar to the true labels. On the contrary, for the entire set of Twitter-Covid, the NMI values (0.308945 - 0.311604) recommend that values are moderate but not particularly high. The NMI values suggest that the clustering algorithm's performance is decent and can able to address 30-50 percent of the community according to the length of the documents, but may not be as strong as in cases with higher NMI values.

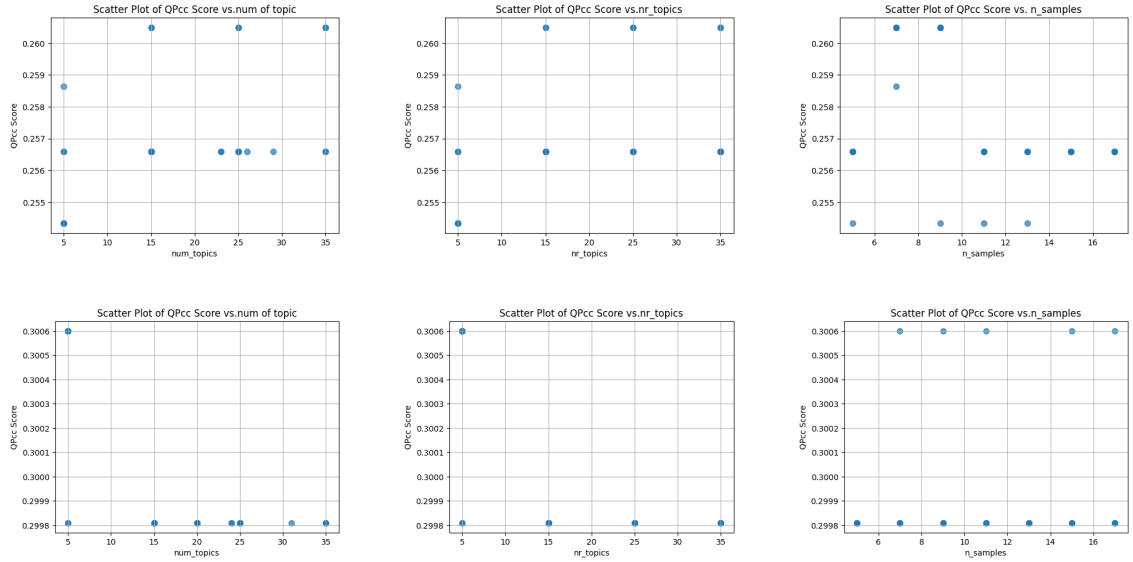


Figure 7.6: Graph representation among the QPcc score, number of topics, min_samples (n_samples) and the nr_topics for the Subset of the Twitter-metoo and Twitter-Covid with the 2 parameters.

In this study, we estimated two types of modularity, in particular, a traditional approach and the remaining one is TF-IDF-based modularity. Traditional community detection mainly focuses on link analysis or the topological structure of the network. Communities identified by those works often incorporate different topics since stronger connections represent the interactions that occur across several different topics, which confuses the meaning of the community and even misleads or mixes the meaning of the community [10]. On the other hand, the TF-IDF-based approach can identify communities from the perspective of both topics and link structure. From this experiment, we wanted to compare the performance of these two modularity approaches within the same framework.

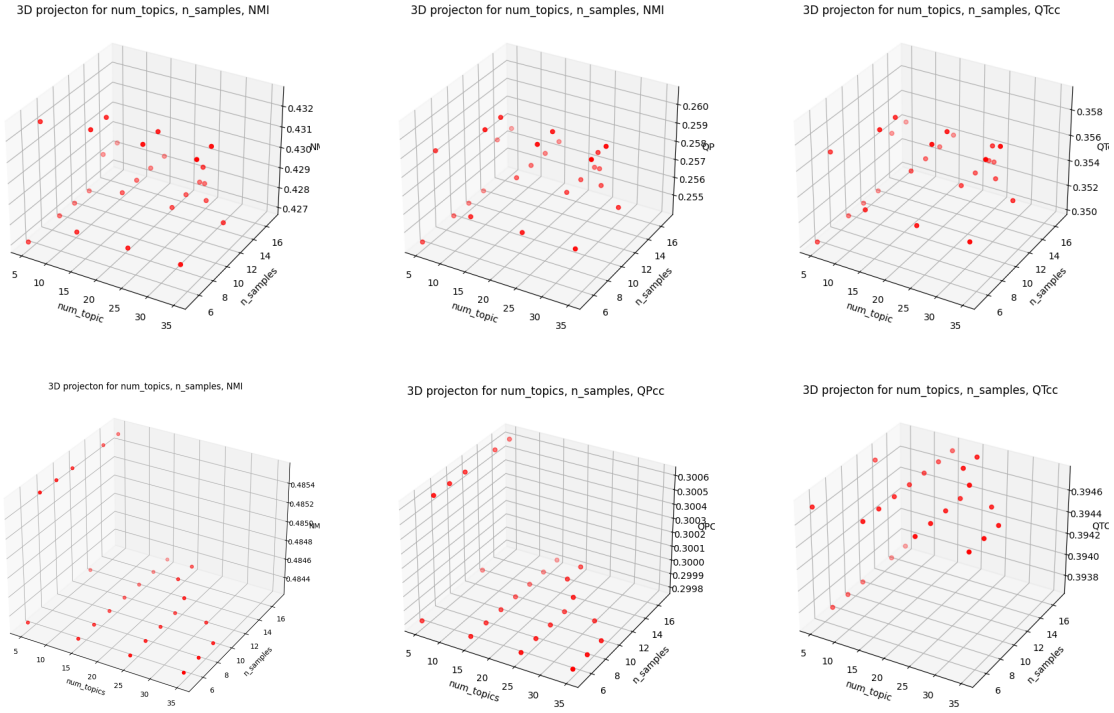


Figure 7.7: 3D projection among the min_samples (n_samples), num_topics and NMI, QPcc, QTcc for the Subset of the Twitter-metoo and Twitter-Covid with the 2 parameters.

Entire Dataset of the Twitter-Covid with the 2 parameters

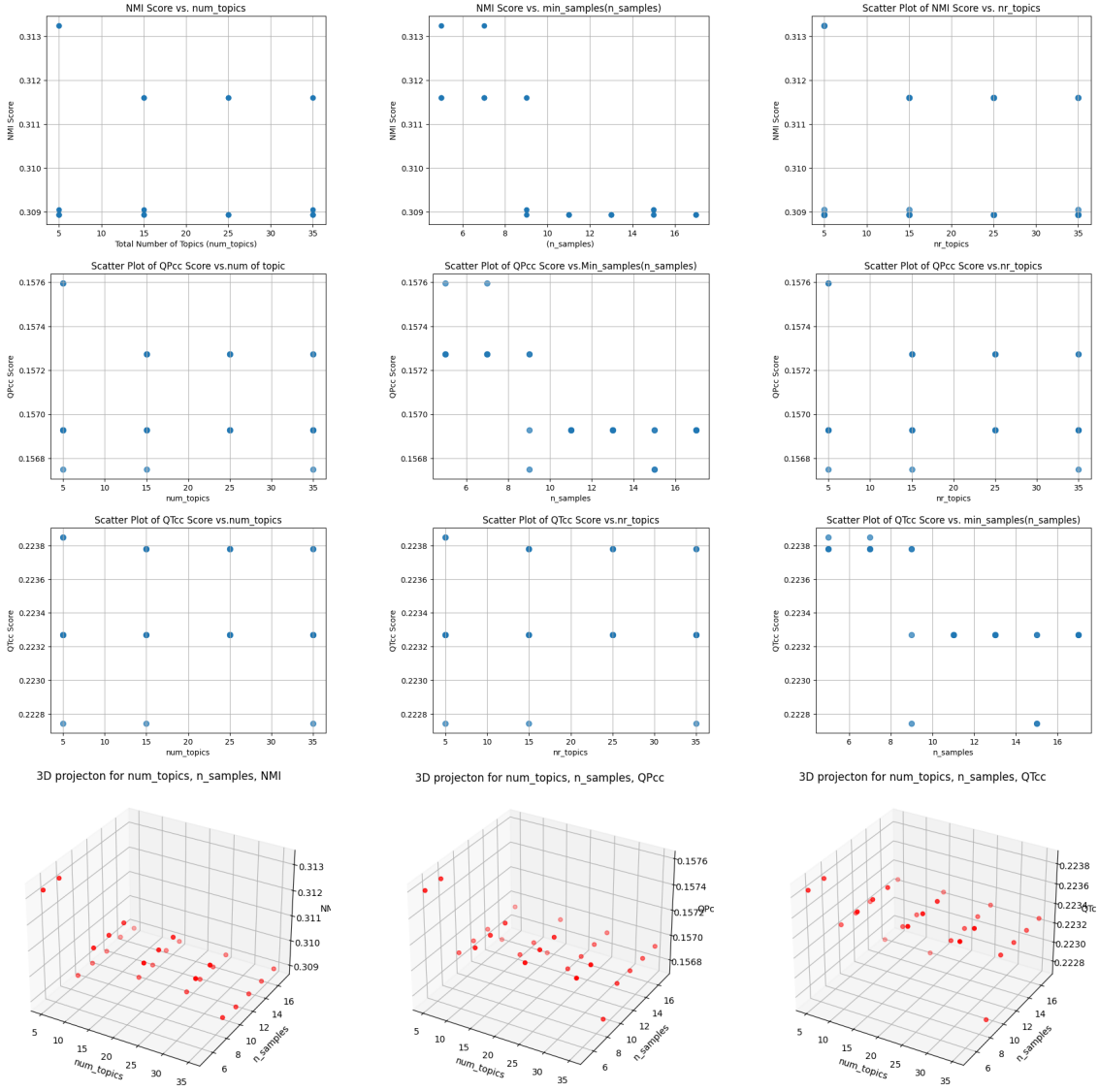


Figure 7.8: Graph Representation for the entire dataset of the Twitter-Covid with the 2 parameters.

For that purpose, we have gathered traditional modularity (QPcc) values between 0.274198 - 0.261223 for the subset of Twitter-metoo while running the combination of four parameters. This value is positive, indicating the presence of community structure in the network and the network is moderately partitioned into communities. The communities are relatively cohesive internally but may have some connections to nodes outside their communities. Higher modularity values are generally desirable as they indicate a more clear separation of nodes into distinct communities. When we were making more specific observations with the same subset i.e. keeping two variables fixed and experimenting with combinations with the other two, we found the value of traditional modularity to be 0.254338 -0.260494. The min_samples (n_samples) value of the mid-range due to the modularity has increased from 0.261223 to 0.274198, indicating an improvement in the clarity and cohesion of communities in the network.

On the other hand, When we were making more specific observations with the Covid subset i.e. keeping two variables fixed and experimenting with combinations with the other two, we found the value of traditional modularity to be 0.299809 -0.300601. The modularity has increased from 0.299809 to 0.300601, indicating that a modularity of 0.300601 indicates a very slight improvement in the community structure compared to the modularity of 0.299809. The communities are slightly more cohesive internally, and there is a slightly better separation between communities. Furthermore, When we were making specific observations with the entire COVID dataset, i.e. keeping two variables fixed and experimenting with the other two variables, we found the value of traditional modularity to be 0.156750 - 0.157597. The min_sample value of the mid-range and due to the modularity has increased from 0.156750 to 0.157597, indicating that the values of the modularity suggest that the network is partitioned into communities. Still, the cohesion within the communities and the separation between communities may not be firm.

On the contrary, we have obtained TF-IDF-based modularity values too. For the subset of Twitter-metoo while running the combination of four parameters we have gathered TF-IDF-based modularity (QTcc) values between 0.356972 - 0.369101. The difference between these two modularity values suggests that 0.369101 has a more well-defined and distinct community structure compared to 0.356972. Higher modularity values generally correspond to better-defined communities within the network. When we were making more specific observations with the same subset i.e. keeping two variables fixed and experimenting with combinations with the other two, we found the value of TF-IDF-based modularity to be 0.350252 - 0.359130. A TF-IDF-based modularity of 0.359130 indicates an improvement in the community structure compared to the TF-IDF-based modularity of 0.350252. The communities are more cohesive internally, and there is a better separation between communities based on the topics. On the other hand, when we were making more specific observations with the covid subset i.e. keeping two variables fixed and experimenting with combinations with the other two, we found the value of TF-IDF-based modularity to be 0.393703- 0.394741. These values are positive, indicating the presence of community structure based on the TF-IDF representation. TF-IDF-based modularity of 0.393703 suggests that the documents are partitioned into communities based on the TF-IDF values, which highlights the importance of terms in the documents.

Moreover, When we were making specific observations with the entire COVID dataset, i.e. keeping 2 variables fixed and experimenting with the other 2 variables, we found the value of TF-IDF-based modularity to be 0.222744 -0.223850. The min_samples (n_samples) value of the mid-range due to the modularity has increased from 0.222744 to 0.223850, indicating that a small improvement is seen in the community structure compared to the TF-IDF-based modularity of 0.222744. A higher modularity value generally indicates better-defined communities within the network based on the TF-IDF features.

Consequently, in comparison to the two types of modularity, it is clearly evident that TF-IDF-based modularity has a better performance than the traditional modularity-based method when the importance of terms in the documents is at least as important as the link. It is certainly apparent that TF-IDF-based modularity contains much less noise than traditional modularity and is more robust in the context of Topic modeling.

8 Conclusion

In summary, this research delved into the time efficiency and performance aspects of BERTopic, particularly concerning the parameter configurations of UMAP and HDBSCAN. The evaluation process was pivotal in assessing the effectiveness of the UMAP and HDBSCAN within the context of BERTopic, focusing on goal-oriented actions and their relevance in analyzing efficacy.

The study meticulously outlined a comprehensive planning methodology encompassing dataset acquisition, preprocessing techniques, the Topic Modeling phase, evaluation metrics, and result analysis. Employing various data visualization techniques, such as graphs, bolstered the depth of the analysis. The outcomes of this research yield valuable insights into the intricate relationship between BERTopic's performance and community elucidation. Our experimentation involved real datasets, demonstrating that our approach effectively discerned more meaningful communities. The quantitative evaluation served as a yardstick for assessing the quality of clusters or topics generated by clustering algorithms, specifically BERTopic.

No model is perfect, and BERTopic is definitely no exception. TF-DF plays a significant role in BERTopic by selecting essential terms for topic modeling, improving topic interpretability, reducing noise, and enhancing the overall performance of the topic modeling process by focusing on terms that provide more discriminative power across the corpus. While TF-IDF has notable strengths, it does have limitations, particularly in capturing the relationships between words, especially synonyms.

As part of future research, we want to investigate some new techniques that can overcome the limitations of TF-IDF aiming to improve its performance and effectiveness.TF-IDF can be combined with other techniques such as TextRank when used as an unsupervised approach to enhance the performance or achieve specific goals in the context of topic modeling with BERTopic. This integration can lead to more accurate, comprehensive, and contextually relevant results in various NLP tasks such as topic modeling, especially in BERTopic.

References

[1] M. Frąckiewicz. A deep dive into chatgpt's approach to topic modeling. [Online]. Available: https://ts2.space/en/a-deep-dive-into-chatgpts-approach-to-topic-modeling/

[2] E. Seaman. (2019) What is a dataset? the definitive guide. [Online]. Available: https://brightdata.com/blog/web-data/what-is-a-dataset

[3] R. Agrawal. (2021) Most conventional techniques for text preprocessing in manuscripts. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/must-known-techniques-for-text-preprocessing-in-nlp/

[4] S. Gharatkar, A. Ingle, T. Naik, and A. Save, "Review preprocessing using data cleaning and stemming technique," in *Proceedings of the Conference Name*, 2017, pp. 1–4.

[5] M. M. Queiroz, "A framework based on twitter and big data analytics to enhance sustainability performance," *Environ. Qual. Manag.*, vol. 28, pp. 95–100, 2018.

[6] S. N. A. N. Ariffin and S. Tiun, "Rule-based text normalization for malay social media texts," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, p. 21, 2020.

[7] A. Krishna, P. Satuluri, S. Sharma, A. Kumar, and P. Goyal, "Compound type identification in sanskrit: what roles do the corpus and grammar play?" in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing*, Osaka, 2016, pp. 1–10.

[8] J. Kasperiuniene, M. Briediene, and V. Zydziunaite, "Automatic content analysis of social media short texts: scoping review of methods and tools," in *Advances in Intelligent Systems and Computing*, A. P. Costa, L. P. Reis, and A. Moreira, Eds. Berlin: Springer, 2020, vol. Computer Supported Qualitative Research, pp. 89–101.

[9] B. Khemani and A. Adgaonkar, "A review on reddit news headlines with nltk tool," in *Proceedings of the International Conference on Innovative Computing Communication (ICICC) 2021*, 2021.

[10] Universal pos tags. [Online]. Available: https://universaldependencies.org/u/pos/

[11] N. T. (Ph.D.). (2023) Topic modeling with bertopic: A cookbook with an end-to-end example (part 1). [Online]. Available: https://medium.com/@nick-tan/topic-modeling-with-bertopic-a-cookbook-with-an-end-to-end-example-part-1-3ef739b8d9f8

[12] M. Grootendorst, "BERTOPIC: Neural topic modeling with class-based TF-IDF procedure," 2022.

[13] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2020.

[14] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," 2018.

[15] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," 2022.

[16] Amy. Hyperparameter tuning for bertopic model in python. [Online]. Available: https://medium.com/grabngoinfo/hyperparameter-tuning-for-bertopic-model-in-python-104445778347

[17] Maarten. Outlier reduction. [Online]. Available: https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html

[18] L. Rita. Normalized mutual information: A measure to evaluate network partitioning. [Online]. Available: https://luisdrita.com/normalized-mutual-information-a10785ba4898

Acknowledgments

To everyone who believed I could when I didn't think it was possible. Especially to my Professor, Tomaso Erseghe, and my two friends who are always with me even in a negative condition.

And most importantly, to the one who raised me up from rock bottom and taught me how to start. Thank you for being with me until the end of a new beginning.