Enhancing Social Media Content Analysis with Advanced Topic Modeling Techniques: A Comparative Study

A.C.Nanayakkara, G.A.D.M.Thennakoon

Abstract*— Topic modeling, a pivotal unsupervised machine learning approach, serves as a valuable tool for uncovering latent themes within vast document repositories. It aids in the organization, comprehension, and simplification of extensive textual data while revealing distinctive underlying themes across a corpus of documents. The intrinsic characteristics of social media content, marked by brevity, text-heavy nature, and a lack of structure, often pose methodological challenges in data collection and analysis. In an effort to bridge the realms of computer science and empirical social sciences, this research aims to assess the effectiveness of three distinct topic modeling methodologies: Bidirectional Encoder Representations from Transformers (BERTopic), Non-negative Matrix Factorization (NMF), and Latent Dirichlet Allocation (LDA). While NMF relies on a matrix factorization paradigm and LDA employs a probabilistic framework, BERT-based techniques, which utilize sentence embeddings for topic generation, represent a contemporary innovation. In this study, BERTopic is evaluated with multiple pre-trained sentence embeddings, and the outcomes are rigorously compared with those derived from LDA and NMF methodologies. The study leverages C_V and U_MASS, two vital coherence measures, to evaluate the efficacy of these topic modeling strategies. The research delves into the analysis of various algorithms, elucidating their strengths and limitations within the context of social sciences, using YouTube comments as a benchmark dataset. Notably, this investigation sheds light on the utility of BERTopic and NMF for evaluating YouTube video content disclosure based on specific attributes, thereby enhancing the analysis process and addressing performance concerns.

***Keywords*— Topic Modeling, Social media analysis, BERTopic, Comparative Study**

I. INTRODUCTION

Topic modelling serves as a dominant method for grouping texts and words that share similar meanings, thus facilitating the extraction of coherent topics from document collections through unsupervised machine-learning techniques. This research explores a diverse array of topic modeling techniques, to offer fresh insights and alternatives for those seeking a deeper comprehension of human relationships. These methodologies encompass a historical trajectory, commencing with the advent of Probabilistic Latent Semantic Analysis (PLSA) in 1999 [1] and the subsequent development of LDA in 2003 [2], which ultimately gained widespread recognition. Another noteworthy technique, NMF, has been instrumental in unsupervised dimension reduction of non-negative matrices [3]. It has found extensive application in uncovering latent topics and revealing underlying connections between textual data [4].

It is important to note that many of these topic modeling methods, even though they operate without the need for predefined labels, require an upfront assessment of the number of categories around which to cluster the data. Nevertheless, in the pursuit of generating meaningful topics, an increasing number of topic modeling techniques, building upon the foundations laid by LDA and NMF, demand substantial effort in terms of hyperparameter tuning and adjustment.

Conversely, in recent years, topic modeling has witnessed significant advancements, thanks to the evolution of Pre-trained Language Models (PLMs). These sophisticated machine learning models have undergone extensive training on vast corpora of text data and can be fine-tuned to address specific Natural Language Processing (NLP) tasks. This transformative shift has been propelled by the limitations of traditional statistical topic modeling tools, leading to the emergence of innovative technologies, such as BERT[5].

In order to provide a more in-depth understanding, we embarked on an exploration of the BERTopic method [6], which has represented a remarkable leap forward in the domain of topic modeling. BERTopic leveraged the capabilities of three essential components to accomplish its objectives: Class-Based Term Frequency-Inverse Document Frequency (c-TF-IDF) weights, BERT Embeddings, and Dimensionality Reduction through UMAP. These methodologies are elaborated upon in the research methodology section for clarity.

It is noteworthy to highlight that the preliminary assessments of the BERTopic approach have yielded highly encouraging results, validating its potential to significantly enhance topic modeling practices [7].consequently, throughout this study, the BERTopic method is subjected to testing across various Pre-trained Language Models (PLMs), and the outcomes are juxtaposed with those obtained from established techniques such as LDA and NMF. The subsequent sections of this paper are structured as follows:

The Literature Review section provides an overview of prior research pertaining to topic modeling for social media disclosure. The Research Methodology section delineates the techniques employed and the dataset utilized in this study. In the Results and Discussion section, we analyze and discuss the outcomes. Finally, the Conclusion section

Correspondence: Amila Chethana Nanayakkara (e-mail: chethana@ccs.sab.ac.lk) Received: 26-02-2024 Revised: 07-04-2024 Accepted: 10-04-2024

Amila Chethana Nanayakkara and Danuka Mahesh Thennakoon are from Sabaragamuwa University of Sri Lanka (chethana@ccs.sab.ac.lk, dhanuka@adm.sab.ac.lk).

DOI: https://doi.org/10.4038/icter.v17i1.7276

encapsulates the paper with its concluding remarks. The importance of this study is elucidated in the following manner:

1. BERTopic-based embedding models versus LDA/NMF: A comparison: This involves evaluating the effectiveness of BERTopic, a topic modeling method using BERT-based embeddings, in comparison to traditional techniques like Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) for topic discovery and analysis. It aims to determine which method performs better in identifying and clustering topics within a given dataset.

2. Analysing the differences between the usage of several sentence transformers: This task entails examining the distinctions and nuances between the utilization of various sentence transformers, namely "all-MiniLM-L6-v2," "all-mpnet-base-v2," "all-distilroberta-v1," and "roberta-base-nli-stsb-mean-tokens." These are different models used for encoding sentences and text data. The analysis seeks to understand how these models differ in their ability to represent and understand text, which is crucial for various NLP tasks.

3. Identifying the best topic model for social media disclose analysis: This refers to the process of finding the most suitable topic modeling approach for analyzing disclosures on social media platforms. It involves comparing and evaluating different topic modeling techniques to determine which one is most effective in uncovering and categorizing the themes or topics within social media content related to disclosures.

II. LITERATURE REVIEW

Text mining approaches play a crucial role in the extraction of key features from vast volumes of textual data [8]. Among these methods, topic modeling [9] stands out as the most frequently employed technique. Topic modeling is an algorithmic approach commonly used in machine learning and natural language processing to reveal hidden thematic patterns within a collection of texts [10]. Its applications extend to various domains, including the social sciences, where it has been instrumental in uncovering consumers' latent preferences [11], identifying semantic structures on platforms like Instagram [7], and enhancing recommendation systems [12].

A. Latent Dirichlet Allocation (LDA)

For discrete datasets such as text corpora, the prevailing topic modeling method LDA, relies on a generative probabilistic model [13]. To aid in the comprehension of the information derived from a trained LDA model, the data is often transformed into an intertopic distance map using tools like pyLDAvis [14].

However, it is important to note that LDA has had criticism, especially when applied to the analysis of social media data, despite its strong track record in social science research [7] [15]. Some researchers [18] highlight that LDA may not be well-suited for "dirty" and limited datasets, which lack the necessary features for robust statistical learning. Consequently, this has prompted a shift in focus among researchers toward cutting-edge algorithms that demonstrate superior performance, with a particular emphasis on analyzing the concise text data commonly found in social media [7].

B. Non-negative Matrix Factorization (NMF)

NMF stands out as a non-probabilistic technique that diverges from LDA by employing matrix factorization as its core methodology. NMF falls within the domain of linear-algebraic algorithms [16]. When applied to data that has been transformed using TF-IDF, NMF dissects a matrix into two lower-ranked matrices [17]. In essence, TF-IDF serves as a metric for assessing the significance of a word within a collection of documents.

However, it is crucial to recognize that NMF necessitates preliminary data preprocessing to function effectively. This involves various preparatory steps such as converting text to lowercase, removing stop words, lemmatization, stemming, and the elimination of punctuation and numeric characters, among other tasks. These preprocessing steps are essential to ensure that the data is in the appropriate format for NMF analysis.

C. BERTopic

Researchers have recently been exploring innovative techniques like BERTopic, Corex, Top2Vec, and NMF [18][17][19] to enhance text analysis and topic modeling. Among these, BERTopic stands out as a noteworthy approach. BERTopic draws inspiration from Top2Vec's methodologies, and as a result, the two share similar algorithmic frameworks.

BERTopic is particularly intriguing due to its multilingual capabilities, supporting over fifty languages. It employs BERT, which is a powerful language model, as the foundation for document embedding extraction. This means that it represents text in a way that captures its nuanced meanings. BERTopic utilizes HDBSCAN for document clustering, a method that identifies dense regions in data, and UMAP for dimension reduction, which simplifies the data while preserving its essential structure.

What makes BERTopic even more appealing is its provision of an interactive intertopic distance map. This tool allows users to visualize the relationships between topics, making it easier to explore and comprehend the landscape of topics within a text corpus. BERTopic, with its multilingual capabilities, robust embeddings, and visualization aids, is a promising and versatile tool in the field of text analysis and topic modeling.

III. RESEARCH METHODOLOGY

A. Topic Model Implementation

Recognizing the transformative potential of social media in disaster intervention [20], our study strategically leverages a prominent YouTube video comment thread, one that resonates with the "black lives matter" movement sparked by George Floyd's tragic demise. This carefully

selected thread serves as a litmus test for evaluating the efficacy of three distinct topic modeling techniques: LDA, NMF, and the innovative BERTopic.

LDA and NMF were initially employed as benchmarks, serving as established reference points in our investigation. Subsequently, the formidable BERTopic framework, augmented by several word embedding models, entered the scene. The computational heavy lifting was executed through the Gensim's LDA Multicore implementation and Sklearn's NMF implementation in Python. To facilitate this intricate exploration, our experiments unfolded within the dynamic realm of a Jupyter notebook, nested securely in an Anaconda environment. The meticulously planned steps in this research proceeded as follows.

Fig. 1, a visual guide to our methodology, delineated two distinctive paths to unearth topics from the data. The first route adhered to traditional Topic Modeling methodologies, typified by LDA and NMF. This approach entailed the indispensable tasks of removing stop words and lemmatization, essential for text normalization. Conversely, the second method used a more advanced BERTopic architecture. In this pursuit, the semantic nuances of words were artfully preserved, while finding topics that were connected to each other and made sense in the context of the information we were studying.

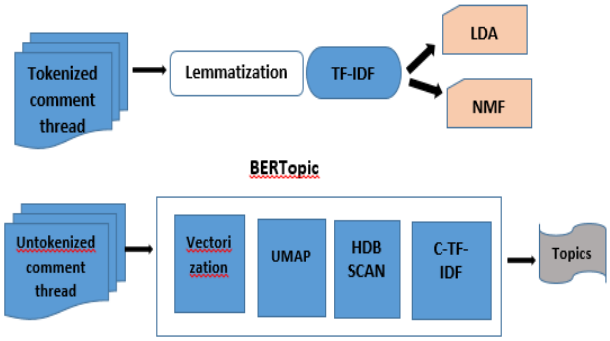

Fig. 1 Proposed Model Architecture

Three fundamental components of BERTopic architecture:

1) *Class-Based Term Frequency-Inverse Document Frequency (c-TF-IDF):* The technique assesses the importance of words within documents by taking into account both their frequency within a specific document (Term Frequency or TF) and their scarcity across the entire corpus (Inverse Document Frequency or IDF). Essentially, it distinguishes words that are not only frequently used in a document but are also distinctive to that document when compared to the corpus as a whole.

2) BERT Embeddings: This represent words in a multi-dimensional vector space, capturing their nuanced contextual meanings. These embeddings offer a more sophisticated representation of words compared to traditional word representations, which often treat words as isolated entities.

3) *Dimensionality Reduction with UMAP:* Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction method that helps visualize and represent high-dimensional data in a more manageable form. In the context of BERTopic, UMAP contributes to condensing the complex embeddings, making them more amenable to clustering and topic analysis.

For the first step of producing document embedding's, sentence transformers were employed to extract document embeddings. BERTopic first extracts document-level embedding's using the "all-MiniLM-L6-v2" model as its default sentence transformer. Subsequently 'all-mpnet-base-v2', 'all-distilroberta-v1', and 'roberta-base-nli-stsb-mean-tokens' sentence transformers have been experimented with. In the subsequent stage, the dimensions of the document embedding's are reduced using the UMAP technique, and the semantically related documents are clustered using the HDBSCAN algorithm. In the third stage, significant terms for each cluster were extracted using c-TF-IDF.

B. *Data gathering and pre-processing*

The study employed a YouTube video comments dataset derived from a highly popular video titled "Violent George Floyd protests at CNN Centre unfold live on TV." As of July 10, 2021, this video had garnered 9,25K views, 98K likes, 16K dislikes, and a remarkable total of 70,272 comments. This selection was deliberate, driven by the video's exceptional comment count, making it a prime candidate for analysis.

The initial step involved downloading the dataset in the form of a .CSV file, achieved through the utilization of the YouTube API V3. These comments, being the product of human interaction, encompassed a diverse array of textual elements, including punctuation marks, HTML tags, numerical figures, and special characters. In preparation for subsequent analysis, comprehensive data preprocessing was executed. This encompassed various operations, such as converting text to lowercase, eliminating URLs and email addresses, removing common stop words, stripping away punctuation, and discarding emojis.

It is noteworthy to highlight that, up to this stage, the BERTopic approach exclusively worked with the original sentences. This approach relies on embedding techniques, and for transformer models like BERT, preserving the original form of the text is paramount to retain the contextual richness.

Moreover, the text data underwent additional preprocessing for LDA and NMF analysis using Python's natural language processing (NLP) tools. This included stemming and lemmatization processes. Furthermore, the text's Term Frequency-Inverse Document Frequency (TF-IDF) weight was calculated to facilitate keyword-based information retrieval.

In sum, the study's initial phases focused on gathering and refining the data to prepare it for advanced textual analysis, ensuring the dataset's readiness for subsequent topic modeling and analysis techniques.

C. *Evaluation matrices*

Evaluating the effectiveness of unsupervised models presented a challenge due to their inherent lack of structure and the absence of established evaluation methods. In our

study, one key criterion for assessing a topic model was to minimize the internal distance between topics while maximizing the inter-distance between clusters, indicating model efficiency.

To quantitatively compare the performance of the topic models, we employed two widely-used evaluation metrics, namely "C V" and "U MASS." The "C V" metric was utilized to measure topic similarity, with success reflected in achieving the maximum value, indicating an effective topic model. Conversely, the "U Mass" measure assessed the intra-word distance within topics, with optimal results achieved when this measure reached its minimum value.

The Genism library played a crucial role in our study, facilitating the implementation of these coherence measures. It allowed for a systematic and objective evaluation of the topic models considered in our research.

D. *Baseline*

The study involved an extensive evaluation of results derived from four distinct BERTopic embedding models, which were compared against the outcomes obtained through two of the most prevalent and established topic modeling techniques, namely LDA (Latent Dirichlet Allocation) and NMF (Non-negative Matrix Factorization).

The experimental process was initiated with the stipulation of a minimum of two topics for analysis. Subsequently, the number of topics was incrementally increased. It is important to note that both LDA and NMF require the prior specification of the number of topics, necessitating a stepwise exploration of topic counts during the experimentation. This systematic approach allowed for a comprehensive comparative analysis of the performance of BERTopic against the conventional methodologies.

IV. RESULTS AND DISCUSSION

In summary, topic models can significantly augment the domain of social science research by incorporating statistical analysis techniques. However, it is crucial to acknowledge that each algorithm within the topic modeling framework is distinct and founded on a unique set of underlying hypotheses. In alignment with the distinct nature of these algorithms, the subsequent sections will be structured into two segments for comparative analysis:

I. Comparison of LDA and NMF.
II. Evaluation of the impact of different word embeddings within the context of BERTopic.

These sections will serve to delineate the differences and relative merits of these methodologies, allowing for a comprehensive examination of their effectiveness in the context of social science research.

A. *Comparison of LDA and NMF*

In this section, a comparative analysis has been conducted to assess the performance of two statistical topic models, LDA and NMF. This evaluation has been approached from two distinct perspectives to provide a comprehensive understanding of the relative strengths and limitations of these models.

1) *Comparison of the topical contents of the LDA and NMF models:* The findings pertaining to the top 10 topics identified by the LDA and NMF topic models have been comprehensively summarized in Appendix 01, Figure 2. To ascertain the most significant terms contributing to each topic, TF-IDF weights were employed.

It is worth noting that when examining the list of topics produced by LDA, certain words, such as "people," "police," and "protest," tend to reappear across multiple topics, resulting in a degree of redundancy. In contrast, the NMF-generated topic list offers a more diverse array of topics, alleviating the issue of repeated terms.

Upon careful evaluation, this study arrives at the conclusion that the results obtained from the NMF model exhibit superior alignment with human judgment. Consequently, NMF outperforms LDA, providing more coherent and varied topics, enhancing the quality and interpretability of the outcomes.

2) *Evaluation metrics comparison of LDA and NMF models :* In accordance with the evaluation metrics outlined in the research methodology section, the coherence of the proposed topic model was systematically assessed using the C_V and U_MASS coherence measures.

As depicted in Figure 3, a quick comparative analysis of the two traditional topic models reveals that, for the majority of topic numbers, NMF demonstrates superior performance. This is evident in the NMF model's attainment of the highest C_V score and the lowest U_MASS score, underlining its commendable performance with the given dataset.

Taking into account these coherence scores (C_V and U_MASS), the empirical evidence leads to the conclusion that the NMF topic model outperforms the alternative models in terms of coherently capturing the underlying topics within the dataset.

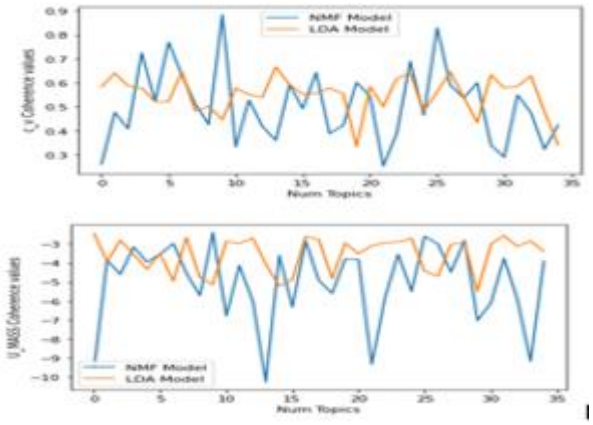

Fig. 3 Coherence comparison of LDA and NMF Models

B. *Comparison of different word embeddings on BERT*

Within this section, a comprehensive comparative analysis has been conducted to assess the performance of four distinct word embeddings applied to the BERTopic model. This evaluation has been approached from three distinct perspectives to provide a thorough understanding of

the relative strengths and limitations of these embeddings within the BERTopic framework.

1) *Generated topics comparison of different word embedding's on BERTopic :* As indicated in Appendix 02, Figure 4, the topic lists denoted as (a, b, c, and d) each incorporate a topic labeled with -1. This label is used to represent all the outliers in the dataset that do not have an assigned topic. The inclusion of documents within these outlier topics could lead to a decrease in overall model performance. Therefore, the topic labeled as -1 has been omitted from consideration.

Notably, when examining topic list d, it becomes evident that it contains the largest number of outliers. This suggests that the topics generated by the d algorithm (roberta-base-nli-stsb-mean-tokens) possess a higher degree of coherence and relevance compared to the other three algorithms.

Furthermore, it is discernible that even in terms of topic diversity, topic list d stands out. It encompasses a broader array of topics compared to the three other algorithms, including subjects such as "George Floyd," "Trump America," "King Luther," and "peaceful protest." This diversity underscores the comprehensive coverage and semantic richness achieved by the d algorithm in generating meaningful topics.

2) *Inter topic distance map comparison of different word embedding's on BERTopic:* As detailed in Appendix 03, Figure 5, the inter-topic distance map is visually represented with circles, each circle symbolizing a specific topic. The size of these circles corresponds to the frequency of the topic's appearance across all documents. In the provided inter-topic distance maps (labeled as a, b, c, and d), the most prominent circle, marked in red as Topic 0, is indicative of the topic with the highest c-TF-IDF weight. The circle's size is directly proportionate to the assigned c-TF-IDF weights for each topic.

It is noteworthy that Topic 0, Topic 1, and Topic 2 carry the highest c-TF-IDF weights, and consequently, they are represented by larger circles. In contrast, Topic 8 and Topic 9 have relatively lower c-TF-IDF weights, and as a result, their corresponding circles are smaller in size.

These inter-topic distance maps reveal the existence of two primary clusters within the 10 identified topics. A closer examination of their characteristics, particularly in the d map, unveils a noteworthy pattern. The d map exhibits the most substantial inter-distance between the two clusters, while simultaneously displaying the lowest internal distances between the topics. In essence, many of the topics are closely intertwined, with significant overlap.

Considering these observations and the overall assessment, it can be conclusively stated that the inter-topic distance map generated by the 'roberta-base-nli-stsb-mean-tokens' algorithm outperforms the maps produced by the other three algorithms. This is indicative of its capacity to offer a more distinct and coherent visualization of the underlying topic structures.

3) *Coherence measures comparison of different word embedding's on BERTopic :* Parallel to the aforementioned experiment involving the conventional topic models LDA and NMF, we employed the C_V and U_MASS evaluation metrics to systematically gauge the efficacy of various word embeddings. Our objective was to ascertain which word embedding method yielded the most coherent and interpretable results.

TABLE I. Coherence values of Bert embedding's

	Metrics/Algorithm	U_MASS	C_V
1	all-MiniLM-L6-v2 (default for bertopic)	-12.0905	0.3749
2	all-mpnet-base-v2	-12.8738	0.3891
3	all-distilroberta-v1	-13.3205	0.4078
4	roberta-base-nli-stsb-mean-tokens (Sentence Transformer)	-14.3149	0.5056

The findings, as presented in TABLE I, unequivocally reveal that among the four approaches, the 'Roberta-base-nli-stsb-mean-tokens' algorithm stands out as the top performer. It achieved the highest C_V score, registering an impressive value of '0.5056,' while simultaneously obtaining the lowest U_MASS score, denoted by '-14.3149.' It is important to note that 'Roberta-base-nli-stsb-mean-tokens' is specifically tailored as an application of sentence transformer, meticulously designed for a designated social media dataset. This customized approach appears to have greatly contributed to its superior performance, highlighting the potential of custom data embeddings for specialized datasets.

In light of these results and the meticulous evaluation of coherence scores (C_V and U_MASS), it is judicious to conclude that 'Roberta-base-nli-stsb-mean-tokens' word embedding, when applied within the BERTopic model, excels in effectively capturing the inherent structure and semantics of the dataset under consideration.

V. CONCLUSION

This study conducted a comprehensive evaluation of BERTopic, utilizing various pre-trained sentence transformers as embeddings, and compared the outcomes with well-established topic modeling methodologies like LDA and NMF in the realm of social media content analysis. The Results and Discussion section unambiguously demonstrates that, in the realm of traditional topic modeling, the NMF model surpasses LDA, as indicated by higher coherence values and greater topic diversity.

Moreover, the Results and Discussion reveal that BERTopic, through its distinctive embeddings, delivers topics that are not only more interpretable but also exhibit higher coherence scores than conventional natural language processing topic modeling approaches. Remarkably, the transformer-based "Roberta-base-nli-stsb-mean-tokens" algorithm within the BERTopic model excelled, primarily due to its training on social media data. It offered more distinct topics and enhanced coherence ratings across the entire spectrum of generated BERT topics. In sum, BERTopic emerged as the superior choice when compared to NMF and LDA, underscoring its overall efficacy.

However, while BERTopic has shown promising initial results and remarkable flexibility in accommodating various topic sizes, the task of reliably assessing the quality of topics it generates remains a challenging frontier, necessitating further exploration in future research endeavors. Additionally, the performance of the model can be further enhanced through the integration of domain-

specific pre-trained word embeddings, representing an exciting avenue for future studies in this domain.

REFERENCES

[1] T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Mach. Learn.*, vol. 42, no. 1–2, pp. 177–196, 2001, doi: 10.1023/A:1007617005950.

[2] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," *Art Sci. Anal. Softw. Data*, vol. 3, pp. 139–159, 2015, doi: 10.1016/B978-0-12-411519-4.00006-9.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999, doi: 10.1038/44565.

[4] S. Arora, R. Ge, and A. Moitra, "Learning topic models - Going beyond SVD," *Proc. - Annu. IEEE Symp. Found. Comput. Sci. FOCS*, pp. 1–10, 2012, doi: 10.1109/FOCS.2012.49.

[5] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, no. May, pp. 1–16, 2022, doi: 10.3389/fsoc.2022.886498.

[6] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia CIRP*, vol. 189, pp. 191–194, 2021, doi: 10.1016/j.procs.2021.05.096.

[7] R. Egger and J. Yu, "Identifying hidden semantic structures in Instagram data: a topic modelling comparison," *Tour. Rev.*, vol. 77, no. 4, pp. 1234–1246, 2022, doi: 10.1108/TR-05-2021-0244.

[8] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A review of text corpus-based tourism big data mining," *Appl. Sci.*, vol. 9, no. 16, 2019, doi: 10.3390/app9163300.

[9] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," *SOMA 2010 - Proc. 1st Work. Soc. Media Anal.*, pp. 80–88, 2010, doi: 10.1145/1964858.1964870.

[10] Y. Guo, S. J. Barnes, and Q. Jia, *Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation*, vol. 59. 2017.

[11] H. Q. Vu, G. Li, and R. Law, "Discovering implicit activity preferences in travel itineraries by topic modeling," *Tour. Manag.*, vol. 75, no. June, pp. 435–446, 2019, doi: 10.1016/j.tourman.2019.06.011.

[12] W. Shafqat and Y. C. Byun, "A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis," *Sustain.*, vol. 12, no. 1, 2020, doi: 10.3390/SU12010320.

[13] S. J. Blair, Y. Bi, and M. D. Mulvenna, "Aggregated topic models for increasing social media topic coherence," *Appl. Intell.*, vol. 50, no. 1, pp. 138–156, 2020, doi: 10.1007/s10489-019-01438-z.

[14] T. Islam, "Yoga-Veganism: Correlation Mining of Twitter Health Data," 2019.

[15] "Psychology and Marketing - 2021 - S nchez-Franco - Do travelers reviews depend on the destination An analysis in coastal.pdf." .

[16] S. Si, J. Wang, R. Zhang, Q. Su, and J. Xiao, "Federated Non-negative Matrix Factorization for Short Texts Topic Modeling with Mutual Information," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2022-July, 2022, doi: 10.1109/IJCNN55064.2022.9892602.

[17] A. Obadimu, E. L. Mead, N. Agarwal, and E. Mead, "Identifying Latent Toxic Features on YouTube Using Non-negative Matrix Factorization," *Ninth Int. Conf. Soc. Media Technol. Commun. Informatics*, no. October, 2019.

[18] P. J. Sheldon and D. R. Fesenmaier, *Tourism on the Verge Series editors*, no. February. 2022.

[19] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022.

[20] F. Femenia-Serra, U. Gretzel, and A. Alzua-Sorzabal, "Instagram travel influencers in #quarantine: Communicative practices and roles during COVID-19," *Tour. Manag.*, vol. 89, no. April, 2022, doi: 10.1016/j.tourman.2021.104454.

Appendix 01, Fig. 2

Topics identified by LDA

Topic No	Major words in the cluster (Top 10 words)
0.	people, knee, protest, police, america, neck, george, floyd, world, protesters
1.	people, protest, go, america, looting, cops, get, way, country, us
2.	people, get, one, need, us, cops. black, time, american, think
3.	Looting, george, floyd, people, protesting, stop, violence, peaceful, us, get
4.	People, black, one, cops, us, going, thats, get, violence, floyd
5.	People, police, know, get, protests, us, one, lol, time, protest
6.	People, protest, police, black, looting, george, us, cant, protesters, white
7.	America, people, floyd, need, live, get, george, one, protesters, media
8.	Police, people, peaceful, see, white, protest, black, right, one, man
9.	People, police, us, get, stop, need, make, white, world, good,

Topics identified by NMF

Topic No	Major words in the cluster (Top 10 words)
0.	People,stop,looting, protesting, dont, stupid, need, innocent, stealing, stores
1.	America, great, make, country, world, trump, love, continent, live, states
2.	George, floyd, death, justice, looters, rip, care, soros, looting, murder
3.	Protest, peaceful, looting, protests, peacefully, protesters, looters, tear, gas,
4.	Knee, neck, officer, cop, cops, guys, putting, guy, white, mans
5.	Lol, look, peacefully, really, brush, say, american, want, know, dont
6.	Like,look, purge, dont, going, sounds, act, movie, riot, acting
7.	Police, need, brutality, protesters, officers, man, officer, killed, racist
8.	Just, want, dont, excuse, im, cops, looting, bad, loot, shoot
9.	Black, lives, matter, white, man, guy, person, racist, power, dont

Fig. 2 Topics identified by two statistical topic models

Appendix 02, Fig. 4

a) 'all-MiniLM-L6-v2' embedding

Topic	Count	Name
-1	4796	-1_people_not_just_will
0	419	0_floyd_george_death_not
1	269	1_peaceful_protest_tear_peacefully
2	244	2_covid_virus_corona_coronavirus
3	203	3_knee_neck_officer_cop
4	166	4_hong_kong_hk_beautiful
5	146	5_looting_protesting_not_protest
6	133	6_cops_police_guard_national
7	105	7_looters_shoot_protesters_looter
8	99	8_stealing_steal_stores_people
9	97	9_riots_history_riot_rioting
10	90	10_america_continent_great_falling

b) '' all-distilroberta-v1'' embedding

Topic	Count	Name
-1	5276	-1_people_not_just_police
0	225	0_dont_just_cant_know
1	215	1_looting_looters_protesters_people
2	180	2_lol_que_de_la
3	173	3_america_purge_great_world
4	166	4_beautiful_sight_quot_pelosi
5	110	5_knee_neck_officer_off
6	93	6_gas_bullets_rubber_tear
7	89	7_american_americans_america_african
8	88	8_protests_riots_peaceful_rioting
9	77	9_george_floyd_rip_soros
10	75	10_hong_kong_hongkong_us

c) 'all-mpnet-base-v2' embedding

Topic	Count	Name
-1	4706	-1_people_not_just_will
0	405	0_looting_protest_protesting_not
1	339	1_floyd_george_death_floyds
2	288	2_covid_virus_corona_que
3	213	3_police_cops_cop_officers
4	178	4_knee_neck_officer_off
5	153	5_black_white_people_blacks
6	148	6_lol_omg_smith_ikr
7	122	7_kong_hong_hk_us
8	111	8_purge_joker_movie_true
9	104	9_media_news_nbc_cnn

d) 'roberta-base-nli-stsb-mean-tokens'

Topic	Count	Name
-1	5585	-1_people_not_just_police
0	208	0_knee_neck_officer_knee neck
1	193	1_george_floyd_george floyd_death
2	131	2_america_america great_trump_great
3	112	3_purge_movie_joker_true
4	108	4_bar_king_luther king_luther
5	103	5_peaceful_protest_protests_peaceful protest
6	96	6_people_trump_like_america
7	79	7_beautiful_sight_beautiful sight_nancy
8	76	8_crossbow_bow_not crossbow_lmao
9	76	9_protest_riots_riot_protesting

Fig. 4 Topics identified by two by Bertopic embedding's

Appendix 03, Fig. 5

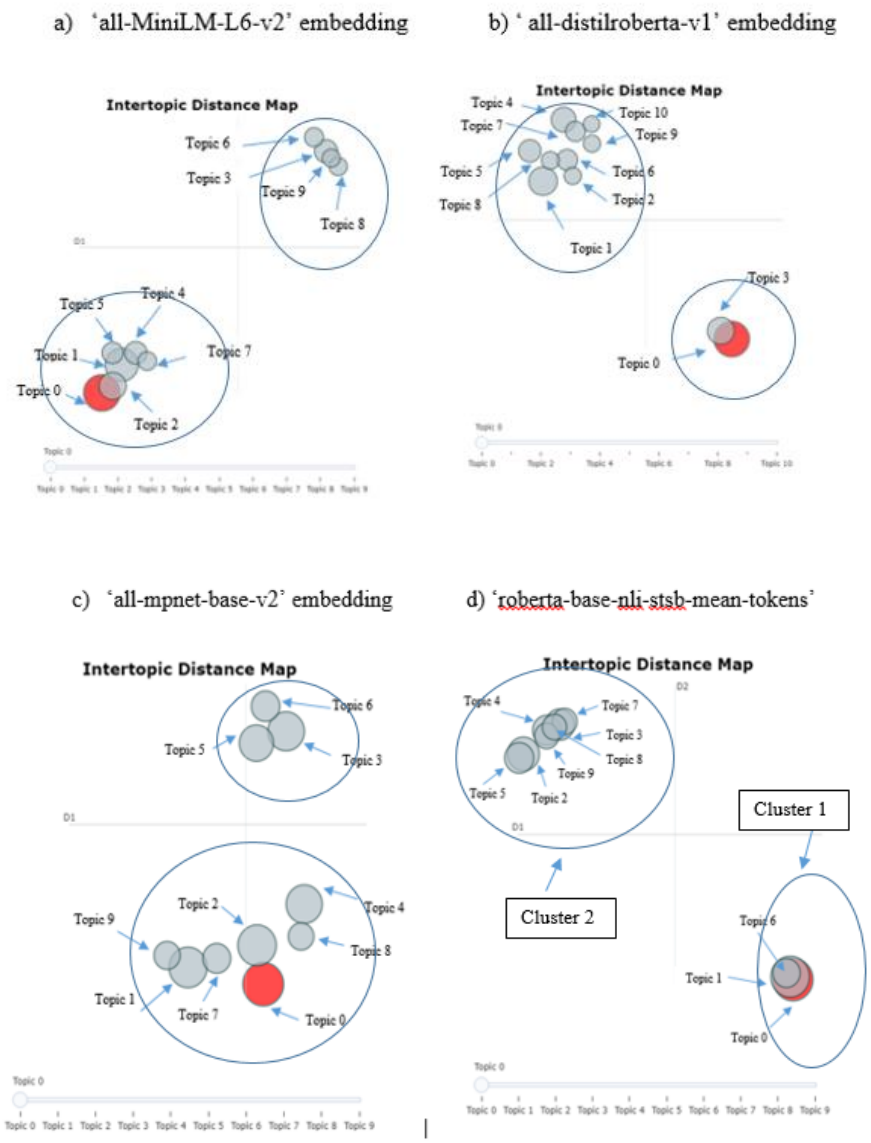

Fig. 5. Inter-topic distance map generated by bertopic embedding's