Exploring Deep Learning Techniques for Topic Modeling of Natural Disaster Tweets

Haardhik Kunder[1], Varun Nair[1], Dhruv Uberoi[1], Prof. Khushali Deulkar[2], Dr. Meera Narvekar[3]

Department of Computer Engineering, Dwarkadas Jivanlal Sanghvi College Of Engineering, Mumbai, India.

Contributing authors: kunderhaardhik@gmail.com; varunr.nair@outlook.com; dhruv.uberoi@gmail.com; khushali.deulkar@djsce.ac.in; meera.narvekar@djsce.ac.in;

Abstract

The sheer quantity and depth of social media data have opened a gateway to understand more about human behavior during certain conditions. The advent of topic modeling models has significantly helped uncover underlying hidden patterns and offered new perspectives on interpreting social phenomena. However, social media content is often brief, text-based and unstructured in nature, presenting difficulties for data collection and analysis. In this paper, we assess four topic modeling techniques: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), BERT Transformer, and Llama3 with BERTopic. Using Twitter posts and tweets made during hurricanes, wildfires, blizzards, and flooding, we implemented a comprehensive preprocessing pipeline that included stop-word removal, word vectorization, data cleaning, and the manual exclusion of unnecessary phrases.This research evaluates the benefits of each technique, highlighting each algorithm's strengths and weaknesses, particularly in the context of social science applications. Furthermore, it sheds light on the efficacy of using traditional topic modeling models and generative models in conjunction with topic modelling to analyse Twitter data.

Keywords: Topic modeling, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), BERT Transformer, Llama3, BERTopic

1 Introduction

Social media has emerged as a dominant channel for communities to gather and spread information during crises [1]. Social media data contain rich information about human activities, environmental conditions, and public sentiment, which geographic information scientists, computer scientists, and domain scientists can use for data analysis [2–4]. The sudden rise in user-generated data during natural disasters like hurricanes, floods, and wildfires is driven by the accessibility of said social media applications. Social media not only generates massive volumes of data, but also a wide variety of data types, such as text, images, and videos [5]. In 2020, there were 3.5 billion social media users worldwide, equivalent to about 45% of the world's population [6]. This surge offers a rich dataset for analyzing public discourse. Social media has opened an entirely new path for social science research, especially when it comes to the overlap between human relations and technology [7]. The user-generated content lets researchers tap into its potential using various advanced machine learning algorithms. Being a source of spatio-temporal information, it can aid us to comprehend emergency situations. Additionally, it can provide rapid and immediate real-time information about events that helps provide greater situational awareness leading to better decision-making [3].

Lots of tweets and images are posted on social media platforms, but many of them are not useful, as these data may not include the information needed by response teams [8]. For this, we have to apply preprocessing on the user-generated data. We can then proceed to perform sentiment analysis and topic modeling on said data.

2 Literature Review

The increasing frequency and intensity of natural disasters necessitate efficient and effective disaster management strategies. In recent years, social media sites have emerged as essential resources for real-time information dissemination and crisis communication. But the vast amounts of unstructured data produced during these events present huge challenges and opportunities for researchers and practitioners in the field.

Chauhan et al. [9] introduced the basics of topic modelling techniques to extract meaningful patterns and insights and review their extensions, such as domain-specific modelling, hierarchical models, word-embedded models, and multilingual models. Additionally, it explored topic modelling in distributed environments and advancements in visualization. It discussed implementation and evaluation techniques, along with comparison matrices of various topic modelling categories. Phengsuwan et al. [5] discussed the approaches for classifying social media data, identifying events, and extracting spatial and temporal information.

The use of BERT in disaster management provides advanced text analysis and sentiment classification, considerably improving real-time situational awareness and responsiveness. Julanta et al. [10] presented a framework that includes the use of specific techniques to analyse unprocessed Twitter data, such as keyword filtering, location-based filtering, and tweet property analysis. BERT was then used to classify tweets from disaster-affected areas into discrete topics that are important and useful for emergency management. Ningsih et al. [11] used BERT to assist rescuers and emergency responders in developing efficient knowledge management procedures for

a quickly changing catastrophe environment. Varghese et al. [12] demonstrated the efficacy of using BERT-based sentiment analysis to optimise disaster management tactics.

Understanding sentiment analysis and communication patterns in disaster management complements topic modelling techniques by providing deeper insights into public reactions and information dissemination during crises. Detera et al. [13] provides a method for analysing the sentiment of locals as well as tourists in Tokyo during earthquake and typhoon scenarios. Jin et al. [14] described the communication patterns that occurred during the various phases of Hurricane Maria, using the crisis and emergency risk communication paradigm. Topic model analysis, LSA, and word-cloud analysis were used to capture the intricacy of communication during this severe occurrence. Ahn et al. [15] collected and analysed tweets from chosen organisations on the July 2019 Ridgecrest earthquake in Southern California, and then used topic modelling to identify both similar and diverse subjects for those organisations. Karimiziarani et al. [16] analyzed twenty million tweets to identify key discussion topics and classify them into humanitarian categories using AI algorithms in NLP, including sentiment analysis and topic modelling.

3 Data Collection and Preprocessing

3.1 Data Collection

Two sets of twitter dataset are used here. They contain tweets made during events of natural disasters and are feed into the topic modeling models after preprocessing.

- Twitter Dataset: This is the primary dataset for our analysis. It consists of comprehensive collection of 49,816 tweets scraped from Twitter related to natural disasters in the United States. This dataset includes specific summaries for major disaster events such as the 2011 tornadoes, Hurricane Sandy in 2012, the 2013 floods, the 2016 blizzard, Hurricane Matthew in 2016, the 2017 hurricanes, Hurricane Michael in 2018, the 2018 wildfires, and Hurricane Dorian in 2019. To prevent bias for one type of disaster, events like hurricanes are omitted from the final dataset. The data was collected using TwitterScraper, a Python library that allows for the extraction of tweets beyond the limitations of Twitter's standard API. The tweets are labeled with sentiment information for a comprehensive grasp of public sentiment during those critical times.

3.2 Data Preprocessing

Initially all tweets are transformed to lowercase ensuring consistency and prevent treating similar words with different cases as separate entities. Common English stopwords were eliminated using the nltk stopword module, an essential step to remove frequently occurring but insignificant words that do not add to the text's overall meaning. Next, noise removal is conducted using regular expressions to eliminate URLs, mentions, hashtags, and other Twitter specific elements, focusing on the substantive content of the tweets. Words considered irrelevant for analysis, like, and, were manually excluded through an exploratory analysis-driven custom word removal step.

Tokenization and lemmatization are carried out in order to simplify the text for analysis by splitting it into separate tokens and converting words to their root forms. However, these steps were omitted in the final preprocessing function due to an error in the provided code snippet. TextBlob is used to perform POS tagging in order to analyze the grammatical structure and classify various word types in the tweets. Examining the labeled texts provided insights into the distribution of different parts of speech, unveiling linguistic features. The text was converted to numerical data using CountVectorizer from sklearn, generating a matrix where every row represents a tweet and every column represents a different word from the text. This matrix records the occurrence of words in tweets and allows for the implementation of machine learning algorithms.

4 Implementation of Topic Models

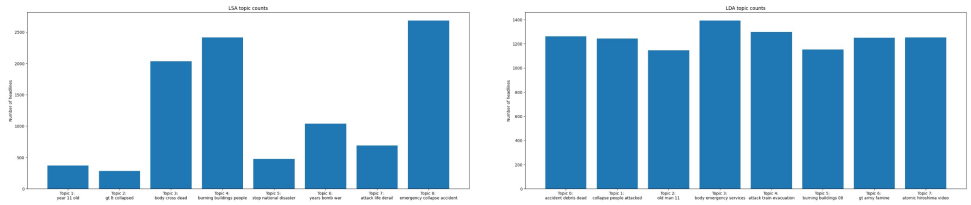
4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a well known generative probabilistic model in the domain of topic modeling. It is considered as a 3-level hierarchical Bayesian model that represents every group item as a finite mixture and represents each topic with a mix of probabilities.

In our research, we conducted a grid search to optimize the values for the three key parameters needed for LDA through a comprehensive grid search. This search involved varying the number of topics (K) as well as the parameters beta and alpha. A higher beta results in topics consisting of more words, while a higher alpha indicates greater diversity among the topics. The search for the optimal number of topics ranged from two to fifteen, with increments of one. Initially, the value of K was set, and the values of alpha and beta were searched for accordingly. Only one hyperparameter was changed during the process, while the others remained constant to obtain the highest coherence score. Finally, pyLDAvis was used to create an intertropic distance map in order to make the extracted information interpretable.

4.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a linear algebra-based natural language processing technique that reveals the latent structure in a text corpus by leveraging Singular Value Decomposition (SVD). It helps uncover the underlying relations between words and concepts. It works on the principle that if words are being used in similar contexts, they tend to have related meanings. In our research, we initialized the model with truncated Singular Value Decomposition (SVD) to lower the term-document matrix's dimensionality. We then assigned each document to the topic of most relevance, based on the highest value in its topic representation matrix. We employed a two pronged approach to identify the most prevalent topics and their characteristic terms. First, we calculated the topic frequencies. Second, we identified the representative terms for each topic by aggregating the term vectors for all documents assigned to that specific topic. To facilitate clear visualization, we leveraged t-SNE to compress the topic vectors in a two-dimensional space.



(a) (b)

Fig. 1 Topic modeling result of LSA and LDA models

4.3 BERTopic

BERTopic is an advanced topic modeling technique that builds upon the mechanisms of Top2Vec, which results in them having similar structures. Utilizing BERT as an embedder, it provides document embedding extraction, and supports a sentence-transformers model for over 50 languages. It also incorporates UMAP for dimensionality reduction and HDBSCAN for clustering. By using the class-based term frequency-inverse document frequency (c-TF-IDF) technique, it sets itself apart from Top2Vec. In order to generate term representations, this method evaluates the relevance of terms inside a cluster; this suggests that a term's greater value indicates its more representativeness of its issue. Unlike LDA, BERTopic provides continuous and not discrete topic modeling. Its stochastic nature leads to different results with repeated modeling. On computation, the most significant topics can be identified.

4.4 Llama 3 with BERTopic

Llama 3 with BERTopic integrates the advanced language understanding capabilities of Llama 3 with the robust topic modeling framework provided by BERTopic.

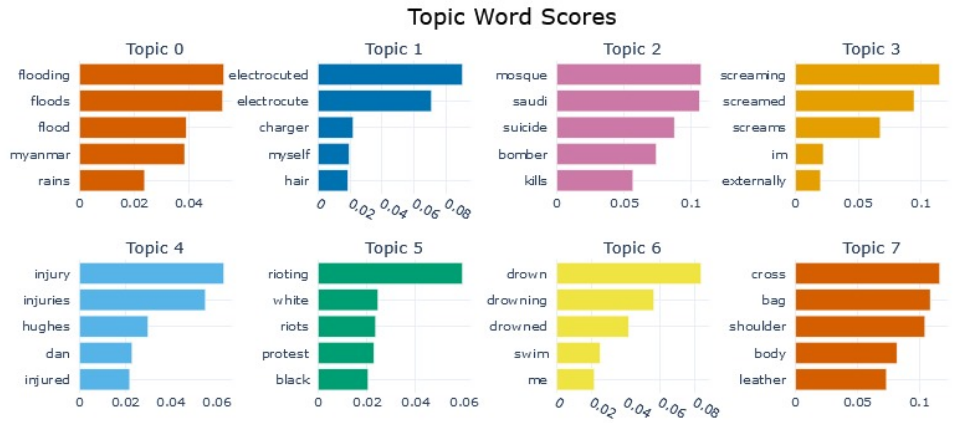


Fig. 2 BERT with Hierarchical Clustering

Fig. 3 Llama3 with BERTopic

Llama 3, a state-of-the-art transformer model, enhances the embedding process by capturing deeper semantic relationships and contextual information within the text. In this study, Llama 3 was used to generate document embeddings, which were subsequently processed by BERTopic. This combination utilized UMAP for dimensionality reduction and HDBSCAN for clustering. The enhanced embeddings from Llama 3 improved topic coherence and representation quality. The class-based TF-IDF algorithm was employed to determine the most relevant terms for each topic. This integrated approach provided a sophisticated tool for analyzing complex textual data, yielding detailed and accurate topic models. An interactive visualization was created to aid in the interpretation and exploration of the results.

5 Results

5.1 Comparison between LSA and LDA

The LSA Model takes the help of singular value decomposition (SVD) to minimize the dimensionality of the data which allows it to capture the underlying structure of words used across tweets. The prevalence of certain major themes like “burning buildings people” and “emergency collapse accident” by the LSA model from Figure 1 displays their significance within the collected tweets. However a noteworthy drawback of the LSA model is its tendency to combine related phrases, which can result in overlapping topics. The t-SNE map makes this overlap clear by showing how several clusters may have common areas, which reflects the diverse contexts in which specific phrases occur together. For instance, there may be a lot of overlap between the terms ”burning buildings people” and ”emergency collapse accident” from Figure 1 as these terms frequently occur together in tweets covering different parts of disaster scenarios.This

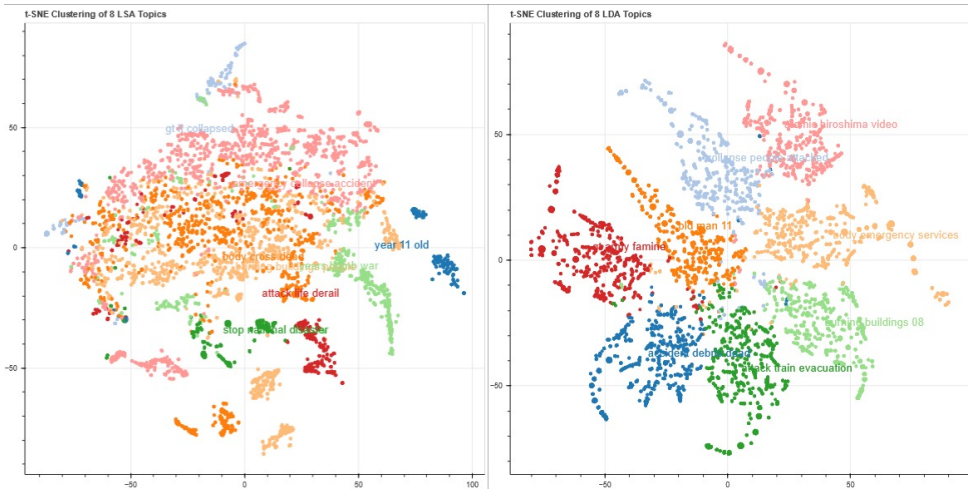


Fig. 4 LSA vs LDA Clustering

overlap indicates that phrases are blended together within a single topic cluster, indicating that the dimensionality reduction process of LSA does not fully maintain the diverse contexts in which terms are employed. A spatial representation of the topics degree of separation is provided by T-SNE clustering Figure 4.

In contrast, the LDA model operates on a generative process, assuming that each tweet arises from a mixture of multiple topics. By modeling the probability distributions of words over topics and topics over documents, LDA provides a more nuanced and balanced distribution of topics. Topic terms identified by the model such as "body emergency services" and "attack train evacuation" from Figure 4 forms well-defined clusters, which therefore shows clearer and more distinct topic boundaries in comparison. This is particularly clear in Figure 4, where LDA distinguishes themes more successfully than LSA in the t-SNE plot. The LDA t-SNE plot's distinct clusters show that, as compared to the LSA model, LDA is better at identifying separate themes in the data, allowing for a more accurate analysis of the underlying subjects. The models LSA and LDA provides insightful analysis of terms and topics hidden in the tweets and have their own unique set of advantages and disadvantages . LSA is better for identifying broad, general themes, making it suitable for analyses that require a broad summary of prevalent topics. When precise topic delineation is required, its tendency to overlap themes might be a disadvantage. On the other hand, LDA excels in providing distinct and well-separated topics, making it ideal for detailed examinations of specific themes.

5.2 Comparison between BERTopic and Llama3 with BERTopic

The BERT model offers insights at the word level for topics in tweets and, when combined with hierarchical clustering, shows the hierarchical connections among these

topics. This method provides a detailed look at the key terms and their organization by similarity. Within each topic group, BERT highlights certain terms, such as "flooding", "Myanmar" and "rains" or in discussions concerning electrocution, "electrocuted", "charger" and "myself". This level of depth makes it easier to understand each subject's subtleties and essential ideas. Additionally, the hierarchical clustering dendrogram shows how subjects are grouped based on similarity, highlighting themes that are closely connected and possible overlaps. "Injuries" and "panic attacks" often come up together when people talk about health problems in groups. This approach enhances the analysis by emphasizing the hierarchical structure of conversations or tweets, which is especially valuable for grasping the wider context of interconnected topics.

BERTopic subjects are effectively clustered into discrete groups by Llama 3 with BERTopic, providing a clear visual depiction of many themes in the dataset. The scatter plot displays distinct clusters and draws attention to emotional outliers. Distinct clusters such as "Flood Struggles and Drama", "Blizzard Emergency and Food Needs" and "Hurricane Michael Relief and Recovery" are easily identifiable. Furthermore, groups such as "Angry Hurricane Michael Survivors" and "Angry FEMA Survivors" imply intense emotional content, which is essential for comprehending public opinion and pressing concerns during natural disasters.

6 Discussion and Conclusion

In our research, we delved into four different topic modeling techniques to analyze how people communicate on social media during natural disasters. We looked at Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), BERT with Hierarchical Clustering, and Llama 3 with BERTopic. Each of these methods showed unique strengths and weaknesses when it came to making sense of the messy, unstructured data from Twitter. What we found was pretty interesting. Although they are both used for topic modeling, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) have different functions.LSA gives a thorough thematic summary but lacks the ability to conduct detailed analysis because of restrictions on topic overlap. LDA on the other hand is better in pinpointing particular topics, making it more suitable for thematic analysis.

Although BERT combined with Hierarchical Clustering may have shortcomings in some areas, it outperformed other techniques in capturing the general thematic contexts. Llama 3 using BERTopic performed well in differentiating theme clusters and offers a broad perspective of the information by highlighting shared patterns and intense emotions. Its capacity to recognize nuanced relationships between subjects makes it especially valuable in comprehending the intricate aspects of public communication in times of crisis. Additionally, Llama 3 and BERTopic's distinct and intersecting topics ensures a thorough examination of the changing discussions following natural calamities.

In conclusion, there are distinct benefits associated with each topic modeling technique when analyzing social media content on natural disasters. Specific research questions, the necessary degree of analytical depth, and the intended application of

the insights should all be taken into consideration when choosing a technique. Future areas for study might look into hybrid approaches, which combine the best features of several methods to offer more comprehensive and nuanced insights into public opinion and the distribution of information during natural disasters. As social media continues to play a pivotal role in disaster communication, the refinement and application of these topic modeling techniques will be instrumental in enhancing our understanding and management of natural disasters. The present work highlights the potential applications of advanced topic modeling techniques in catastrophe management and provides new directions for future investigations in this important field.

References

[1] Burel, G., Alani, H.: Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media, Rochester, NY, USA (2018). https://oro.open.ac.uk/55139/

[2] Li, L., Goodchild, M.F., Xu, B.: Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. Cartography and Geographic Information Science **40**(2), 61–77 (2013) https://doi.org/10.1080/15230406.2013.777139

[3] Wang, Y., Wang, T., Ye, X., Zhu, J., Lee, J.: Using social media for emergency response and urban sustainability: A case study of the 2012 beijing rainstorm. Sustainability **8**(1), 25 (2015) https://doi.org/10.3390/su8010025

[4] Zhao, P., Qin, K., Ye, X., Wang, Y., Chen, Y.: A trajectory clustering approach based on decision graph and data field for detecting hotspots. International Journal of Geographical Information Science, 1–27 (2016) https://doi.org/10.1080/13658816.2016.1213845

[5] Phengsuwan, J., *et al.*: Use of social media data in disaster management: A survey. Future Internet **13**(2), 46 (2021) https://doi.org/10.3390/fi13020046

[6] Mohsin, M.: 10 Social Media Statistics You Need to Know in 2020 [Infographic]. Available online (2020). https://www.oberlo.com/blog/social-media-marketing-statistics

[7] Egger, R., Yu, J.: A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in Sociology **7** (2022) https://doi.org/10.3389/fsoc.2022.886498

[8] Li, X.: Disaster tweet text and image analysis using deep learning approaches (2020). https://krex.k-state.edu/items/4f96534d-34ef-463c-9831-0b9eba8a8680

[9] Chauhan, U., Shah, A.: Topic modeling using latent dirichlet allocation. ACM Computing Surveys **54**(7), 1–35 (2021) https://doi.org/10.1145/3462478

[10] J, J.L.R., A, B., M, K.: Topic modeling based clustering of disaster tweets

using bertopic. In: 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, pp. 1–6 (2024). https://doi.org/10.1109/MITADTSoCiCon60330.2024.10575555

[11] Ningsih, A.K., Hadiana, A.I.: Disaster tweets classification in disaster response using bidirectional encoder representations from transformer (bert). IOP Conference Series Materials Science and Engineering **1115**(1), 012032 (2021) https://doi.org/10.1088/1757-899x/1115/1/012032

[12] Varghese, S.R., Juliet, S., Anitha, J.: Social media analytics for disaster management using bert model. In: 2023 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, pp. 1–6 (2023). https://doi.org/10.1109/ICERCS57948.2023.10434012

[13] Detera, B.J., Kodaka, A., Kohtake, N., Nishino, A., Onda, K.: An english-japanese twitter-based analysis of disaster sentiment during typhoons and earthquakes. In: 2021 IEEE International Symposium on Systems Engineering (ISSE), Vienna, Austria, pp. 1–8 (2021). https://doi.org/10.1109/ISSE51541.2021.9582473

[14] Jin, X., Spence, P.R.: Understanding crisis communication on social media with cerc: topic model analysis of tweets about hurricane maria. Journal of Risk Research **24**(10), 1266–1287 (2020) https://doi.org/10.1080/13669877.2020.1848901

[15] Ahn, J., *et al.*: Understanding public engagement on twitter using topic modeling: The 2019 ridgecrest earthquake case. International Journal of Information Management Data Insights **1**(2), 100033 (2021) https://doi.org/10.1016/j.jjimei.2021.100033

[16] Karimiziarani, M., Moradkhani, H.: Social response and disaster management: Insights from twitter data assimilation on hurricane ian. International Journal of Disaster Risk Reduction **95**, 103865 (2023) https://doi.org/10.1016/j.ijdrr.2023.103865