Group 1

CE 784 Project Report: Semester 2024-25 (II)

Topic Modelling of Incident Reports using Large Language Models

Saurabh Srivastava, Pallab Mandal, Hrishikesh Prasad, Aditya Meena

Abstract

This project explores the use of Large Language Models (LLMs) and Latent Dirichlet Allocation (LDA) for analyzing incident reports collected from news sources. The goal is to extract concise summaries and identify underlying themes across hundreds of reports. Using LLMs, each report was summarized for clarity, and LDA was used to find five major topics. Each topic was then named using the LLM based on the most relevant keywords. This report outlines the dataset, preprocessing steps, methodology, results, and key insights from the analysis.

Keywords: Topic Modeling, NLP, Sentence Transformers, UMAP, HDBSCAN, LDA, pyLDAvis, Road Safety

Introduction

Traffic incidents represent a critical challenge for public safety and urban planning, impacting lives, infrastructure, and economic stability. Traditionally, structured data such as accident counts, locations, and casualty statistics have been used to understand and address these events. However, detailed descriptions of accidents - known as incident narratives - contain deeper and more meaningful information that is often ignored. These unstructured texts can show patterns related to how people behave, environmental conditions, and other important issues that structured data, like numbers and categories, might miss.

In this project, we leverage advancements in Natural Language Processing (NLP) by utilizing Large Language Models (LLMs) to analyze and extract insights from these narratives. Specifically, we use Sentence Transformers to generate high-dimensional semantic embeddings of incident descriptions. Unsupervised clustering techniques such as UMAP and HDBSCAN are then applied to uncover hidden structures within the data. To further interpret and label these clusters, Latent Dirichlet Allocation (LDA) is employed, helping to identify dominant themes and keywords characterizing each group.

The primary objective of this study is to automate the discovery of recurring accident scenarios across Bangladesh, providing a thematic overview that can assist policymakers and urban planners. By highlighting common patterns, this analysis can inform more targeted interventions aimed at improving road safety and reducing traffic-related fatalities.

Dataset Overview

Two datasets were combined from leading Bangladeshi news portals, totaling 401 incident reports. The cleaned dataset included columns for URL, date, headline, content, and portal. The main analysis was performed on the Content column. After preprocessing, 200 records were embedded and analyzed to extract topics.

Key Features (Columns) in the Dataset

Feature Name	Description
URL	Web address of the original news article reporting the road accident.
Date	Date on which the news article was published.
Headline	The title of the news article summarizing the incident.
Content	Full text or snippet of the article describing the road accident.
Portal	Name of the news source (e.g., *Dailystar*).

Initial Exploration of the Dataset

The initial exploration of the dataset involved the following steps:

- Loading the Dataset: The dataset was imported using the Pandas library. It contained columns such as Headline, Content, and Date, with a total of 301 rows (incident reports).
- Checking for Missing Values: We examined the dataset for any missing or null values in key columns like Headline, Content, and Date. Missing values, if any, were handled by:
    - Filling them with empty strings for text columns
    - Parsing dates with error handling for invalid formats
- Previewing and Understanding the Structure: Sample entries were reviewed to understand the type of content in the Headline and Content columns. This helped in confirming the dataset's focus on incident-related news.
- Combining Text Columns: To prepare the data for further analysis, the Headline and Content columns were merged into a new column called Combined_Text, which provided a unified representation of each incident report.
- Basic Statistics and Shape: The dataset was checked for:
    - Total number of rows and columns
    - Distribution of word counts per report
    - Common words appearing in the corpus
- Date Formatting and Parsing: The Date column was converted into a standard datetime format to enable trend analysis over time. Any parsing errors were safely handled using coercion.

Data preprocessing

Before analyzing the data, several preprocessing steps were performed to prepare it for topic modeling. First, the Headline and Content fields were merged to form a single Combined Text column, giving a unified view of each incident report. The text was then cleaned by converting all characters to lowercase, tokenizing the sentences using NLTK, and removing common stopwords such as "the" and "and" to focus on meaningful words. Additionally, only alphabetic tokens were retained, filtering out numbers and punctuation. For time-based analysis, the Date column was converted into a standard datetime format, handling any parsing issues gracefully. Finally, a dictionary and a corpus were created from the cleaned text, capturing the frequency of words in each document. These were essential inputs for building the LDA topic model.

Methodology

Topic modeling is a method in natural language processing (NLP) that aims to discover the underlying themes or subjects within a large set of textual data. This study focused on analyzing road traffic accident reports, topic modeling serves as a valuable tool to detect recurring patterns and significant themes within the descriptions of various incidents. These insights can help highlight potential causes and contributing factors of accidents.

One of the most widely used algorithms for topic modeling is Latent Dirichlet Allocation (LDA). Introduced by Blei et al. in 2003, LDA is a generative probabilistic model that represents each document as a mixture of topics and each topic as a mixture of words. The model operates under several key assumptions:

- Each document contains a random number of words, with that number drawn from a Poisson distribution.
- Topic proportions for each document are sampled from a Dirichlet distribution, denoted by the parameter θ_n.
- Each word within a document is associated with a topic, assigned according to a multinomial distribution based on the document's topic proportions.
- Each topic is characterized by a distribution over words, represented by β_k.
- The actual words in the documents are drawn from a multinomial distribution based on these topic-specific word distributions.

LDA works by assigning and refining the probability distributions of words across topics and documents. Through multiple iterations, it identifies a set of hidden topics that effectively explain the observed textual data, offering a structured interpretation of otherwise unorganized text.

LLM Summarization

The first step in our methodology involved generating concise summaries of each incident report using a Large Language Model (LLM). We utilized the LLaMA 3 (70B) model integrated through the LangChain framework. The model was prompted with the full content of each article and instructed to produce a one-line summary capturing the key details of the incident. This summarization step served two main purposes: it helped reduce the length and complexity of the text for downstream processing, and it provided quick insights into each report without needing to read the entire article

These summaries made the dataset more interpretable and streamlined the topic modeling phase by providing a cleaner, more focused input.

Topic Modeling with LDA

To uncover the hidden thematic structure of the dataset, we applied Latent Dirichlet Allocation (LDA), a widely-used unsupervised topic modeling technique. The cleaned and preprocessed text data, including the tokenized and filtered word lists, was used to create a dictionary and a document-term matrix (corpus). The LDA model was trained on this corpus with the number of topics set to five, as this number was found to balance interpretability and granularity.

The model was run for 15 passes to ensure better convergence and stability of the topics. Once trained, the LDA model produced:

- A distribution of topics for each document
- A list of top 10 keywords that best represent each topic

Each report was then assigned its dominant topic—the one with the highest probability in its topic distribution. This classification allowed for grouping similar incidents and further exploration of common themes within the dataset.

Topic Naming using LLM

While the LDA model successfully identified clusters of similar reports, the topics themselves were represented by lists of keywords, which were not always easy to interpret. To make the topics more user-friendly and descriptive, we again employed the LLaMA 3 language model. For each topic, the top 10 keywords identified by LDA were passed to the model, which was prompted to suggest a meaningful and unique name for the topic.

For example, given keywords like “police,” “bus,” “accident,” and “fire,” the model suggested the topic name “Safety Watch”. This process helped transform abstract keyword sets into intuitive and human-readable labels, making the analysis more accessible and insightful for users and stakeholders.

Results

Identified Topics and Names

Topic No.	Top Keywords	Assigned Name
1	police, bus, fire, students, area	Safety Watch
2	road, bus, killed, accidents, dhaka	Wheels of Tragedy
3	road, accident, transport, blood	StreetBeat
4	injured, killed, upazila, people	Road to Tragedy
5	hospital, injured, urban, news	Urban Crisis

These names capture the essence of the underlying incidents and make interpretation easier.

Topic Distribution

After assigning each document its dominant topic using the LDA model, we analyzed how the incident reports were distributed across the five topics. The results revealed that the majority of the articles fell under Topic 3 and Topic 1, indicating that a significant portion of the news reports focused on road-related incidents, such as traffic accidents, and general public safety events, including police and emergency responses.

- Topic 3 (named "StreetBeat") includes keywords like accident, road, transport, and blood, reflecting a strong emphasis on street-level incidents and traffic issues.
- Topic 1 (named "Safety Watch") is characterized by keywords such as police, bus, fire, and students, representing broader safety concerns often involving public transport or emergency services.

The remaining topics, while smaller in volume, still covered important themes:

- Topic 2 ("Wheels of Tragedy") concentrated on fatal road accidents, especially in urban areas.
- Topic 4 ("Road to Tragedy") focused on injuries and casualties in specific regions (e.g., upazilas).
- Topic 5 ("Urban Crisis") dealt with hospitalizations, city-specific issues (like those in Dhaka), and emergency medical responses.

To better understand the distribution of reports across these topics, a bar chart was generated. This visualization clearly showed the number of documents associated with each topic, helping to identify which themes were most prevalent in the dataset.

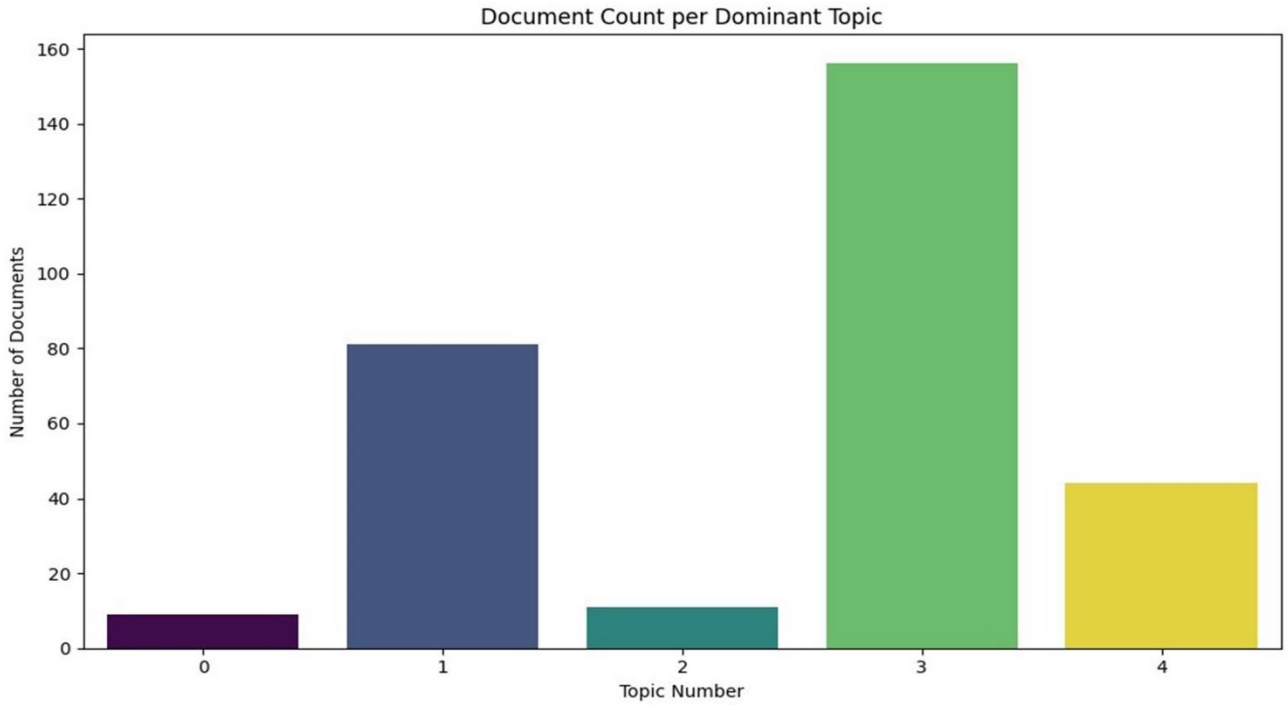

Figure 1: Document Count per Topic

Topic Trends Over Time

To gain deeper insights into how the focus of incident reporting changed over time, a temporal analysis of topic trends was carried out. The publication dates of the reports were used to observe how the proportion of each topic varied on a monthly basis throughout the dataset. This approach helped in understanding the changing prominence of different types of incidents over time. By grouping the reports by month and calculating the average topic probabilities, it became possible to track which topics became more or less frequent during specific periods. This analysis offered valuable information about seasonal patterns, event-driven spikes, and long-term shifts in incident themes.The process involved the following steps:

- Assigning Topic Probabilities: Instead of using only the dominant topic per document, we retrieved the full topic distribution for each report using the LDA model. This means each document was assigned a probability score for all five topics, reflecting how strongly it related to each theme.
- Aggregating by Month: The Date column, which had already been parsed into datetime format during preprocessing, was grouped by month using pandas' Period functionality. The topic probabilities were averaged for each month, resulting in a time series of average topic relevance.
- Visualizing Topic Trends: A multi-line plot was created to show the trend of each topic across months. The x-axis represented time (monthly intervals), while the y-axis showed the average topic probability. Each line corresponded to a topic, making it easy to observe rises and falls in topic prominence over time.

This time-series visualization allowed us to identify:

- Spikes in specific topics, such as sudden increases in reports related to traffic accidents (e.g., during holidays or adverse weather seasons).
- Seasonal patterns, such as more reports involving urban incidents during certain months.
- Shifts in focus, showing when attention might have moved from general safety issues to more specific road or health-related events.

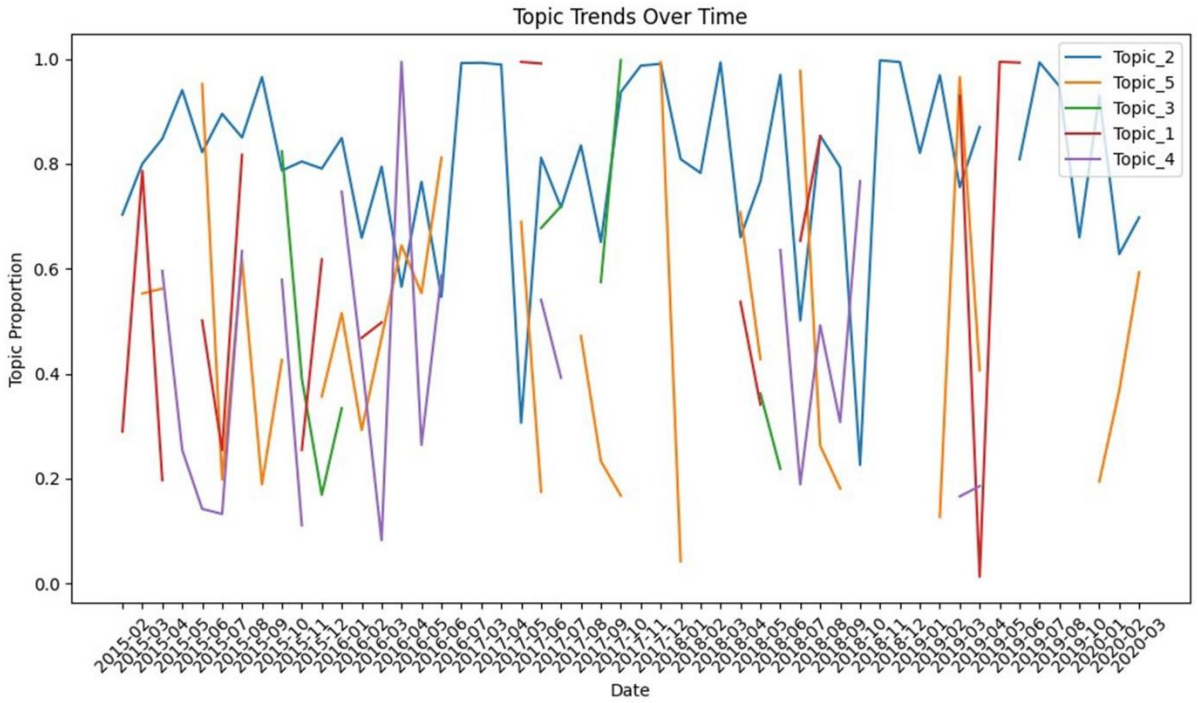

Figure 2: Topic Trends Over TimeConclusion

Conclusion

This study demonstrates the effectiveness of integrating advanced Natural Language Processing (NLP) techniques—specifically Large Language Models (LLMs) and Latent Dirichlet Allocation (LDA)—for the analysis of large-scale incident reports. By applying these tools to a corpus of over 300 news articles, meaningful thematic insights were extracted without the need for manual review of each individual report.

The use of LLMs enabled the generation of concise, human-readable summaries, which improved the readability of the dataset and provided a high-level understanding of each report. This was particularly beneficial for handling lengthy and information-dense articles. Subsequently, LDA topic modeling was employed to uncover latent themes within the corpus. The model effectively grouped semantically similar reports and identified key terms representative of each topic.

To enhance interpretability, the LLM was further utilized to generate descriptive names for each identified topic based on the top keywords provided by the LDA model. This step transformed abstract topic representations into more accessible and meaningful labels.

In addition, visualizations were developed to illustrate the distribution of documents across topics and to analyze temporal trends. These graphical representations highlighted patterns in incident reporting, including dominant themes and fluctuations in topic prominence over time.

Overall, the combined methodology offers a scalable and interpretable framework for exploring unstructured textual data. Beyond incident reports, this approach holds potential for application in diverse domains such as healthcare, aviation, and customer service, where large volumes of textual records require efficient summarization and thematic categorization.

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
[2] Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

Github Link: https://github.com/pallabm22/News2Topics-LDA-based-Topic-Modeling

Contribution:

Saurabh Srivastava –

- Conducted topic labeling using LLM by providing top keywords from each LDA topic
- Conducted initial data loading and preprocessing, including merging headline and content fields
- Assisted in refining the summarization prompts to generate concise one-line descriptions
- Collaborated in interpreting the LDA topic modeling output and assigning descriptive topic names using LLM
- Contributed to writing the Methodology, Results, and Conclusion sections of the report

Pallab Mandal –

- Led the integration of the Large Language Model (LLaMA 3) using LangChain for summarizing incident reports
- Performed text cleaning, tokenization, and stopword removal using NLTK
- Created the dictionary and corpus required for LDA modeling
- Built and optimized the LDA model by experimenting with the number of topics and passes
- Drafted the Dataset Overview, Initial Exploration, and Data Preprocessing sections

Hrishikesh Prasad –

- Extracted and processed publication dates for use in time-series analysis of topic trends
- Computed full topic probability distributions and grouped documents by month for temporal trend analysis
- Generated the multi-line plot showing topic trends over time
- Helped interpret seasonal and event-driven changes in topic prominence
- Contributed to writing the Topic Trends Over Time and Interpretation of Results sections

Aditya Meena –

- Validated and refined the LDA-generated topics with real-world context and examples
- Participated in visual design of the final figures for report inclusion
- Handled report formatting and final compilation
- Wrote the Abstract, Topic Naming, and formatted the References section