Group-03, Data Science Section (C), Final Term Project

1st Faysal Ahmmed   
*Dept. of Computer Science and Engineering   
American-International University-Bangladesh*  
22-47069-1@student.aiub.edu

4th Asef Rahman Antik  
*Dept. of Computer Science and Engineering   
American-International University-Bangladesh*  
22-47106-1@student.aiub.edu 2nd Ajmy Alaly  
*Dept. of Computer Science and Engineering   
American-International University-Bangladesh*  
22-46733-1@student.aiub.edu

3rd Samanta Mehnaj  
*Dept. of Computer Science and Engineering   
American-International University-Bangladesh*  
22-47081-1@student.aiub.edu

*Abstract*—Topic modeling is a widely used technique for uncovering latent themes in textual data, offering valuable insights across various domains. However, the vast amount of unstructured textual data available on online news platforms remains underexplored for identifying coherent topics in real-time. This study aims to address this gap by applying Latent Dirichlet Allocation (LDA) to perform topic modeling on data scraped from CNN news articles. Using web scraping, a dataset of recent articles will be collected, preprocessed, and analyzed using LDA. The identified topics will be evaluated for coherence and relevance, providing a clear depiction of underlying themes in the news corpus. The findings are expected to reveal prominent topics discussed in CNN news, demonstrating the utility of LDA in extracting meaningful insights from unstructured data. This study contributes to the field of natural language processing by showcasing the application of topic modeling on contemporary news data.

Keywords—Topic modeling, Latent Dirichlet Allocation (LDA), CNN news, web scraping, natural language processing, text analysis, unstructured data, topic extraction

# Introduction

The digital era has ushered in an unprecedented surge of information, particularly from online news platforms. This vast expanse of unstructured textual data presents both opportunities and challenges for researchers aiming to extract meaningful insights. Traditional manual analysis methods are often inadequate due to the sheer volume and velocity of data generation. Consequently, automated techniques have become indispensable for processing and understanding large-scale text corpora.

Topic modeling has emerged as a pivotal technique in natural language processing (NLP) to identify hidden thematic structures within extensive text datasets. Among the various methods available, Latent Dirichlet Allocation (LDA) is particularly prominent. Introduced by Blei et al. in 2003, LDA is a generative probabilistic model that posits each document as a mixture of topics, with each topic being a distribution over words [1]. This approach allows for the discovery of latent topics that pervade a collection of documents, facilitating a deeper understanding of the underlying themes.

The application of LDA in analyzing news articles has been explored in various studies. For instance, Alghamdi and Alfalqi conducted a comprehensive survey on topic modeling in text mining, highlighting the efficacy of LDA in extracting topics from large text corpora [2]. Similarly, Wang et al. utilized LDA to analyze news content related to the COVID-19 pandemic, uncovering significant themes and trends in media coverage [3]. These studies underscore the potential of LDA in distilling complex information from vast textual datasets.

To harness the power of LDA for real-time news analysis, it is imperative to have access to current and comprehensive datasets. Web scraping serves as a potent tool in this regard, enabling the automated extraction of data from websites. By deploying web scraping techniques, researchers can systematically collect up-to-date articles from news platforms such as CNN. This process involves navigating web pages, identifying relevant content, and extracting information such as headlines, publication dates, authors, and full texts [4]. The amalgamation of web scraping and LDA facilitates the dynamic analysis of news content, allowing for the identification of emerging topics and trends.

However, the integration of web scraping and LDA is not without challenges. Web scraping must be conducted responsibly, adhering to legal and ethical guidelines, and ensuring compliance with the terms of service of target websites. Moreover, the preprocessing of scraped data is crucial to enhance the quality of topic modeling outcomes. This includes tasks such as removing HTML artifacts, normalizing text, and filtering out irrelevant content [5]. Addressing these challenges is essential to ensure the reliability and validity of the analysis.

In this study, we aim to develop a comprehensive framework that combines web scraping of CNN news articles with LDA-based topic modeling. By systematically collecting and analyzing news content, we seek to identify and interpret the dominant themes present in the data. This approach not only demonstrates the practical application of LDA in processing large-scale, unstructured textual data but also contributes to the understanding of prevalent topics in contemporary news media. The insights garnered from this analysis can inform various applications, including trend analysis, sentiment assessment, and information retrieval.

# Literature review

M. Zheng et al. [6] address the challenge of determining the optimal number of topics (K) in Latent Dirichlet Allocation (LDA) models for news text analysis. Traditional approaches for selecting K often rely on heuristic methods, which lack flexibility and adaptability to varying datasets. The authors propose a novel adaptive method that dynamically determines K by analyzing perplexity and coherence scores across different values. This method identifies the point where increasing the number of topics no longer improves model performance significantly. Experimental validation on diverse news datasets demonstrates that the adaptive approach outperforms fixed-K methods in terms of topic coherence and relevance. The study highlights the flexibility and robustness of the adaptive method, making it suitable for real-world applications in news topic identification. By improving the interpretability of LDA results, this work contributes significantly to enhancing topic modeling for large-scale unstructured text data[6].

In the study conducted by D. Buenaño-Fernandez et. al. [7] explore the application of Latent Dirichlet Allocation (LDA) to analyze open-ended responses from university educators' self-assessments. Traditional methods of evaluating such qualitative data often involve manual coding, which can be time-consuming and subjective. By employing LDA, a generative probabilistic model commonly used in topic modeling, the researchers aim to uncover latent themes within the textual data, providing a more objective and scalable analysis. The study demonstrates that LDA can effectively identify prevalent topics in educators' self-reflections, offering insights into common areas of focus and concern.

S. Zoya et al. [8] explore the effectiveness of Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) in extracting topics from Urdu-language tweets. Given the unique linguistic characteristics of Urdu and the brevity of tweets, traditional topic modeling approaches face challenges in maintaining coherence and relevance. The researchers conducted experiments using various feature extraction methods, including Term Frequency-Inverse Document Frequency (TF-IDF) and short-text pooling techniques. Their findings indicate that NMF, when applied with TF-IDF feature vectors, outperforms other combinations in capturing coherent topics from Urdu tweets. Conversely, LDA shows improved performance when short texts are merged prior to modeling. This study underscores the importance of selecting appropriate modeling techniques and preprocessing strategies tailored to the specific characteristics of the language and dataset in question.

In their 2016 study, Naskar et al. [9] introduce a methodology for analyzing sentiments in social network conversations by integrating topic modeling with sentiment analysis. The authors employ the Latent Dirichlet Allocation (LDA) model to uncover latent topics within conversations. To associate these topics with specific emotions, they utilize the Affective Norms for English Words (ANEW) dictionary, which provides affective ratings for words. This approach allows for a nuanced understanding of the emotional context of topics discussed in social networks. The study further explores the formation of user communities based on shared sentiments, providing insights into how users with similar emotional expressions cluster together in social networks.

# Methodology

This section outlines the steps undertaken to perform web scraping, preprocess textual data, and implement topic modeling for analyzing news articles sourced from CNN. The methodology is divided into three primary components: data collection, text preprocessing, and topic modeling.

## Data Collection via Web Scraping

The first step in this research involved collecting news articles from CNN’s official website. Web scraping was utilized to extract relevant content from the news website, ensuring a comprehensive dataset for analysis. The scraping process targeted article links from the homepage collecting all the types of news articles.

Each link was accessed programmatically, and the article content was extracted using HTML parsing techniques. Care was taken to focus only on the primary content of the articles while avoiding irrelevant sections such as advertisements, navigation menus, or footer information. The extracted data included the main body text of the articles, which was stored for further preprocessing.

Additionally, the dataset was structured into a tabular format containing the article URLs and their corresponding text content. This structure facilitated ease of manipulation during subsequent stages. Duplicate or incomplete entries were removed to ensure data consistency.

## Text Preprocessing

The next step involved preprocessing the textual data to prepare it for topic modeling. Preprocessing is a crucial stage in natural language processing (NLP) that aims to clean, normalize, and tokenize the text, ensuring that the data is suitable for advanced analysis. The following steps were performed during preprocessing:

### Text Cleaning:

Text cleaning was performed to remove extraneous elements such as HTML tags, special characters, and numerical data. This ensured that only meaningful textual content was retained for analysis. Additionally, the text was converted to lowercase to standardize the input.

### Tokenization:

The cleaned text was tokenized, splitting it into individual words or tokens. Tokenization allows for the isolation of meaningful units of text, which are essential for downstream processing and analysis.

### Stop Words Removal:

Common stop words (e.g., “and,” “the,” “is”) were removed from the text to focus on more meaningful terms. Stop words often contribute little to the overall semantic understanding and are usually excluded to improve computational efficiency and accuracy.

### Stemming and Lemmatization:

Words were reduced to their base or root form through stemming and lemmatization. This process ensures that variations of a word (e.g., “running,” “runs”) are treated as a single term, improving the interpretability of the results.

### Handling Contractions:

Contractions (e.g., “can’t,” “won’t”) were expanded into their full forms (e.g., “cannot,” “will not”) to preserve the semantic meaning of the text.

### Handling Emojis and Emoticons:

Emojis and emoticons were replaced with textual descriptions or removed, depending on their relevance to the context of the analysis.

### Spell Checking:

Spell-checking was applied to correct any typographical errors in the text. Suggested corrections were applied to ensure accuracy and consistency in the dataset.

The result of the preprocessing stage was a clean and normalized corpus of text that was ready for topic modeling.

## Topic Modeling

After preprocessing the data, topic modeling was conducted to uncover latent themes within the dataset. Latent Dirichlet Allocation (LDA) was employed as the primary topic modeling technique. LDA is a probabilistic model that assumes that each document in a corpus is a mixture of topics, and each topic is a distribution of words. The following steps were performed during topic modeling:

### Creating the Document-Term Matrix (DTM)

The preprocessed text was converted into a document-term matrix (DTM), a numerical representation where each row corresponds to a document and each column corresponds to a term. Each cell in the matrix represents the frequency of a term within a document. The matrix was carefully filtered to ensure that only relevant terms with significant frequency were included. Removing sparse terms ensured that the model remained efficient and focused on meaningful data.

### Calculating TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) weighting was applied to the DTM to measure the importance of terms in the corpus. This method assigns higher weights to terms that appear frequently in a document but are rare across the corpus, emphasizing their relevance to specific topics.

### Topic Modeling with LDA

LDA was applied to the DTM to extract topics. The model was initialized with a pre-determined number of topics (K). The iterative nature of the LDA algorithm allowed for the distribution of topics across documents and terms.

### Examining the Topics

Once the LDA model was trained, the resulting topics were examined and interpreted. Two key steps were involved in this process:

*4.1) Extracting the Most Probable Words*

For each topic, the most probable words were identified. These words provided insights into the semantic structure of each topic and facilitated their interpretation.

*4.2) Calculating Topic Proportions*

The proportion of each topic within each document was calculated, enabling the identification of dominant themes across the corpus.

### Interpreting the Results

The final step involved interpreting the identified topics. Each topic was assigned a meaningful label based on its most probable words. The results were analyzed to uncover patterns and trends within the dataset, providing valuable insights into the themes present in CNN news articles.

The methodology outlined above integrates web scraping, text preprocessing, and topic modeling to analyze news articles. The combination of robust preprocessing techniques and the LDA algorithm ensures accurate and interpretable results. This workflow not only uncovers latent themes in the dataset but also demonstrates the utility of NLP methods in analyzing large-scale textual data. The insights derived from this analysis can inform future research and applications in text analytics. **Fig. 1** Summarizes the Methodology Structure.

**Fig. 1.** Proposed Methodology for Topic Modeling using Web-Scrapped data

# Result and Discussion

The topic modeling analysis produced 12 distinct topics, each represented by the most probable terms.

The topic modeling analysis was performed on 60 news documents extracted from CNN, using Latent Dirichlet Allocation (LDA). This analysis produced some distinct topics. The key results are summarized as follows:

### Topic Identification

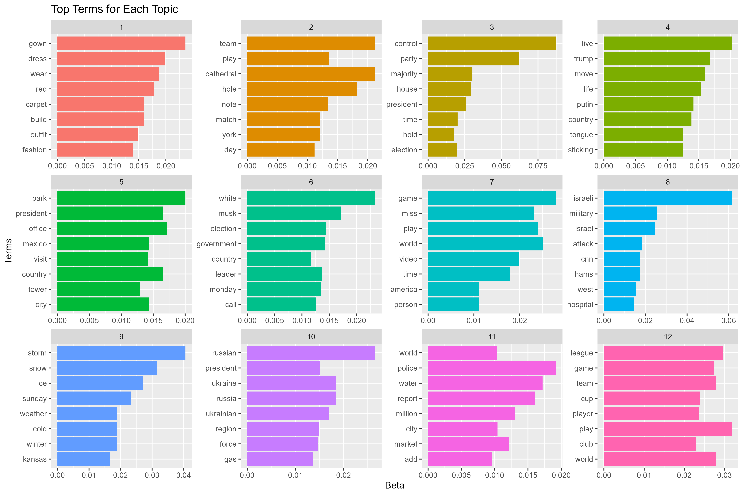
The analysis identified 12 distinct topics, each represented by the most probable terms. These topics reflect a range of themes, including fashion, politics, sports, weather, geopolitics, and societal issues. The most probable terms for each topic were visualized in a bar graph in **Fig. 2**, clearly highlighting the dominant terms within each topic.

The results in **Fig 2.** highlight the key themes within the dataset, as outlined below:

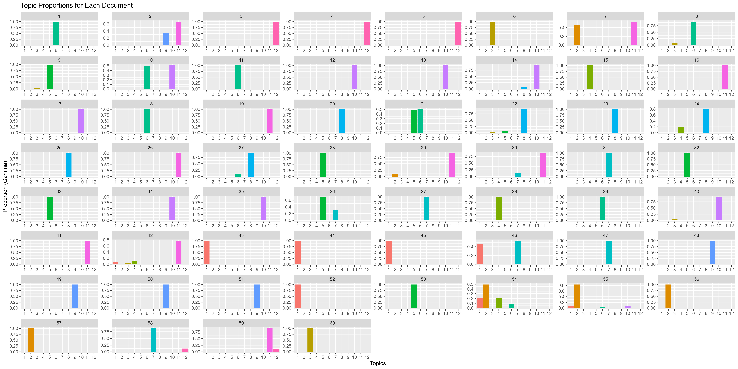
1. Topic 1: This topic appears to be related to fashion and red carpet events, with terms such as "gown," "dress," "wear," and "outfit" dominating.
2. Topic 2: Keywords like "team," "play," and "match" suggest this topic revolves around sports or teamwork.
3. Topic 3: This topic focuses on governance and political events, with words like "control," "party," "majority," and "election."
4. Topic 4: Words such as "live," "Trump," and "Putin" indicate this topic involves political figures and current affairs.
5. Topic 5: With terms like "president," "office," "Mexico," and "visit," this topic points to diplomatic relations and political activities.
6. Topic 6: Keywords "White," "Musk," and "election" suggest discussions surrounding prominent figures, governmental actions, and leadership.
7. Topic 7: This topic is linked to entertainment and digital media, indicated by terms like "game," "video," "time," and "America."
8. Topic 8: Focused on military and conflict-related discussions, terms like "Israeli," "military," and "attack" dominate this topic.
9. Topic 9: The topic highlights weather-related themes with terms such as "storm," "snow," and "winter."
10. Topic 10: Words like "Russia," "Ukraine," and "gas" point to geopolitical issues and energy politics.
11. Topic 11: This topic reflects societal and environmental concerns, with words like "water," "market," and "report."
12. Topic 12: Keywords such as "league," "game," and "team" suggest a focus on sports events and tournaments.

### Topic Proportions Across Document:

Another important output of the analysis was a graph showing topic proportions for each document. Each of the 60 documents was represented as an individual bar graph in **Fig. 3**, illustrating how strongly it was associated with different topics. For example, Document 2 was found to be 30% related to Topic 9 (weather-related) and 70% to Topic 11 (environmental concern). Similarly, other documents exhibited varying proportions of relevance to one or more topics.



**Fig. 2.** Top Words for 12 Different Topics



**Fig. 3.** Topic Proportion for Each Document

The topic modeling results effectively identified 12 coherent themes within the CNN news dataset, including fashion, politics, sports, weather, and geopolitics. The bar graph of top terms for each topic highlighted the key terms, ensuring interpretability. Document-topic proportions revealed varying levels of association between documents and topics, showing that some documents were highly focused on a single topic, while others covered multiple themes.

These insights demonstrate the versatility of the LDA model in uncovering thematic structures in unstructured data. While the results are coherent, some overlap in themes suggests opportunities for refinement, such as increasing the number of topics or enhancing preprocessing steps. Future applications could include sentiment analysis or dynamic topic modeling to track thematic trends over time. The findings are valuable for tasks like document categorization, media analysis, and understanding narrative diversity in news articles.

# Conclusion

This study successfully applied topic modeling using Latent Dirichlet Allocation (LDA) to analyze a dataset of CNN news articles, identifying 12 distinct and coherent themes. The preprocessing steps, including text cleaning, tokenization, and stop-word removal, ensured the dataset was well-prepared for analysis. The visualization of top terms for each topic and document-topic proportions provided valuable insights into the thematic structure of the dataset, showcasing the diversity and complexity of news content.

The results demonstrate the effectiveness of LDA in uncovering latent themes and their distribution across documents. Topics such as politics, weather, sports, and geopolitics were accurately identified, highlighting the model's robustness. Additionally, the document-topic proportions revealed the multifaceted nature of many documents, emphasizing the richness of the dataset and the flexibility of the model.

This work contributes to advancing text analysis in real-world applications, such as media monitoring, content categorization, and public sentiment analysis. While the results were interpretable and meaningful, future enhancements could include dynamic topic modeling to track thematic changes over time and sentiment analysis for a deeper understanding of public perception. Overall, this approach provides a scalable and reliable method for extracting actionable insights from large-scale unstructured text data.

##### Acknowledgment

The We extend our heartfelt gratitude to American International University-Bangladesh for providing the infrastructure and resources required to conduct this research. We are deeply thankful to Tohedul Islam for his invaluable guidance and encouragement throughout this study.

##### References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
2. R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 6, no. 1, 2015.
3. M. Wang and P. Mengoni, "How pandemic spread in news: text analysis using topic model," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2020, pp. 764-770.
4. M. Dallabetta, C. Dobberstein, A. Breiding, and A. Akbik, "Fundus: A Simple-to-Use News Scraper Optimized for High Quality Extractions," *arXiv preprint arXiv:2403.15279*, 2024.
5. S. Sahu, K. Divya, N. Rastogi, P. K. Yadav, and Y. Perwej, "Sentimental Analysis on Web Scraping Using Machine Learning Method," *Journal of Information and Computational Science (JOICS)*, ISSN 1548-7741, 2022.
6. M. Zheng, K. Jiang, R. Xu, and L. Qi, "An Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification," *IEEE Access*, vol. 11, pp. 92273-92284, 2023, doi: 10.1109/ACCESS.2023.3308520.
7. D. Buenaño-Fernandez, M. González, D. Gil, and S. Luján-Mora, "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach," *IEEE Access*, vol. 8, pp. 35318-35330, 2020, doi: 10.1109/ACCESS.2020.2974983.
8. S. Zoya, F. Latif, F. Shafait, and R. Latif, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," *IEEE Access*, vol. 9, pp. 127531-127547, 2021, doi: 10.1109/ACCESS.2021.3112620.
9. D. Naskar, S. Mokaddem, M. Rebollo, and E. Onaindia, "Sentiment analysis in social networks through topic modeling," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 46-53.

A person with a beard wearing a blue shirt

Description automatically generated**Faysal Ahmmed** was born in Rajshahi, Bangladesh in 2001. He is currently pursuing a Bachelor of Science degree in computer science and engineering (CSE) at the American International University-Bangladesh (AIUB), Dhaka, Bangladesh. He is in his final year of study, with a focus on advancing his expertise in cutting-edge technologies.

His research interests include machine learning, deep learning, robotics, and the development of user-friendly software solutions. He is passionate about leveraging these technologies to create impactful and innovative applications that enhance user experience and improve lives.

**A person wearing a head scarf

Description automatically generated**Hailing from Faridpur, Bangladesh, and born in 2001, **Ajmy Alaly** is pursuing her Bachelor of Science in Computer Science and Engineering (CSE) at AIUB, Dhaka. With a flair for artificial intelligence and deep learning, she envisions a future where technology reshapes society in transformative ways. Her academic pursuits are focused on computer vision and innovative problem-solving through advanced computing technologies. Ajmy is devoted to tackling pressing societal issues, and her research endeavors are driven by a commitment to develop impactful AI solutions that address real-world challenges efficiently.

**A person in glasses and a red dress

Description automatically generated**Born in Jamalpur, Bangladesh, in 2002, **Samanta Mehnaj** is completing her Bachelor of Science degree in Computer Science and Engineering (CSE) at the American International University-Bangladesh (AIUB). Known for her passion for technological innovation, Samanta thrives on mastering complex concepts like machine learning, artificial intelligence, and deep learning. Her research focuses on leveraging these technologies to develop pioneering solutions that solve real-world problems. She brings a creative perspective to her work and believes in using advanced computing tools to inspire progress and societal betterment.

A person with a badge on his neck

Description automatically generatedBorn in 2001 in Kushtia, Bangladesh, **Asef Rahman Antik** is a final-year student pursuing a Bachelor of Science in Computer Science and Engineering (CSE) at the American International University-Bangladesh (AIUB). With aspirations to lead in the fields of artificial intelligence, computer vision, and deep learning, Asef is driven by the transformative potential of technology. His work embodies a vision of using innovation to bridge the gap between technological advancements and societal needs. Asef is particularly interested in advancing data science methodologies to create meaningful and impactful solutions for real-world challenges.