**Unsupervised Context Detection in Articles via Topic Modeling**

# Mushfequr Rahman, Apon Kumar Das, Mst.Ummay Fatema Mim, Mst Mafroza Haque Iqra

American International University - Bangladesh (AIUB)

anantamushfiq7@gmail.com, apondas007890@gmail.com,[ummay.fatema.2021@gmail.com,](mailto:mrebollo@dsic.upv.es) mafrozahaque462@gmail.com

***Abstract—* This study explores the application of unsupervised topic modeling techniques for automatic context detection in textual articles.** **Using Latent Dirichlet Allocation (LDA)​, the research identifies thematic structures within unstructured text, enabling the categorization of articles based on inferred topics [1]. The approach involves preprocessing steps such as tokenization, stop word removal, stemming, lemmatization, and topic assignment, followed by probabilistic topic distribution analysis​ [2]. Additionally, automatic labeling methods are employed to assign meaningful context tags​ [3]. The findings highlight the effectiveness of unsupervised learning in uncovering latent themes and contextual patterns in diverse textual corpora. This research contributes to advancements in text mining and automated content analysis, offering applications in information retrieval, recommendation systems, and knowledge management. Furthermore, the proposed methodology enhances document classification accuracy by leveraging topic distributions rather than relying solely on keyword-based approaches. Future work can explore deep learning-based enhancements to further improve topic coherence and interpretability.**

***Index Terms*—** **Topic Modeling, Latent Dirichlet Allocation (LDA), Unsupervised Learning, Context Detection, Machine Learning.**

**INTRODUCTION**

With the rapid growth of digital content, efficient methods for organizing and analyzing large-scale textual data have become increasingly important. One of the key challenges in natural language processing (NLP) is the automatic detection of contextual themes within unstructured textual data. Traditional rule-based or supervised approaches require extensive labeled datasets and domain-specific knowledge, making them less scalable for large and dynamic text corpora. In contrast, unsupervised topic modeling techniques offer a promising alternative by automatically discovering latent themes without the need for labeled training data [1].

Among various topic modeling methods, Latent Dirichlet Allocation (LDA) has emerged as a widely used approach for discovering hidden topics in documents by analyzing word co-occurrence patterns [2]. LDA represents each document as a mixture of topics, where each topic is a probability distribution over words. This enables context detection and document classification based on underlying themes rather than relying solely on surface-level keyword matching [3]. Through probabilistic inference, LDA facilitates topic extraction from large, unstructured text datasets, making it particularly useful in applications such as news categorization, sentiment analysis, content recommendation, and trend detection [2].

The effectiveness of topic modeling largely depends on data preprocessing techniques such as tokenization, stopword removal, stemming, and lemmatization. These preprocessing steps ensure that redundant and irrelevant elements are filtered out before conducting probabilistic topic distribution analysis [3]. Additionally, automatic topic labeling enhances the interpretability of extracted topics by associating them with meaningful descriptors derived from high-probability words within each topic cluster [1].

Recent studies have demonstrated that LDA-based unsupervised topic modeling can significantly improve the efficiency of text mining and information retrieval tasks [4]. Unlike supervised classification models that require extensive human annotations, LDA learns the hidden thematic structure of documents without manual intervention, allowing it to be applied to vast collections of text from diverse domains [5]. Furthermore, topic modeling techniques have been successfully employed in analyzing social media conversations, detecting fake news, summarizing research articles, and understanding user-generated content trends [6].

Despite its advantages, LDA has certain limitations, including difficulty in handling short texts and a dependence on optimal hyperparameter tuning to ensure meaningful topic distributions [7]. Recent advancements in deep learning-based topic models and hybrid approaches integrating LDA with word embeddings (e.g., Word2Vec, BERT) aim to address these challenges and further enhance topic coherence and classification accuracy [8].

This study investigates the effectiveness of unsupervised context detection in textual articles using LDA. By leveraging probabilistic modeling techniques, the proposed approach enables the automatic identification and classification of topics within large and diverse document collections. The findings contribute to advancements in automated content analysis, recommendation systems, and knowledge management, paving the way for future research in hybrid and deep learning-enhanced topic modeling.

**Literature Review**

Numerous studies have explored the effectiveness of unsupervised topic modeling, particularly Latent Dirichlet Allocation (LDA), in identifying hidden thematic structures and improving sentiment analysis. Griffiths and Steyvers [3] laid the foundation for LDA by demonstrating its capability in uncovering latent topics within large text corpora. Building upon this, Mehrotra etal. [2] addressed the challenges of applying LDA to short-text data, such as microblogs, by introducing tweet pooling strategies and automatic topic labeling to enhance coherence and interpretability. Further advancing the field, Naskar et al. [1] integrated sentiment analysis with topic modeling, leveraging context-aware sentiment detection to analyze emotional trends in social networks. Collectively, these studies highlight LDA’s potential in context detection, sentiment classification, and social media analysis, while also pointing toward future improvements through hybrid models and deep learning-based enhancements.

Topic modeling, particularly LDA, has been widely adopted for discovering latent topics in text corpora without requiring labeled data. The study by Griffiths and Steyvers (2004) introduced LDA as a probabilistic method for identifying hidden thematic structures in large text collections​ [3]. Their research demonstrated how documents could be represented as mixtures of topics, where each topic consists of probabilistic distributions of words. The study emphasized that LDA can significantly improve document classification, information retrieval, and text clustering. However, they acknowledged challenges such as choosing the optimal number of topics and maintaining topic coherence.

Building on this foundation, Mehrotra et al. (2013) explored improvements to LDA, particularly for short-text documents such as microblogs and tweets [​2]. The authors pointed out that traditional LDA struggles with sparse text sources, where individual tweets or short posts do not contain enough words to effectively define a topic. To address this, they proposed tweet pooling strategies, where multiple related tweets were aggregated before applying LDA. Their experiments showed that hashtag-based pooling and burst-score pooling significantly improved topic coherence and retrieved more interpretable topic structures in social media datasets.

**2. Sentiment Analysis and Context Detection using Topic Modeling**

The study by Naskar et al. (2021) focused on sentiment analysis in social networks using LDA​ [1]. Unlike traditional sentiment classification techniques that rely on polarity-based lexicons, they introduced Sent-LDA, a hybrid approach integrating LDA with sentiment analysis to detect context-dependent sentiment trends. Their model leverages ANEW dictionaries to map topic distributions onto emotion categories, allowing for fine-grained sentiment classification rather than just positive, negative, or neutral sentiments. This approach provided more contextual awareness in sentiment detection, particularly useful in social media discourse analysis.

Moreover, their findings confirmed that sentiment diffusion in online communities is strongly influenced by topic relevance, meaning that users discussing similar topics tend to express similar sentiments. This aligns with Mehrotra et al. (2013), who also observed that users participating in specific topic clusters exhibit homogeneous linguistic patterns, making LDA-based topic modeling an effective tool for understanding social discourse dynamics [​2].

**3. Advancements in Automatic Topic Labeling**

One of the key limitations of LDA is the interpretability of discovered topics, as LDA provides topics as word distributions without explicit labels. Mehrotra et al. (2013) addressed this issue by proposing automatic topic labeling, where high-probability topic words were matched against predefined domain-specific lexicons​ [2]. They found that domain-specific labeling improved human interpretability of topics, making LDA-based insights more actionable for real-world applications like recommendation systems and text summarization.

Similarly, Naskar et al. (2021) emphasized sentiment-aware topic labeling, incorporating contextual emotional tags into LDA-derived topic clusters [​1]. Their approach aligned sentiments with topic categories, enhancing thematic coherence in sentiment classification tasks. This was particularly effective in social media applications, where topics often carry emotional connotations that influence user engagement patterns.

**4. Applications and Future Directions**

The reviewed studies highlight several promising applications of LDA-based topic modeling and sentiment analysis, including:

* Social media monitoring and fake news detection​ [1].
* Trend identification in real-time textual data​ [2].
* Scientific document classification and knowledge management [​3].
* Context-aware recommendation systems​ [2].

However, Griffiths and Steyvers (2004) noted that choosing the optimal number of topics remains a fundamental challenge [​3]. Both Mehrotra et al. (2013) and Naskar et al. (2021) proposed improvements by enhancing topic coherence using context aggregation and sentiment alignment [​2] [​1]. Future research directions could include hybrid models integrating deep learning and topic modeling, such as BERT-based topic classification or word embeddings to refine LDA outputs.

The reviewed studies establish LDA as a powerful tool for unsupervised context detection and topic-based sentiment analysis. While traditional LDA provides a strong probabilistic foundation, recent advancements—such as automatic labeling, tweet pooling, and sentiment-aware topic alignment—enhance its applicability to real-world datasets. Future research should focus on hybrid deep learning approaches that further refine topic coherence, interpretability, and domain-specific sentiment classification.

**DATA COLLECTION**

In the data analysis work, data collection is often the most time-consuming part of the research. In this study, data has been scraped from two reputable news websites: BBC News and Daily Star English. The scraping process involved collecting sports-related articles from Daily Star and health-related articles from BBC News.

To extract data from these websites, a web scraping technique called web extraction was used. Web scraping involves programmatically extracting data from web pages and structuring it into a usable format. The extracted data was then stored in a structured dataset for further analysis. This approach ensured that the collected data was relevant, comprehensive, and up to date.

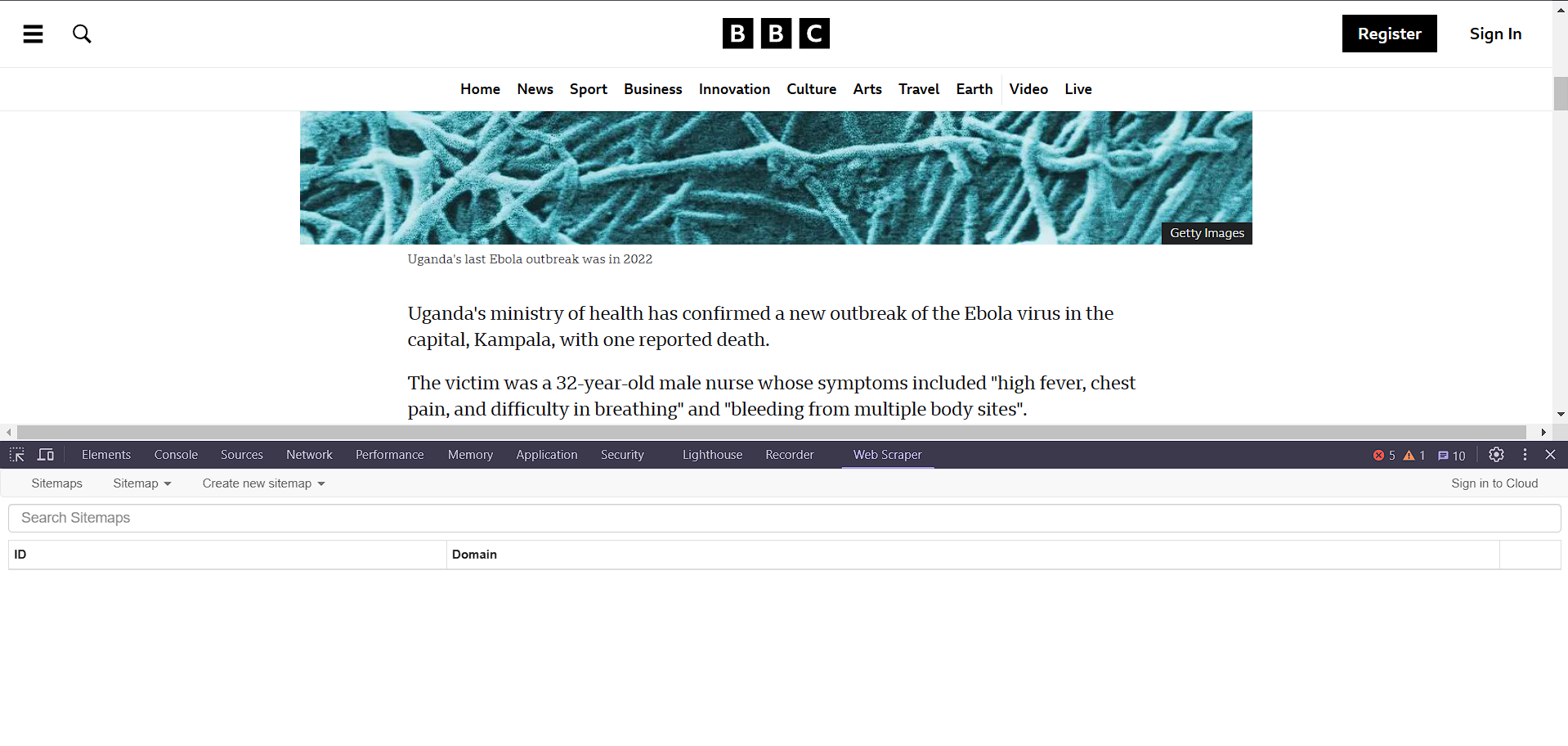


Figure 1: Web Scraping Interface via which Web Scraping was Done.

**PREPROCESSING**

Once the data was collected, several preprocessing steps were applied to clean and prepare it for analysis. The R programming language was used for preprocessing due to its powerful text processing and data manipulation capabilities. The following steps were performed:

1. **Data Cleaning**: The dataset was merged from multiple sources using the rbind function, ensuring a consolidated dataset.
2. **Text Cleaning**:
   * Converted text to lowercase to maintain uniformity.
   * Removed punctuation, numerical values, and HTML tags.
   * Eliminated non-alphabetic characters.
3. **Tokenization**: The text was broken down into individual words using the tokenizers package.
4. **Stopwords Removal**: Common words that do not contribute to the meaning (e.g., "the", "is", "and") were removed using the tm package.
5. **Stemming and Lemmatization**: Words were reduced to their root forms using the SnowballC and textstem libraries.
6. **Spelling Correction**: The hunspell package was used to correct misspelled words.

After these preprocessing steps, the dataset was transformed into a structured and refined form suitable for analysis.

A screenshot of a computer screen

Description automatically generated

Figure X: Corrected Words After Preprocessing in Web Scraped Data.

**DATA ANALYSIS & TOPIC MODELING**

To identify key topics in the collected articles, Latent Dirichlet Allocation (LDA) was applied using the topicmodels package in R. LDA is a probabilistic model that helps in identifying topics from a large corpus of text. The following steps were taken:

1. **Document-Term Matrix (DTM) Creation**: The text corpus was converted into a document-term matrix using the tm package.
2. **TF-IDF Weighting**: Term frequency-inverse document frequency (TF-IDF) was applied to give importance to significant words.
3. **LDA Topic Modeling**: The LDA algorithm was implemented with k=2 (two topics: Sports and Health) to extract dominant topics.
4. **Topic Labeling**:
   * The extracted words from the topics were matched with predefined keyword lists for Sports and Health.
   * A function was developed to assign appropriate labels based on the highest-scoring category.

# A computer screen shot of text Description automatically generated

# Figure: Labeled Topics Identification Using LDA.

1. **Topic Distribution Analysis**:The topic distribution was visualized using the ggplot2 package, providing insights into the frequency of sports and health topics in the collected data.

A graph of a number of bars

Description automatically generated with medium confidence

Figure X: Topic Distribution of Health and Sports Using LDA.

Topic Distribution of Health and Sports Using LDA illustrates the number of articles categorized into the topics "Health" and "Sports" based on the Latent Dirichlet Allocation (LDA) model. The chart reveals that the "Health" category contains a slightly higher count of articles compared to "Sports," indicating a stronger emphasis on health-related content in the analyzed dataset. This distribution reflects the nature of the data collected from the sources, suggesting that BBC News, from which health-related articles were scraped, may prioritize health topics more frequently, while Daily Star's focus on sports contributed to the sports-related articles. The visualization provides insight into the relative prevalence of these topics, helping to understand the thematic trends in the web-scraped dataset.

**Processed Dataset with Topic Probabilities from LDA**



Figure: Updated Dataset with Topic Probabilities and Labels.

This figure showcases the dataset after undergoing text preprocessing and topic modeling. Each row corresponds to an article, showing the transformations such as tokenization, stopword removal, stemming, lemmatization, and spelling corrections. The columns for "Topic 1" and "Topic 2" display the probabilities calculated by the Latent Dirichlet Allocation (LDA) model for each article belonging to the respective topics (e.g., Sports and Health). The "Topic\_Label" column indicates the final category assigned to each article based on the highest topic probability.

# Limitations

# While the proposed method for unsupervised context detection using topic modeling is effective, it has certain limitations. The results heavily depend on the quality of the input data and the preprocessing steps. Errors in data scraping, such as missing or irrelevant content, can negatively impact the accuracy of topic modeling. Additionally, the number of topics (k) must be pre-defined in Latent Dirichlet Allocation (LDA), which might lead to suboptimal results if not chosen carefully. The model also struggles with overlapping topics or articles that cover multiple contexts. Moreover, the predefined keyword lists used for labeling topics can introduce bias, limiting the adaptability of the approach to datasets with different themes.

# RELATED WORK

# Unsupervised topic modeling has been widely applied in various domains for context detection. Naskar et al. explored sentiment analysis in social networks through topic modeling, demonstrating its utility in identifying underlying themes [1]. Mehrotra et al improved LDA for microblogs by pooling tweets and automatic labeling, emphasizing the importance of preprocessing for short text data [2]. Griffiths and Steyvers discussed the use of LDA for finding scientific topics, showcasing its adaptability in academic contexts [3]. Similarly, Blei et al introduced Latent Dirichlet Allocation as a foundational model for topic detection [5]. The application of LDA to traditional media and social media was compared by Zhao et al. highlighting differences in topic distributions [6]. These studies underline the versatility of topic modeling for unsupervised context detection in diverse datasets.

# CONCLUSION

# This study demonstrates the effectiveness of unsupervised context detection in articles using topic modeling. By leveraging Latent Dirichlet Allocation (LDA) and a structured preprocessing pipeline, articles from BBC News and Daily Star were categorized into health and sports topics. The analysis highlights the importance of preprocessing, topic labeling, and visualization in extracting meaningful insights from text data. Despite limitations such as dependence on predefined parameters and keyword lists, the approach provides a scalable solution for thematic analysis in unstructured text. Future work could focus on integrating advanced embeddings or neural topic models to enhance accuracy and flexibility in context detection.

# References

1. D. Naskar, S. Mokaddem, M. Rebollo, and E. Onaindia, "Sentiment Analysis in Social Networks through Topic Modeling," Universitat Politecnica de Valencia and University of Oran, 2012. [Online]. Available: debashis@drtc.isibang.ac.in.
2. R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling," in *Proceedings of SIGIR'13*, Dublin, Ireland, July 2013, pp. 20-30. [Online]. Available: ResearchGate.
3. T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," in *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228-5235, Apr. 2004. [Online]. Available: ResearchGate.
4. B. Chaney and D. M. Blei, “Visualizing topic models,” in *International Conference on Machine Learning (ICML)*, 2012, pp. 1–8.
5. D. Blei, A. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
6. W. X. Zhao et al., “Comparing Twitter and traditional media using topic models,” in *European Conference on Information Retrieval (ECIR)*, 2011, pp. 338–349.
7. A. Dieng, F. J. Ruiz, and D. Blei, “Topic modeling in embedding spaces,” *Transactions of the Association for Computational Linguistics (TACL)*, vol. 8, pp. 439–453, 2020.
8. J. Gao, D. Chen, and P. Li, “Enhancing topic modeling using pre-trained contextualized embeddings,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 3173–3185.