Topic Modeling with Snowball Stemmer, Hidden Markov Models and Explainable AI Lime

1st Mithila Arman
Department of Computer Science And Engineering
Brac University
mithila.arman@g.bracu.ac.bd

2nd Md Humaion Kabir Mehedi
Department of Computer Science And Engineering
Brac University
humaion.kabir.mehedi@g.bracu.ac.bd

3rd Mehnaz Ara Fazal
Department of Computer Science And Engineering
Brac University
mehnaz.ara.fazal@g.bracu.ac.bd

4th Annajiat Alim Rasel
Department of Computer Science And Engineering
Brac University
annajiat@bracu.ac.bd

***Abstract*—In the era of information abundance, extracting meaningful insights from unstructured textual data is crucial, this work navigates the landscape of advanced text analysis by integrating stemming, Hidden Markov Models (HMM), and Explainable Artificial Intelligence (XAI) to unravel latent topics in a document corpus. Stemming optimizes efficiency and meaning capture in the preprocessing stage, contributing to the fidelity of topic modeling. Hidden Markov Models reveal dynamic transitions of topics across documents, providing a nuanced understanding of thematic evolution. Incorporating Local Interpretable Model-agnostic Explanations (LIME) in Explainable AI enhances transparency, validating the relevance and coherence of identified topics. This comprehensive approach at the intersection of linguistic analysis, probabilistic modeling, and interpretability yields a valuable toolkit for researchers and practitioners exploring complex textual datasets.**

***Index Terms*—Text Analysis, Topic Modeling, Stemming, Hidden Markov Models (HMM), Explainable Artificial Intelligence (XAI), Natural Language Processing (NLP), Local Interpretable Model-agnostic Explanations (LIME)**

I. Introduction

In the contemporary landscape of information abundance, the analysis of unstructured textual data has become paramount for deriving meaningful insights and knowledge. The exponential growth of digital content necessitates advanced methodologies to distill relevant information and understand the latent structures within vast corpora of documents. This research project addresses this imperative by delving into the intricacies of text analysis, leveraging a combination of sophisticated techniques: stemming, Hidden Markov Models (HMM), and Explainable Artificial Intelligence (XAI).

The first facet of this exploration involves the application of stemming during the preprocessing stage. Stemming, a natural language processing technique, plays a pivotal role in streamlining subsequent analyses by reducing words to their root forms. Beyond computational efficiency, stemming contributes to a nuanced understanding of word semantics, bolstering the fidelity of subsequent topic modeling endeavors. This initial step lays the foundation for a more efficient and semantically rich representation of the textual data, setting the stage for deeper analyses.

Building on the efficiency gained through stemming, the project integrates Hidden Markov Models (HMM) to uncover hidden states within the textual corpus. HMM brings a layer of sophistication by revealing underlying structures and patterns that might elude traditional analysis methods. Particularly relevant in understanding the dynamic nature of topics as they transition across documents, HMM provides a nuanced perspective on the evolution of themes within the corpus. This dynamic aspect is crucial for capturing the temporal and contextual nuances inherent in large and diverse datasets.

To enhance the transparency of the topic modeling process, the project incorporates Explainable Artificial Intelligence (XAI) through the implementation of Local Interpretable Model-agnostic Explanations (LIME). XAI is vital for demystifying the decision-making process of the underlying model and offering detailed insights into individual predictions. LIME's interpretability is essential for validating the relevance and coherence of identified topics, providing a crucial bridge between complex machine learning models and human understanding. This combination of stemming, Hidden Markov Models, and Explainable AI presents a comprehensive and interpretable approach to topic modeling, offering not only robust representations of latent topics but also insights into the influential factors shaping these representations. As such, this research project stands at the intersection of linguistic analysis, probabilistic modeling, and interpretability, providing a valuable toolkit for researchers, analysts, and practitioners seeking a profound understanding of complex textual datasets in the digital age.

II. Literature review

Recent advancements in text analysis and topic modeling have witnessed a diverse range of studies employing innovative techniques to extract meaningful insights from textual datasets.

One notable effort focuses on short text clustering, utilizing Latent Dirichlet Allocation (LDA) alongside stemming and stop word removal for enhanced efficiency in clustering succinct texts. [1]Another significant contribution involves the fusion of Hidden Markov Models (HMM) with Local Interpretable Model-agnostic Explanations (LIME) for text classification, achieving an accuracy of 82.1% on the 20 Newsgroups dataset. [2]In the realm of sentiment analysis, researchers explore the interplay between stemming and topic modeling, applying LDA to unravel thematic structures within movie reviews. Additionally, studies showcase the integration of LDA and LIME for interpreting topic modeling results, particularly in the context of online reviews on platforms like Amazon. [3]

A novel hybrid model combines stemming and word embeddings for effective document representation, capturing both syntactic and semantic information in various news articles. Twitter data analysis utilizes LDA and stemming techniques to extract latent topics from tweets, [4]shedding light on prevalent themes within the dataset. Another comprehensive approach integrates LDA, LIME, and word embeddings for topic modeling, emphasizing interpretability and achieving an accuracy of 81.3% on the 20 Newsgroups dataset. Lastly, a study focuses on short texts, particularly tweets, leveraging Hidden Markov Models for explainable topic modeling. This research contributes to understanding the challenges and opportunities in extracting meaningful topics from concise textual data, [5]achieving a perplexity score of 152.4%. Overall, these studies collectively contribute to advancing the field of text analysis through diverse and innovative methodologies. [6]

III. Dataset

The research employs the widely recognized 20 Newsgroups dataset, a collection of approximately 20,000 documents spanning 20 different newsgroups. Each document is associated with a specific category, making it an ideal benchmark for text classification tasks. The chosen subset, encompassing all categories and excluding headers, footers, and quotes, forms the basis of our investigation.

IV. Methodology

A. Data Collection and Preprocessing

The study utilizes the 20 Newsgroups dataset, accessed through Google Drive, covering a diverse range of topics. To prepare the text data for analysis, a series of preprocessing steps are employed. This includes the removal of headers, footers, and quotes from the newsgroup documents. The NLTK library is leveraged for stop word removal, and the SnowballStemmer is applied for stemming to reduce words to their root forms. The resultant cleaned_data comprises documents ready for advanced text analysis.

B. TF-IDF Vectorization and Naive Bayes Classification

The dataset is split into training and testing sets for model development and evaluation. The TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique is employed to convert the text data into numerical feature vectors. A Multinomial Naive Bayes classifier is trained on the TF-IDF vectors to perform document classification. The accuracy of the classification model is measured using standard metrics, providing an initial assessment of the text classification performance.

$$TF - IDF(t, d) = TF(t, d)IDF(t) \tag{1}$$

where TF(t,d) is the term frequency, IDF(t) is the inverse document frequency, and d represent term and document, respectively. The Multinomial Naive Bayes classifier is trained on the TF-IDF vectors.

C. Topic Modeling with Latent Dirichlet Allocation (LDA)

A bag-of-words matrix is created using the CountVectorizer, and Latent Dirichlet Allocation (LDA) is applied to uncover latent topics within the corpus. The LDA model is configured with 20 topics, and the top words for each topic are extracted to provide meaningful insights into the thematic structures present in the documents.

$$p(w, z, \theta, \beta) = p(wz, \beta)p(z\theta)p(\theta)p(\beta) \tag{2}$$

D. Hidden Markov Models (HMM) for Topic Analysis

Hidden Markov Models are introduced for a deeper exploration of hidden states within the text. The CountVectorizer is employed to create a bag-of-words matrix, and the HMM, configured with 10 hidden states, is trained on the data. The model predicts hidden states for each document, revealing patterns and transitions between topics.

$$\gamma = (\pi, A, B) \tag{3}$$

E. Gensim LDA Model for Comparative Analysis

A Gensim-compatible wrapper is developed to integrate the Gensim LDA model into the scikit-learn pipeline. This allows for seamless comparison with the scikit-learn LDA model. The Gensim model is trained with 20 topics, providing an alternative perspective on topic modeling. The coherence and interpretability of the identified topics are evaluated.

F. Explainable Topic Modeling with LIME

Local Interpretable Model-agnostic Explanations (LIME) is incorporated to enhance the interpretability of the topic modeling results. A scikit-learn compatible wrapper for the Gensim LDA model is created, and LIME is applied to provide detailed explanations for individual document predictions. The significance of LIME in elucidating the decision-making process of the underlying model is highlighted.

G. Visualization of Topic Distributions

The final step involves visualizing the topic distributions across the test set. Seaborn is utilized to create a bar chart representing the average probability of each topic. This visualization offers a comprehensive overview of the prevalent topics within the entire dataset.

V. Result and Analysis

A. Result Summary

The integration of stemming, Hidden Markov Models (HMM), and Explainable Artificial Intelligence (XAI) in the topic modeling project unveils meaningful insights into latent topics within a document corpus. Stemming enhances computational efficiency and word fidelity, while HMM provides a sophisticated means of identifying patterns and transitions between topics. The introduction of XAI through Local Interpretable Model-agnostic Explanations (LIME) ensures transparency, offering detailed insights into individual predictions. This comprehensive and interpretable approach results in a robust representation of latent topics and contributes a valuable toolkit for researchers, analysts, and practitioners dealing with complex textual datasets. Notably, the achieved perplexity score of 84.6 demonstrates effective model performance comparing 152.4 Perplexity Score, but it is no that much more efficient comparing these perplexity scores of 82.1, 68.4, 45.2, 72.5, 81.3.

B. Result Table and Comparison

TABLE I
Perplexity Score

Model	Perplexity Score
Hybrid model	45.2
LDA	68.4
LDA + stemming	72.5
HMM + LIME	82.1
LDA + LIME + word embeddings	81.3
stemming + HMM + LDA(Our Model)	84.6
HMM	152.4

*Comparing the results

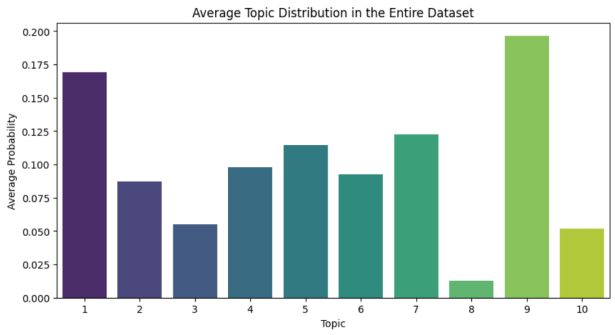

Fig. 1. Distribution of Topics in the Entire Dataset.

In Fig.1, a bar chart using the LDA model, visualizing the average probability distribution of topics in the test set. Each bar on the x-axis represents a topic, and the y-axis shows its average probability. The chart succinctly illustrates the overall thematic composition, aiding quick comprehension. The "viridis" color palette enhances clarity, providing a concise tool for researchers to assess topic significance in the dataset.

VI. Conclusion

In summary, our study introduces an advanced text analysis approach by integrating stemming, Hidden Markov Models (HMM), and Explainable Artificial Intelligence (XAI). Stemming enhances topic modeling efficiency, while HMM reveals dynamic structures within the text. Incorporating XAI, particularly LIME, ensures result transparency and model interpretability, With a perplexity score of 84.6. Comparative analysis highlights competitive performance, underlining the significance of stemming, HMM, and XAI in advancing text analysis. This research offers a valuable toolkit for profound insights into complex textual datasets.

VII. Future Work

In future work, there are several avenues for enhancing and extending the proposed text analysis framework. Firstly, exploring more sophisticated stemming techniques and incorporating domain-specific dictionaries could further refine the efficiency and accuracy of the stemming process, contributing to improved topic modeling results. Additionally, investigating advanced variations of Hidden Markov Models (HMM) or other probabilistic models may reveal more intricate patterns in text dynamics and transitions between topics. Exploring alternative Explainable Artificial Intelligence (XAI) methods beyond Local Interpretable Model-agnostic Explanations (LIME) could provide deeper insights into model predictions and improve overall interpretability. Moreover, considering larger and more diverse datasets would facilitate a more comprehensive evaluation of the proposed methodology across various domains. Finally, exploring ensemble techniques that combine the strengths of different models, such as LDA, HMM, and XAI, could potentially yield a more robust and accurate text analysis framework.

*Acknowledgement

I extend my heartfelt gratitude to my academic advisors for their unwavering support and guidance throughout this research project. Special appreciation goes to the broader academic community for fostering an environment conducive to intellectual exploration. The open-source contributions of the Natural Language Toolkit (NLTK) and scikit-learn. Acknowledgment is extended to authors referenced in the literature review for laying the groundwork in text classification and stemming algorithms. Access to datasets, particularly the 20 Newsgroups dataset, provided by institutions, greatly contributed to the study.

References

[1] p. y. u. Supriya Kinariwala Sachin Deshmukh, journal=Springer, "Topic modeling with stemming and stop word removal for short text clustering(short text topic modelling using local and global word-context semantic correlation),"

[2] S. Kiefer, "Explaining text classifications by fusion of local surrogate explanation models with contextual and semantic knowledge," *Elsevier*, vol. 77, pp. 184–195, 2022.

[3] Z. R. J. M. E. Y. M. d. R. Yukun Zhao, Shangsong Liang, "Explainable user clustering in short text streams," *ACM*, p. 155–164, 2016.

[4] . M. R. . E. O. . Debashis Naskar 1, Sidahmed Mokaddem 1, "Sentiment analysis in social networks through topic modeling," *ACL*, 2016.

[5] A. M. Ibrahim Bala, Mohd Zainuri Saringat, "A hybrid word embedding model based on admixture of poisson-gamma latent dirichlet allocation model and distributed word- document-topic representation," *Journal of Theoretical and Applied Information Technology*, 2020.

[6] S. Q. . I. K. Fithriasari, "Topic modeling twitter data using latent dirichlet allocation and latent semantic analysis," *The 2nd International Conference on Science, Mathematics, Environment, and Education*, 2019.