Evaluating Traditional And Transformer-Based Topic Models For Indian Financial News Analytics

Vrishali A. Chakkarwar[1], Dr. Sharvari C. Tamane[2]

[1] Research Scholar, Department of Computer Science and Engineering, Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar -431004, Maharashtra, India (e-mail: vachakkarwar@geca.ac.in)
[2] Director & Professor, University Department of Information & Communication Technology MGM University, Chhatrapati Sambhajinagar , Maharashtra, India (E-mail- Directoriict@mgmu.ac.in)

Abstract

Topic modeling allows the unsupervised learning of latent thematic representations in large textual collections which can in turn form the basis of semantic analyses and later uses such as trend analysis and classification. Considering the complexities of the Indian financial markets and the limitations of purely quantitative models, the study will conduct a comparative evaluation of the Latent Dirichlet Allocation (LDA), Nonnegative Matrix Factorization (NMF), and BERTopic, on a dataset of 50,000 headlines and descriptions of Indian business news that were scraped from Business Standard between 2003 and May 2020. LDA is described as a probabilistic generative model, NMF as a non-negative matrix factorization model, and BERTopic as a transformer architecture that combines SentenceBERT embedding with UMAP dimensionality reduction, HDBSCAN clustering, and class-based TF-IDF weighting.

The qualitative analysis was conducted through the comparative analysis of salient words and word clouds. The alternative methods may not be as effective as BERTopic when evaluating contextual coherence and defining clear topical boundaries because it allows the discrimination between closely related financial subdomains. The findings emphasize the effectiveness of embedding-based topic modeling approaches in achieving greater interpretability and advancing the scope of financial text analytics.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Non-negative Matrix Factorization, BERTopic, Unsupervised Learning.

1 *Introduction*

1.1 *Topic modeling and its significance in text analysis*

Topic modeling is an unsupervised machine learning technique used to uncover the underlying thematic structure within a large collection of documents. It identifies latent topics by analyz-ing patterns of word co-occurrence across texts, enabling the automatic discovery of themes without requiring labeled data. This process transforms unstructured text data into a structured
representation of topics, which can then be used for various applications such as document classification, information retrieval, and summarization.

Topic modeling plays a significant role in text analysis due to its capacity to uncover latent semantic structures within large textual corpora. By identifying recurring patterns of word co-occurrence, it reveals underlying themes that facilitate conceptual understanding of unstructured text data. Furthermore, topic modeling contributes to dimensionality reduction by representing documents in terms of a limited set of topics rather than thousands of individual terms, thereby simplifying computational processing and analysis. It also enhances interpretability, as the extracted topics correspond to coherent groupings of semantically related words that reflect meaningful subjects within the corpus. In addition, topic distributions derived from such models can be effectively integrated into downstream tasks, including document classification,

clustering, recommendation systems, and trend analysis, thereby improving overall analytical performance.

Traditional methods like Latent Dirichlet Allocation (LDA) use probabilistic frameworks to model documents as mixtures of topics, each represented by word distributions. Non-negative Matrix Factorization (NMF) offers an alternative by factorizing document-term matrices into additive, interpretable topic components. More recent approaches integrate contextual embed-dings from models like BERT to capture deeper semantic and contextual relationships, improving topic coherence and relevance, especially in complex or domain specific texts.

1.2 *LDA (Latent Dirichlet Allocation) and its methodology*

Latent Dirichlet Allocation (LDA) is a generative probabilistic model introduced by Blei et al. It is primarily used to discover the underlying thematic structure in a large collection of documents. In LDA, documents are represented as random mixtures over latent topics, where each topic is characterized by a word distribution. The model assumes that documents arise from a hierarchical Bayesian process: first, a distribution over topics is drawn for each document from a Dirichlet distribution, and then, words in the document are generated by selecting topics from this distribution and drawing words from those topics' respective distributions. This framework allows LDA to uncover the hidden thematic patterns in the data without supervision, enabling applications such as topic discovery, document classification, and information retrieval [1].

The Latent Dirichlet Allocation (LDA) model is grounded in a hierarchical Bayesian frame-work that conceptualizes documents as mixtures of latent topics. The generative process can be formally described as follows:

For each document d in a corpus of M documents:

1. Choose $N \sim \text{Poisson}(\xi)$: the number of words in the document.
2. Choose $\theta_d \sim \text{Dirichlet}(\alpha)$: the topic distribution for document d, where α is a K-dimensional parameter vector (with K being the number of topics).
3. For each of the N words w_n in document d:

(a) Choose a topic $z_n \sim \text{Multinomial}(\theta_d)$.

(b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability distribution conditioned on the topic z_n, where β is a $K \times V$ matrix (with V being the vocabulary size) and $\beta_{ij} = p(w^j = 1|z^i = 1)$.

The joint probability distribution for the observed words and latent topics can be expressed as:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{1}$$

where θ is the topic distribution for the document, $\mathbf{z}$ is the vector of topic assignments for each word, and $\mathbf{w}$ is the vector of observed words.

The marginal distribution of a document is obtained by integrating over θ and summing over $\mathbf{z}$:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{2}$$

The complete corpus likelihood is the product of the marginal probabilities of individual documents:

$$p(\mathcal{D}|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}=1}^{K} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d \tag{3}$$

As shown in Fig. 1, the corpus-level parameters α and β affect word and topic distributions, respectively; θ represents the topic distribution unique to a given document; z denotes a word instance's topic; and w represents the observed word.

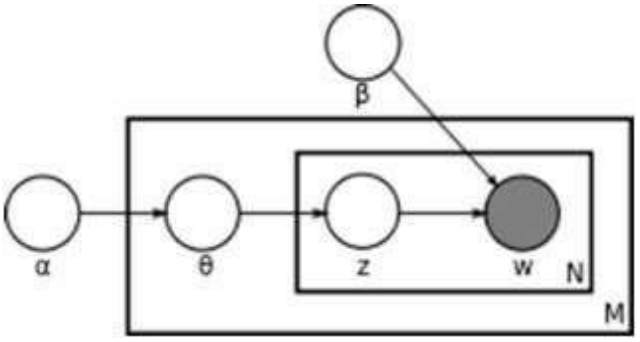


Figure 1: Traditional Generative Process of Topics and Documents in LDA

1.3 *Overview of NMF (Non-negative Matrix Factorization)*

Non-negative Matrix Factorization (NMF) is a prominent topic modeling technique that fac-torizes a non-negative input matrix—such as a document-term matrix—into two lower-rank non-negative matrices. This factorization approximates the original matrix as the product of these two, typically representing topics and their associated weights within documents [2, 3, 4]. Unlike probabilistic approaches, such as Latent Dirichlet Allocation (LDA), NMF employs a matrix decomposition approach that is directly interpretable due to its production of additive, part-based representations, making it suitable for extracting meaningful topics from text data [2].

Non-negative Matrix Factorization (NMF) approaches topic modeling through a linear algebraic lens by decomposing a non-negative document-term matrix into two lower-rank non-negative matrices. These matrices typically represent latent topics and their corresponding weights across documents.

Given a document-term matrix $V \in \mathbb{R}^{n \times m}$ (with n documents and m terms), NMF aims to find matrices $W \in \mathbb{R}^{n \times k}$ and $H \in \mathbb{R}^{k \times m}$ such that:

$$\geq 0 \qquad \geq 0$$

$$V \approx W \times H \tag{4}$$

where k is the number of topics (typically $k \ll \min(n,m)$), W contains document-topic associations, and H contains topic-term associations. The optimization problem can be formulated as:

$$\min_{W \geq 0, H \geq 0} \|V - WH\|_F^2 \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Unlike LDA's probabilistic foundation, NMF produces additive, part-based representations that are directly interpretable, as it models topics as additive combinations of words without requiring probabilistic assumptions. This characteristic makes NMF particularly effective for extracting meaningful and easily understandable

topics from textual data, offering an alternative perspective on topic structure to probabilistic methods like LDA.

Non-negative Matrix Factorization (NMF) operates iteratively to minimize the difference between the original document-term matrix and the product of two factor matrices, typically through alternating non-negative least squares (NLS) optimization. This process maintains non-negativity in the factor matrices, preserving interpretability by representing topic distributions and word contributions to topics [3, 4]. The resulting basis vectors capture underlying semantic patterns (topics), and the corresponding encodings indicate the strength with which each document expresses those topics [5].

With words arranged in rows and documents arranged in columns, let V be the input matrix to be factored. For example, consider 500 documents with 10,000 words indexed. Each document is represented by a column vector v in V. To create a features matrix W with 10,000 rows and 10 columns and a coefficients matrix H with 10 rows and 500 columns, identifying 10 topics. If the factorization is successful, the product W × H approximates the input matrix V:

$$V_{[10000\times500]} \approx W_{[10000\times10]} \times H_{[10\times500]} \quad (6)$$

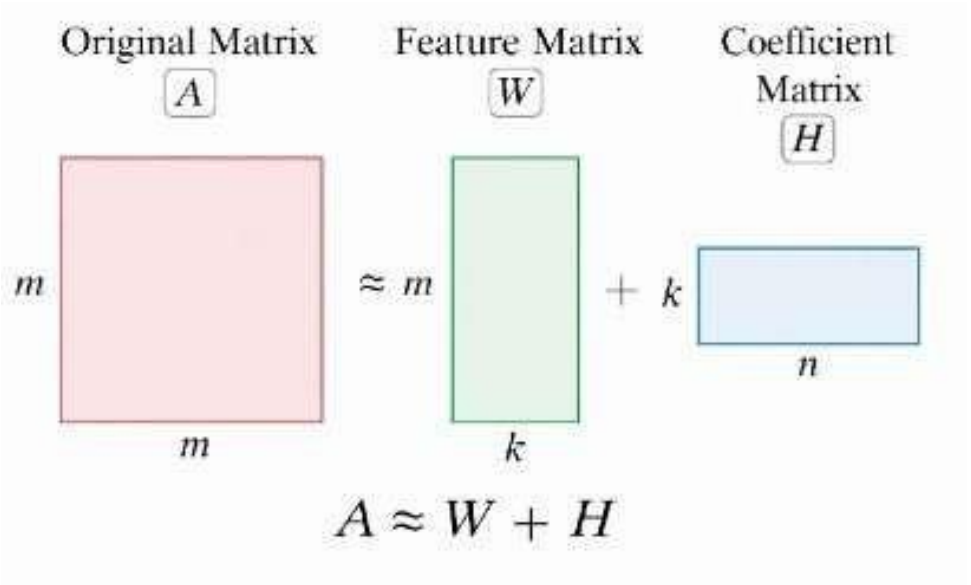


Figure 2: Matrix Factorization View of NMF Topic Modeling

Non-negative Matrix Factorization (NMF) has the advantage of generating sparse, part-based decompositions that are intuitively similar to topics formed of distinct word groups. NMF eliminates some of the independence requirements inherent in probabilistic models by directly factorizing word-document co-occurrence matrices, which relaxes conditional independence assumptions for words given themes [6]. Furthermore, NMF approaches can be made more resilient to noise and outliers by using regularized versions or robust loss functions [6].

Evaluating topic coherence is crucial for selecting optimal parameters for Non- negative Matrix Factorization (NMF), notably the number of topics. Comparative studies indicate that NMF often yields more coherent topics than Latent Dirichlet Allocation (LDA) on corpora, such as those containing COVID-19 documents, as measured by coherence metrics, such as c v [2]. Furthermore, initialization strategies for factor matrices significantly impact the quality of the solution due to the non-convexity of NMF optimization; several methods have been proposed to enhance convergence to meaningful local minima [7].

1.4 *BERT (Bidirectional Encoder Representations from Transformers)*

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model designed for natural language understanding tasks. Unlike traditional embeddings, which gen-erate fixed vectors for words irrespective of context, BERT produces contextual embeddings by considering the entire sentence bidirectionally—that is, it simultaneously analyzes both the preceding and following context. This allows BERT to capture nuanced meanings and disam-biguate polysemous words based on their sentence usage [8].

BERT is pretrained on large corpora using two unsupervised tasks: masked language mod-eling (predicting randomly masked words) and next-sentence prediction (understanding rela-tionships between sentences). After pretraining, BERT can be fine-tuned for specific down-stream tasks, such as question answering, sentiment analysis, or named entity recognition [8].

In topic modeling, BERT's contextual embeddings can generate dense, semantically rich representations of words, sentences, or documents. These embeddings offer more precise input features than traditional count-based vectors, such as document-term matrices used in NMF or LDA. Integrating BERT embeddings with topic modeling can improve the quality and coher-ence of extracted topics by incorporating deeper semantic and contextual information [9].

The BERT architecture is based on the Transformer encoder model. It consists of multiple layers of Transformer encoders that process input text bidirectionally, meaning it considers the context from both left and right of each token simultaneously. This bidirectional approach enables BERT to capture deeper semantic relationships within text [10].

Consider the sentence," The bank can guarantee deposits will be safe." The term "bank" in this context may refer to either a financial organization or the bank of a river. Conventional word embeddings allocate a static vector to "bank," irrespective of context, rendering them incapable of differentiating these meanings.

BERT, however, reads the entire sentence bidirectionally. It analyzes the words before and after "bank" to understand that in this sentence, "bank" refers to a financial institution. This contextual understanding is possible because BERT uses its transformer encoder layers to process the sentence in both directions at once. After pretraining, BERT can generate rich, context-aware embeddings for words or sentences, which can then be fine-tuned for various tasks like question answering or sentiment analysis.

1.5 *BERTopic*

Recently, Grootendorst proposed BERTopic, a modular topic modelling framework that leverages BERT embeddings, dimensionality reduction (UMAP), densitybased clustering (HDB-SCAN), and a novel class-based TF-IDF (c-TF-IDF) procedure to generate coherent and interpretable topics. BERTopic has quickly gained attention for its ability to produce state-of-the-art results on diverse datasets without requiring the number of topics to be specified a priori [11].

BERTopic is a topic modeling methodology that incorporates BERT embeddings to im-prove the quality and coherence of identified topics. In contrast to conventional topic models like LDA or NMF, which depend on document-term matrices or probabilistic frameworks, BERTopic utilizes BERT's sophisticated, contextual embeddings to elucidate more profound semantic and contextual connections within text. BERTopic is a modular topic modeling framework that consists of four stages:

1.5.1 *Document Embeddings*

Each document is mapped to a dense vector using a pre-trained Sentence-BERT (SBERT) model. SBERT fine-tunes BERT with a siamese network architecture to produce semantically meaningful sentence embeddings that are directly comparable via cosine

similarity. Given a document d, we obtain an embedding $\mathbf{e}_d \in R^{768}$ (for the base model).

1.5.2 *Dimensionality Reduction*

High-dimensional embeddings are noisy and computationally expensive to cluster. UMAP (Uniform Manifold Approximation and Projection) [12] is used to project $\mathbf{e}_d$ into a lower dimensional space $\mathbf{e}_{\mathbf{d}'} \in R^m$, typically m = 5. UMAP constructs a fuzzy simplicial com- plex from the high-dimensional data and then optimizes a low-dimensional analog to minimize cross-entropy between the two representations. It preserves both local and global structure more faithfully than PCA or t-SNE.

1.5.3 *Clustering*

To group semantically similar documents, BERTopic applies HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [12]. HDBSCAN extends DBSCAN by building a hierarchy of clusters and extracting the most persistent ones. It does not require a predetermined number of clusters and naturally identifies outliers (noise). Documents assigned to the same cluster form the basis of a candidate topic.

1.5.4 *Topic Representation with c-TF-IDF*

After clustering, each cluster c contains a set of documents. To extract the most representative words for the topic, BERTopic employs class-based TF-IDF (c-TF-IDF). For a term t and cluster c:

$$\text{c-TF-IDF}_{t,c} = \frac{f_{t,c}}{\sum_{t' \in c} f_{t',c}} \cdot \log\left(1 + \frac{N}{\sum_{c'} \mathbb{1}(t \in c')}\right) \tag{7}$$

where $f_{t,c}$ is the frequency of term t in all documents of cluster c, N is the total number of documents, $(\cdot)$ is the indicator function, and the denominator counts how many clusters contain term t. This formulation rewards terms that are frequent in a specific cluster but rare across clusters. The top n terms with the highest c-TF-IDF score form the topic representation. Because BERTopic does not assign each document exclusively to a single topic (documents may be noise), for evaluation we consider only documents with a non- negative cluster assignment.

2 *Dataset Description*

The Indian financial market is characterized by recurring periods of volatility, marked by frequent highs and lows. While numerical data and market indices support statistical analysis and are readily available, they do not fully capture the underlying factors driving market behavior. In this context, analyzing trends through financial news sources becomes essential for understanding market dynamics and predicting overall market sentiment.

To address this need, an exhaustive Indian Financial News Dataset has been compiled from Business Standard (www.business-standard.com), a well-established and reputable source of Indian financial journalism. The dataset focuses specifically on news related to the Indian economy and financial markets, covering a broad spectrum of topics such as stock market movements, banking and finance, corporate developments, government policies, macroeconomic indicators, and sectoral trends [13].

The dataset comprises 50,000 news headlines along with their corresponding descriptions, spanning a time period from 2003 to May 2020. This long temporal coverage enables both short-term and longitudinal analysis of financial trends and sentiment evolution in Indian markets. The dataset is designed to be updated on a monthly basis, ensuring continued relevance and applicability for ongoing research.

This dataset is suitable for various natural language processing and financial text analysis tasks, including sentiment analysis, market trend prediction, event detection, and

financial news classification. Its domain-specific focus and extensive coverage make it a valuable resource for studying the relationship between financial news and market behavior in the Indian context.

3 *Results and Findings*

To complement the quantitative metrics, we conducted a qualitative inspection of the ten largest topics produced by each model. This analysis serves two purposes: (i) to assess the inter-pretability and practical utility of the extracted topics, and (ii) to diagnose specific failure modes that are not captured by automated coherence scores. Table presents the top-10 words for five representative topics from each model, selected to illustrate typical patterns.

Fig. 3 presents a Top 10 Words comparison across LDA, NMF, and BERTopic for five extracted topics from the financial news dataset. The comparison highlights differences in semantic clarity, thematic consistency, and contextual depth among the three models.

For Topic 1, LDA produces relatively broad and frequency-driven terms such as *new, premium, raise, and insurer*, suggesting an insurance-related theme but lacking precise contextual structure. NMF shifts the focus toward monetary policy terms such as rbi, reserve, policy, and inflation, indicating improved thematic consolidation. In contrast, BERTopic generates highly domain-specific and context-rich keywords such as bank, hsbc, rbi, sbi, and lending, clearly identifying a structured banking theme. This demonstrates the ability of transformer-based embeddings to capture deeper contextual relationships.

Top 10 Words Comparison (LDA vs NMF vs BERTopic)

Topic ID	LDA Top Words	NMF Top Words	BERTopic Top Words
Topic 1	new, premium, raise, insurer, norm, month, sale, via, impact, nod	rbi, reserve, india, governor, policy, inflation, norm, today, liquidity, monetary	bank, banking, hsbc, rbi, banks, bankers, india, sbi, idbi, lending
Topic 2	loan, firm, hdfc, deal, call, come, point, repo, customer, even	net, profit, quarter, ended, reported, rise, posted, increase, december, income	insurer, premiums, insurance, pension, companies, irda, investm-ents, investment, premium, equity
Topic 3	insurance, net, lic, life, idbi, development, corporation, stake, week, financial	insurance, life, lic, insurer, irda, premium, company, new, general, regulatory	rupee, dollar, forex, currency, euro, currencies, dollars, indian, trading, rbi
Topic 4	likely, liquidity, banker, kotak, paper, system, soon, line, vysya, spread	rate, cut, lending, bps, point, basis, hike, base, repo, today	profit, crore, bank, canara, canbank, income, q1, india, q2, vijaya
Topic 5	bank, rbi, may, credit, today, cooperative, growth, policy, take, cover	rate, cut, lending, bps, point, basis, hike, base, repo, today	profit, crore, bank, canara, canbank, income, q1, india, q2, vijaya

Figure 3: Top 10 Words Comparison of LDA, NMF, and BERTopic embeddings to capture deeper contextual relationships.

In Topic 2, LDA exhibits a mixture of loan and corporate-related terms (loan, firm, repo), reflecting partial topic overlap. NMF improves clarity by modeling corporate earnings using terms such as net profit, quarter, and income. However, BERTopic distinctly identifies an insurance-investment theme through precise keywords including insurer, premiums, pension, irda, and equity, indicating stronger semantic alignment and reduced topic mixing.

For Topic 3, both LDA and NMF emphasize insurance-related vocabulary, though with some general and overlapping words. BERTopic, however, separates this into a well-defined foreign exchange theme using terms such as rupee, dollar, forex, euro, and trading. This clearer differentiation suggests that contextual embeddings enable better discrimination be-tween closely related financial subdomains.

In Topic 4, LDA generates moderately ambiguous financial terms such as liquidity and banker, while NMF more clearly captures monetary policy rate discussions (rate, cut, hike, repo). BERTopic, on the other hand, identifies a corporate earnings theme centered on specific banking entities (profit, crore, q1, q2, canara), reflecting its capacity to incorporate entity-level semantics.

Similarly, in Topic 5, LDA and NMF present partially overlapping banking and currency terms. BERTopic isolates a coherent banking-profit theme (axis, raises, q3, equity, sector), demonstrating stronger topic purity and interpretability.

Overall, the results indicate that LDA, due to its bag-of-words probabilistic framework, tends to generate broader and occasionally overlapping word distributions. NMF improves topic separation through matrix factorization but remains frequency-dependent. BERTopic consistently produces more semantically coherent, context-aware, and domain specific topics. Although BERTopic may involve higher computational complexity, it provides superior interpretability and clearer thematic boundaries in financial news analysis.

This comparison clearly demonstrates that BERTopic provides superior thematic clarity and interpretability for financial news analysis, while LDA and NMF remain useful but less context-sensitive approaches.

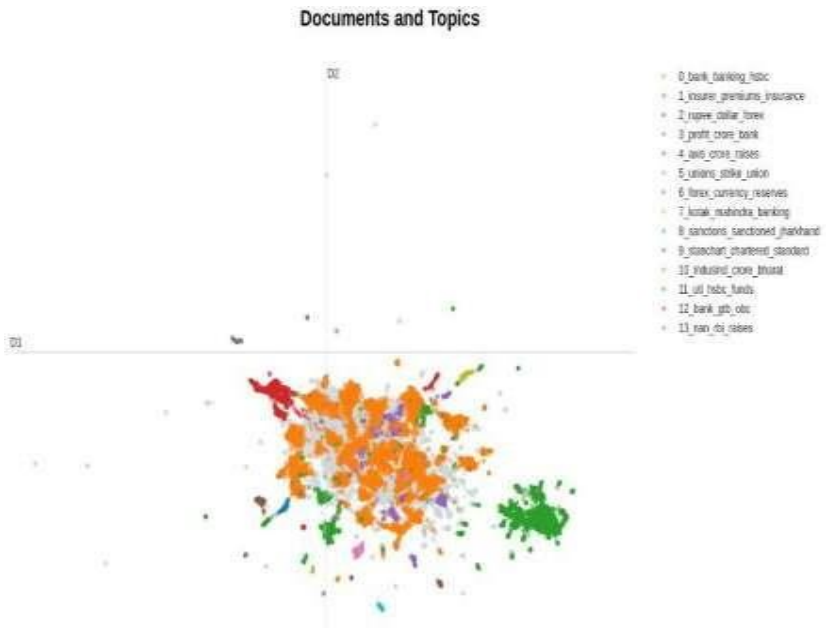


Figure 4: 2-dimensional UMAP projection generated by BERTopic

Fig. 4 represents the 2-dimensional UMAP projection generated by BERTopic for the fi-nancial news dataset. Each point corresponds to a news document, and colors represent distinct discovered topics. The axes (D1 and D2) denote reduced semantic dimensions obtained after embedding and dimensionality reduction.

The BERTopic visualization of the financial news dataset demonstrates strong semantic clustering and meaningful topic separation. Each point in the two-dimensional projection represents an individual news article, grouped based on contextual similarity derived from transformer-based embeddings. The figure shows several well-defined and compact clusters corresponding to key financial themes such as banking, insurance premiums, forex and currency reserves, RBI rate decisions, corporate profits, and industrial growth. Notably, certain clusters appear dense and clearly separated from the central region, indicating high intra-topic coherence and strong semantic similarity among documents within those themes. The presence of a moderately overlapping central region reflects the inherent interconnectedness of financial news, where topics such as banking, monetary policy, and market movements often co-occur. Additionally, a few scattered points suggest niche or less frequent financial events. Overall, the visualization confirms that BERTopic effectively captures contextual relationships within financial text data, producing coherent, interpretable, and domain-relevant topics.

Fig. 5 presents the comparative evaluation of topic modeling approaches—LDA, NMF, and BERTopic—using coherence score (C v) and topic diversity as performance metrics. The coherence results indicate that BERTopic achieves the highest coherence score (0.69), followed by NMF (0.65), while LDA records the lowest value (0.31).

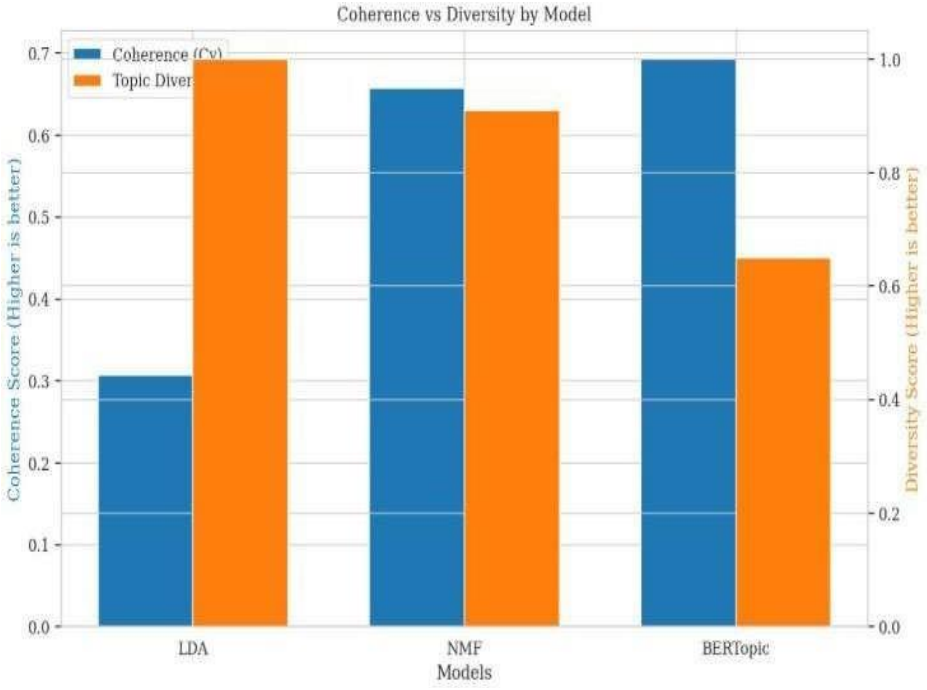


Figure 5: Comparative analysis of Coherence and Diversity Score by Model

Higher coherence reflects better semantic consistency among top words within each topic. The substantially lower coherence of LDA suggests weaker semantic grouping compared to embedding-based and matrix factorization approaches. The improved performance of BERTopic can be attributed to contextual embeddings, which capture semantic relationships more effectively than traditional probabilistic models.

In contrast, topic diversity exhibits an inverse trend. LDA attains the highest diversity score (1.0), followed by NMF (0.90), whereas BERTopic shows comparatively lower diversity (0.64). Higher diversity indicates reduced overlap among topic keywords, suggesting broader topical coverage. The results imply that while BERTopic produces

more semantically coherent topics, it may generate comparatively less diverse topic representations.

Overall, the findings reveal a trade-off between coherence and diversity across models. BERTopic demonstrates superior interpretability through higher coherence, whereas LDA pro-vides broader topical dispersion. NMF offers a balanced performance between the two metrics. These results suggest that embedding-based models are better suited for applications requiring semantically meaningful topics, while probabilistic models may be preferred when maximizing topic diversity is a priority.

The LDA model word cloud as indicated in Fig. 6, visually represents the most frequent and prominent words associated with topics generated by the Latent Dirichlet Allocation (LDA) model. In the context of the financial news dataset, the LDA word cloud tends to display relatively broad and frequency driven terms such as ”new,” ”premium,” ”raise,” and ”insurer.” These words suggest general thematic areas, for example, an insurance-related theme, but the word cloud may lack precise contextual structure or domain-specific clarity.
The NMF model word cloud visually represents the key terms associated with the topics extracted by Non-negative Matrix Factorization from the financial news dataset. The word clouds for each topic highlight prominent, domain-specific keywords that reflect distinct financial themes. For example, topics include terms such as ”reserve,” ”rbi,” ”policy,” and ”inflation” indicating monetary policy; ”net,” ”profit,” and ”quarter” reflecting corporate earnings;” insurance,” ”life,” and ”premium” representing insurance-related topics; and ”loan,” ”home,” and ”bad” pointing to credit and loan issues as shown in Fig. 7.

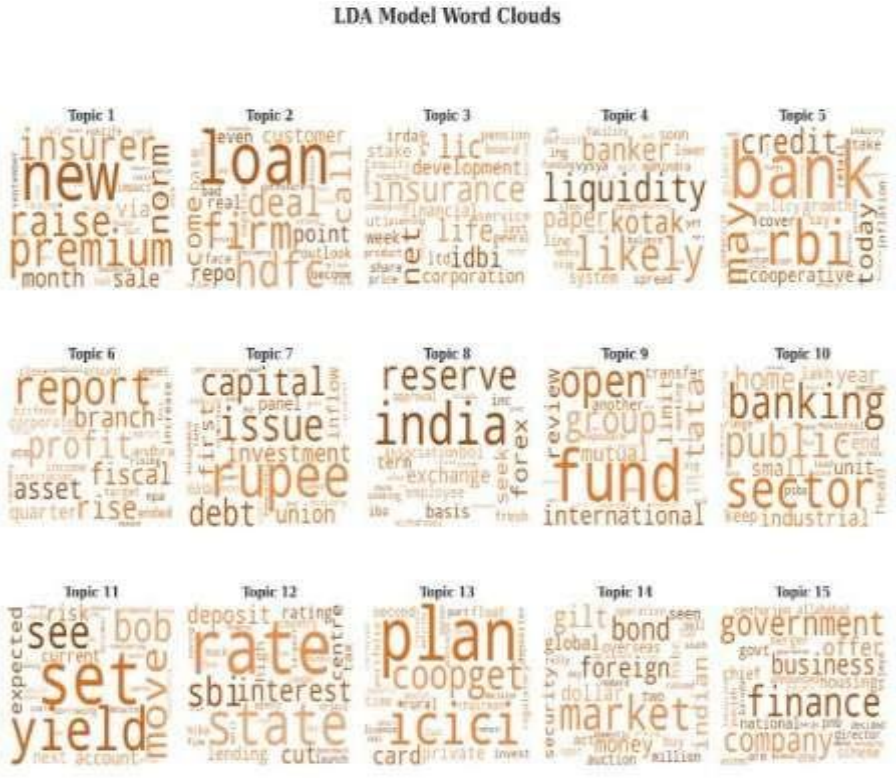


Figure 6: Word Cloud for LDA top words

Compared to LDA, the NMF word clouds show improved thematic consolidation and

clearer topic separation, focusing on more specific financial concepts rather than broad or overlapping terms. This is consistent with NMF's matrix factorization approach, which produces additive, part-based representations that align well with interpretable groups of words. The frequency-driven nature of the word clouds emphasizes terms that are not only common but also semantically coherent within each topic, supporting better interpretability in financial text analysis.

The BERTopic word cloud as indicated in Fig 8. visually represents the most prominent and contextually rich keywords associated with topics extracted using the BERTopic model from the financial news dataset. Compared to LDA and NMF word clouds, BERTopic's word cloud displays highly domain-specific and semantically coherent terms that reflect precise financial themes. For ex-ample, it highlights keywords such as "bank", "hsbc", "rbi", "sbi," and "lending" clearly identifying structured banking themes. This demonstrates BERTopic's strength in capturing deeper contextual relationships and producing topics with stronger thematic clarity and interpretability. The word cloud reflects the model's ability to disambiguate closely related financial subdomains and provide clearer thematic boundaries due to its use of transformer-based contextual embeddings.



Figure 7: Word Cloud for NMF top words

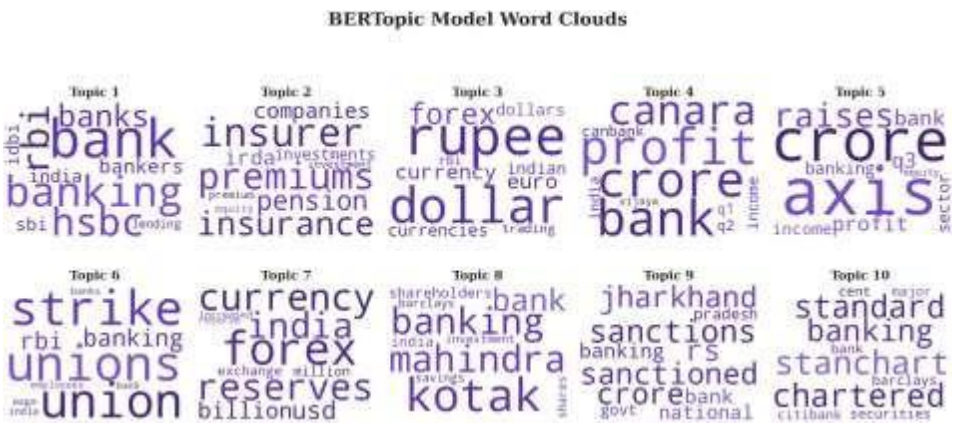


Figure 8: Word Cloud for BERTopic top words

4 *Conclusion*

This study will offer a comparative analysis in detail of three topic modeling techniques, that is, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) and BERTopic applied on a large scale Indian financial news dataset. The results show that as useful as the LDA and NMF are for providing thematic groupings, these are limited by more general and frequency-driven and sometimes overlapping topics. In contrast, BERTopic, using transformer-based contextual embeddings, generates highly domain-specific, semantically coherent, and contextually nuanced topics that are better suited to capture the complex financial subdomains within the data. The qualitative analysis of the word clouds and the topic coherence demonstrate that BERTopic can disambiguate the topics close to each other better, providing more explicit thematic boundaries and making its interpretation easier. Also, the UMAP-based visualization affirms the effectiveness of BERTopic in the creation of clear clusters that are related to major financial topics. These findings highlight the benefits of using a combination of contextual embeddings and advanced clustering methods with topic modeling in the financial text analysis domain. Future work may include the use of BERTopic on other domain-specific corpora, as well as its use in conjunction with downstream financial analytics tasks in order to take advantage of its interpretative power.

5 *Future Scope*

The promising results demonstrated by BERTopic in this study open several avenues for fu-ture research and practical applications. Firstly, extending the application of BERTopic to other domain-specific corpora, such as healthcare, legal, or scientific literature, could validate its versatility and robustness across diverse textual datasets. Secondly, integrating BERTopic with real-time financial analytics systems may enhance market sentiment analysis, event de-tection, and predictive modeling by providing dynamically updated and contextually rich topic insights. Thirdly, exploring hybrid models that combine BERTopic's contextual embeddings with other advanced natural language processing techniques, such as entity recognition or senti-ment analysis, could further improve topic interpretability and downstream task performance. Fourthly, optimizing the computational efficiency of BERTopic, particularly for large-scale streaming data, would facilitate its deployment in operational environments requiring

rapid topic extraction. Lastly, investigating the integration of BERTopic outputs with automated decision-making frameworks in finance could unlock new opportunities for informed invest-ment strategies, risk management, and policy analysis.

References

[1] H. Li, "Latent Dirichlet Allocation," Springer Nature Singapore, 2023, pp. 439–471. doi: 10.1007/978-981-99-3917-6 20.

[2] S. Mifrah, "Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Cōrpus," IJATCSE, vol. 9, no. 4, pp. 5756–5761, Aug. 2020, doi: 10.30534/ijatcse/2020/231942020.

[3] R. Kannan, G. Ballard, and H. Park, "A high-performance parallel algorithm for non-negative matrix factorization," SIGPLAN Not., vol. 51, no. 8, pp. 1–11, Feb. 2016, doi: 10.1145/3016078.2851152.

[4] R. Kannan, G. Ballard, and H. Park, "MPI-FAUN: An MPI-Based Framework for Alternating-Updating Nonnegative Matrix Factorization," IEEE Trans. Knowl. Data Eng., vol. 30, no. 3, pp. 544–558, Mar. 2018, doi: 10.1109/tkde.2017.2767592.

[5] M. W. Berry and M. Browne, "Email Surveillance Using Non-negative Matrix Factorization," Comput Math Organiz Theor, vol. 11, no. 3, pp. 249–264, Oct. 2005, doi: 10.1007/s10588-005-5380-5.

[6] M. Luo, F. Nie, X. Chang, Y. Yang, A. Hauptmann, and Q. Zheng, "Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling," AAAI, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.10832.

[7] S. Fathi Hafshejani and Z. Moaberfard, "Initialization for non-negative matrix factoriza-tion: a comprehensive review," Int J Data Sci Anal, vol. 16, no. 1, pp. 119–134, Nov. 2022, doi: 10.1007/s41060-022-00370-9.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Association for Computational Lin-guistics, Jan. 2019, pp. 4171–4186. doi: 10.18653/v1/n19-1423.

[9] A. C. Nanayakkara and G. A. D. M. Thennakoon, "Enhancing social media content analysis with advanced topic modeling techniques: A comparative study," International

Jour-nal on Advances in ICT for Emerging Countries, vol. 17, no. 1, pp. 40–47, May 2024, doi: 10.4038/icter.v17i1.7276.

[10] L. George and P. Sumathy, ”An integrated clustering and BERT framework for improved topic modeling,” International Journal of Information Technology, vol. 15, no. 4, pp. 2187–2195, Jan. 2023, doi: 10.1007/s41870-023-01268-w.

[11] M. Grootendorst, ”BERTopic: Neural topic modeling with a class-based TF-IDF proce-dure,” Cornell University, Mar. 11, 2022. doi: 10.48550/arxiv.2203.05794.

[12] J. Blanco-Portals, F. Peiró, and S. Estradé, ”Strategies for EELS data analysis: Introducing UMAP and HDBSCAN for dimensionality reduction and clustering,” Microscopy and Mi-croanalysis, vol. 28, no. 1, pp. 109–122, Nov. 2021, doi: 10.1017/S1431927621013696.

[13] K. A. Coulter, ”The impact of news media on Bitcoin prices: Modelling data-driven dis-courses in the crypto-economy with natural language processing,” Royal Society Open Science, vol. 9, no. 4, Apr. 2022, Art. no. 220276, doi: 10.1098/rsos.220276.