Evaluating Short Text Stream Clustering on Large E-commerce Datasets

Cesar Andrade[1,3] 0009-0005-2218-6648, Rita P. Ribeiro[1,3] 0000-0002-6852-8077, and João Gama[2,3] 0000-0003-3357-1195

[1] Department of Computer Science, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal up202101459@edu.fc.up.pt, rpribeiro@fc.up.pt
[2] Faculty of Economics, University of Porto, 4200-464 Porto. Portugal Faculty of Computer Science, University of Porto, Porto, Portugal
jgama@fep.up.pt
[3] INESC TEC, 4200-465 Porto, Portugal

Abstract. Latent Dirichlet Allocation (LDA) is a fundamental method for clustering short text streams. However, when applied to large datasets, it often faces significant challenges, and its performance is typically evaluated in domain-specific datasets such as news and tweets. This study aims to fill this gap by evaluating the effectiveness of short text clustering methods in a large and diverse e-commerce dataset. We specifically investigate how well these clustering algorithms adapt to the complex dynamics and larger scale of e-commerce text streams, which differ from their usual application domains. Our analysis focuses on the impact of high homogeneity scores on the reported Normalized Mutual Information (NMI) values. We particularly examine whether these scores are inflated due to the prevalence of single-element clusters. To address potential biases in clustering evaluation, we propose using the Akaike Information Criterion (AIC) as an alternative metric to reduce the formation of single-element clusters and provide a more balanced measure of clustering performance. We present new insights for applying short text clustering methodologies in real-world situations, especially in sectors like e-commerce, where text data volumes and dynamics present unique challenges.

Keywords: Short Text Stream Clustering · Latent Dirichlet Allocation · Online.

1 Introduction

Text clustering has become an essential technique in natural language processing (NLP), enabling the discovery of hidden structures within large collections of unstructured text data. Among various approaches, stream short text clustering, particularly methods based on Latent Dirichlet Allocation (LDA), has been widely researched due to its effectiveness in grouping similar texts and discovering topical patterns. While these methods have shown promising results in domains such as news feeds and social media platforms, their application to larger and more complex datasets remains less explored.

The performance of traditional LDA-based stream short text clustering methods is well-documented in scenarios involving relatively small datasets, with few exceptions [1]. However, these methods often struggle when scaled to larger datasets commonly found in industries like e-commerce. The dynamic and voluminous nature of e-commerce text data presents unique challenges not typically encountered in other domains such as news or social media. Handling the vast number of product descriptions, user reviews, and query logs requires clustering algorithms that are robust, scalable, and adaptive to the evolving nature of data.

A significant challenge in evaluating the performance of clustering methods is the reliance on metrics like Normalized Mutual Information (NMI). While NMI is useful for measuring the statistical accuracy of clustering results, it does not fully capture the practical aspects of the clusters formed, particularly in the presence of single-element clusters, which can artificially inflate homogeneity scores [2,3]. This limitation calls for a more comprehensive evaluation framework to better understand the practical utility of clustering outcomes.

Our study seeks to extend the application of these clustering methods to understand their scalability and adaptability better, particularly focusing on:

1. how these methods perform with extensive e-commerce datasets;
2. the clustering quality metrics, especially concerning homogeneity scores and single-element cluster formation;
3. the potential for alternative metrics like the Akaike Information Criterion (AIC) to provide more reliable evaluations of clustering outcomes.

This paper is organized as follows. Section 2 reviews related work on similarity-based and model-based stream clustering, detailing the online approaches. Section 3 presents our proposal and details the dataset, the dataset preparation, the evaluated methods, and the experimental setup. Section 4.2 reports the results. Section 5 discusses various aspects of the experiment. Finally, Section 6 concludes the paper and suggests future research directions.

2 Related Work

Two prominent areas of study emerge in clustering short text streams: similarity-based and model-based approaches. Similarity-based clustering relies on pairwise similarities, while model-based clustering relies on statistical models that define data distribution.

2.1 Similarity-based Clustering

Recent advances, incorporating pre-trained language models like BERT [4] coupled with clustering algorithms such as HDBSCAN [5], have showcased as promising in text clustering tasks. BERT generates contextual embeddings, allowing deeper context understanding and language structure modeling [6]. These embeddings serve as inputs for clustering algorithms. Studies [7,8] have demonstrated the superiority of combining BERT embeddings and HDBSCAN for short

text clustering over traditional methods. Additionally, ELINAC [9] introduced an auto-encoder-based method for clustering electronic invoices.

Despite BERT and transformers being a new trend, their use in short text stream clustering with LDA methods brings an additional computational cost [10], posing scale issues, particularly for large datasets.

2.2 Model-based Clustering

Approaches to model-based clustering of short text streams fall into two categories: batch methods and online methods. Batch methods handle data in separate blocks, whereas online methods continuously process new incoming data.

Batch approaches Latent Dirichlet Allocation (LDA)[11] is a model-based stream clustering technique that has inspired numerous extensions. These extensions address challenges related to topic evolution, semantic representation, and the dynamic nature of text streams. Initially, models like DCT[12] aimed to simplify topic assignments by assigning a single topic to each document. While effective for short texts, this approach lacked adaptability to handle changing topic counts within streams. MStream [13] emerged in response to this challenge, managing topic counts by discarding outdated batch documents and adapting to evolving topics. However, MStream's reliance on single-term document representation limited its ability to navigate semantic spaces, impacting cluster purity.

To enhance semantic representation, NPMM [14] utilized pre-learned word embeddings, advancing text semantics understanding. However, these embeddings' static, language-dependent nature constrained their adaptability to evolving semantic landscapes. Building on NPMM's limitations, DP-BMM [15] introduced unordered sequences of bi-terms for semantic representation, dynamically detecting topic counts. Nonetheless, this emphasis on semantic representation did not directly address challenges posed by sparse, high-dimensional data.

The introduction of DCSS [16] and FastStream [17] marked significant milestones. DCSS, utilizing the Dirichlet process, demonstrated superior stability and adaptability by autonomously learning topic counts and adjusting to topic drift. FastStream introduced efficiency and adaptability improvements, employing a novel cluster indexing mechanism and dynamic similarity thresholds to expedite processing. Rakib et al. [18] proposed an iterative classification method, enhancing short text clustering. However, a common limitation among these studies is that they are not focused on complex, large datasets, especially in e-commerce.

Online approaches The Online Semantic-enhanced Dirichlet Model (OSDM) [19] excels in dynamic clustering, integrating word-occurrence semantic data into a graphical model for real-time text processing without fixed cluster numbers or batch sizes. Tested on the News, Reuters, and Tweets datasets and their synthetic versions, OSDM uses metrics like Purity, V-Measure, Homogeneity, and Normalized Mutual Information (NMI) for evaluation, adding accuracy to its metric suite for a thorough analysis of performance.

Following in the footsteps of OSDM, the Online Semantic-enhanced Graphical Model (OSGM) [20] also employs a Poly Urn scheme to dynamically manage document clustering, focusing on semantic smoothing and term co-occurrence for enhanced clarity and ambiguity resolution. OSGM's evaluation mirrors OSDM's, utilizing the same datasets and metrics to ensure consistency in performance assessment and comparability between methods.

The introduction of EINDM [21], a context-enhanced Dirichlet model for online clustering in short text streams, adds another layer to the field. By using a window-based semantic term representation and new clustering metrics like word specificity, EINDM effectively captures the evolving nature of text data. Tested on datasets similar to OSDM and OSGM, EINDM uses the same metrics except for accuracy, focusing on core aspects of clustering effectiveness.

EStream [22] merges online and offline clustering techniques to offer a versatile solution to text stream clustering. Unlike the other methods, EStream expands its testing grounds to include additional datasets like NT and SO-T, along with News-T and Tweets-T. This approach allows for a broader validation of its effectiveness, using the same core metrics to ensure a comprehensive assessment of its performance across various text stream scenarios.

While these methods in clustering short text streams show innovation and effectiveness, particularly when evaluated with News and Tweets datasets. However, the focus on these limited domains raises questions about their adaptability and performance in broader contexts, such as larger datasets or specific fields like e-commerce. Exploring varied datasets could provide deeper insights into the scalability and applicability of these clustering techniques across diverse real-world scenarios.

The evaluation of clustering methods often relies on Normalized Mutual Information (NMI). However, NMI alone does not account for qualitative factors like the formation of very small or single-element clusters. A more nuanced evaluation considering cluster size distribution and the meaningfulness of clusters formed could lead to a more comprehensive understanding of clustering effectiveness, ensuring statistical performance and practical significance.

3 Our Methodology

Recent studies utilizing online LDA approaches for clustering short text streams have reported substantial performance achievements. However, these studies have predominantly focused on smaller datasets, mainly in specific domains such as News and Tweets. This focus has left a significant gap in our understanding of these methods' scalability and effectiveness across larger and more varied datasets, especially in dynamic sectors like e-commerce.

In light of these limitations, our study aims to:

1. Assess the method's effectiveness on an e-commerce dataset that varies significantly in size and number of clusters to determine adaptability and scalability. This evaluation will help us understand how well the method performs

when subjected to the complexities and varied dynamics of e-commerce texts compared to traditional short text streams.
2. Investigate the influence of high homogeneity scores on the reported NMI values, mainly focusing on whether the prevalence of single-element clusters inflates these scores. This aspect is crucial for ensuring that the NMI values reflect clustering effectiveness rather than artefacts of the clustering process.
3. Propose and test using the Akaike Information Criterion (AIC) as an alternative metric to mitigate the formation of single-element clusters. Including AIC is expected to provide a more robust evaluation framework that discourages overly simplistic clustering solutions and promotes more meaningful and practical clustering outcomes.

The methodology involves a systematic grid search across datasets ranging from 10,000 to 150,000 instances to assess how varying parameters impact performance. It will lay the groundwork for a detailed analysis, where each dataset size informs the parameter settings for the next, ensuring continuous refinement and optimization of both AIC and NMI. The AIC optimization aims to minimize the AIC value, while the NMI optimization aims to maximize the NMI score. This approach seeks to provide insights into the method's performance across various scenarios and improve clustering practices.

3.1 Datasets

Our experimental study employed a set of datasets based on the Brazilian NF-e Project, initiated in 2006, which revolutionized tax documentation by transitioning to an electronic system. This shift has generated a large volume of data, including comprehensive invoice details, making it a valuable resource for machine learning applications, such as detecting tax fraud.

We analyzed a subset of the NF-e dataset from November 2021, provided by the Amazonas State Department of Finance. The dataset contains several key features, including:

- **GTIN:** The GTIN is an identifier for trade items developed and controlled by GS1, formerly EAN/UCC. GTINs, formerly called EAN codes, are assigned to any item (product or service) that can be priced, ordered, or invoiced at any point in the supply chain.
- **NCM:** The NCM[4], which is a regional nomenclature for the categorization of goods adopted by since 1995, being used in all foreign trade operations of Mercosur countries. The NCM is based on the HS, a condensed expression of the "Harmonized Commodity Description and Coding System" maintained by the WCO. It was created to improve and facilitate international trade and its statistical control.
- **Product Description:** A free text field on the invoice with a brief product description.

[4] The Southern Common Market (MERCOSUR for its Spanish initials) is a regional integration process, initially established by Argentina, Brazil, Paraguay and Uruguay, and subsequently joined by Venezuela and Bolivia*.

GTIN	NCM (Full)	NCM (4-Digit)	Product Description
7894900011512	22030000	2203	Cerveja pilsen, lata 350ml
7894900011529	22030010	2203	Cerveja de trigo, garrafa 500ml

Table 1: Examples of Brazilian products with GTIN and NCM. NCM codes 22030000 (general beer packaging) and 22030010 (specific types like wheat beer). Both share NCM4 2203 ("beer"), but the additional digits provide specific details about type and packaging

3.2 Data Pre-processing

To better understand the intrinsic properties of our dataset, we have segmented the data into manageable blocks, increasing in increments of 10,000 up to a maximum of 150,000 instances. This segmentation is designed to facilitate analysis and ensure each subset remains a practical size for detailed cluster evaluation.

In our analysis, we distinguish between two coding formats used for dataset generation: NCM4 and NCM8. The NCM4 code signifies broader product categories using just the first four digits, providing a general view of item groups. In contrast, the NCM8 code utilizes all eight digits for a more detailed and specific classification, revealing finer distinctions in product types.

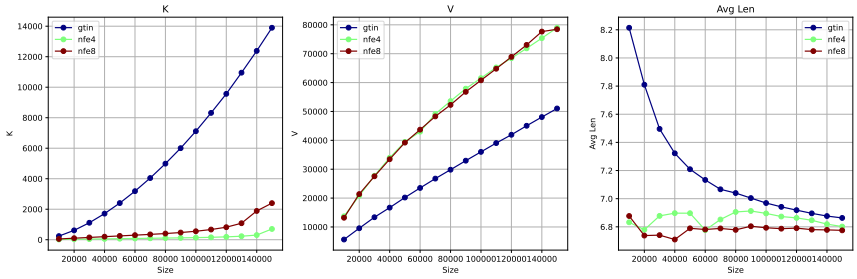

Fig. 1: Characteristics of the datasets: number of documents (D), number of clusters (K), vocabulary size (V), average length of the documents ($AvgLen$).

The behavior of the metric 'K', which represents the number of clusters, varies significantly across different data groupings. For NCM4, there is a gradual increase in 'K', marked by sporadic larger increases that hint at emerging complexities within the data as the volume grows. The NCM8 displays a more uniform rise in 'K', suggesting a linear enhancement in topic diversity corresponding to data size. Tailored to specific product descriptions, GTIN clusters show a steep upward trend in 'K'. This indicates a widening range of topics and reflects GTIN's detailed and expansive nature, where adding more product specifics leads to greater diversity in the identified topics.

Regarding the vocabulary size 'V', the GTIN dataset shows a pronounced growth, indicating an expanding lexicon that accommodates the increasing diversity of products. However, for NCM subsets—NCM4 and NCM8—the growth in 'V' is more measured, underscoring that while GTIN descriptions might capture a larger number of clusters due to unique product variations, a single NCM code can encompass a broad range of products, hence a larger vocabulary.

Finally, in our data preparation, we carefully control the average document length ('Avg Len') to ensure that clusters are not too small or too large, which could skew the analysis. The 'Avg Len' metric shows minor variations across GTIN and NCM datasets, with a slight decrease for GTIN, suggesting that product descriptions become more concise as the dataset grows. However, for NCM4 and NCM8, there are only slight fluctuations, reflecting the diverse complexity within their broader item categories as they expand. This control of 'Avg Len' is crucial for maintaining the integrity and comparability of clusters across different dataset sizes.

3.3 Clustering Methods

- **Online Semantic-enhanced Dirichlet Model (OSDM) [19]:** Features a dynamic clustering approach, integrating word-occurrence semantic information into a graphical model. It allows for real-time text processing without the need for predetermined cluster numbers or batch sizes, showcasing its adaptability to varying data streams.[5]
- **Online Semantic-enhanced Graphical Model (OSGM) [20]:** Uses a Poly Urn scheme for dynamic document clustering, focusing on semantic smoothing and term co-occurrence to resolve term ambiguity and enhance clarity, proving its strength in semantic-based text clustering.[6]
- **EINDM [21]:** A Dirichlet model using window-based semantic term representation to capture the evolving nature of text data through enhanced semantic understanding and novel clustering metrics.[7]
- **EStream [22]:** Employs both online and offline clustering to effectively adapt to various text streaming scenarios.[8]

3.4 Evaluation Metrics

To evaluate our experimental results, we resorted to metrics widely used in the literature [23] to provide a comprehensive assessment of clustering performance and model evaluation. The metrics are the following:

- **Homogeneity** (H) measures how well each cluster contains only members of a single class. It is defined as:

$$H = 1 - \frac{H(U|V)}{H(U)} \tag{1}$$

[5] https://github.com/JayKumarr/OSDM
[6] https://github.com/JayKumarr/OSGM
[7] https://github.com/JayKumarr/EINDM
[8] https://github.com/rashadulrakib/short-text-stream-clustering/tree/master/OnlineClustering

where $H(U|V)$ is the conditional entropy of the class distribution given the cluster assignments, and $H(U)$ is the entropy of the class distribution.

- **Completeness** (C) evaluates how well each class is represented within a single cluster. It is defined as:

$$C = 1 - \frac{H(V|U)}{H(V)} \tag{2}$$

where $H(V|U)$ is the conditional entropy of the cluster distribution given the class assignments, and $H(V)$ is the entropy of the cluster distribution.

- **Normalized Mutual Information** (NMI) combines homogeneity and completeness, measuring the similarity between the true class labels and the assigned cluster labels. It is defined as:

$$NMI = \frac{2 \cdot I(U;V)}{H(U) + H(V)} \tag{3}$$

where $I(U;V)$ is the mutual information between class and cluster distribution.

- **Purity** (P) assesses the clustering quality by calculating the proportion of the dominant class in each cluster. It is defined as:

$$P = \frac{1}{N} \sum_{k} \max_{j} |c_k \cap t_j| \tag{4}$$

where N is the total number of samples, c_k is the set of samples in cluster k, and t_j is the set of samples in class j.

- **Akaike Information Criterion** (AIC) helps balance model fit and complexity.
  It is defined as:

$$AIC = 2PTR - 2\ln(L) \tag{5}$$

where PTR represents the Predict/True Size Rate, and L is the likelihood of observing the Homogeneity value H given the mean (μ) and standard deviation (σ) of the H values across the datasets. The likelihood is computed using the normal distribution:

$$L = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(H-\mu)^2}{2\sigma^2}\right) \tag{6}$$

The Predict/True Size Rate (PTR) assesses how the optimization of NMI and AIC affects the ratio between the number of predicted clusters ($\hat{n}_c$) and the real number of clusters (n_c) and is determined by:

$$PTR = \left|\frac{\hat{n}_c}{n_c} - 1\right| \tag{7}$$

These metrics provide a balanced evaluation of clustering performance and model selection, ensuring robustness in our analysis.

4 Experimental Study

4.1 Experimental Setup

We evaluated various techniques for clustering product descriptions, focusing on methods based on Latent Dirichlet Allocation (LDA). For methods such as OSDM, OSGM, and EINDM, we carefully tuned the LDA parameters α (document-topic distribution) and β (topic-word distribution) through a grid search to optimize clustering performance. Additionally, for these methods, the decay rate λ was consistently set to 1e$-$6, with all other parameters kept at their default values.

ESTREAM requires only one parameter: Update-interval (UI). This parameter, set to 500 in our experiments, is essential for maintaining cluster quality by updating and removing outdated clusters. These parameter settings were crucial for achieving effective and stable clustering results.

4.2 Results

In this section, we present our findings through three primary graphs. The initial graph focuses on identifying the best parameter by illustrating the performance of metrics like Homogeneity, Completeness, NMI, Purity, and the cluster-to-label ratio. This investigation is crucial for determining the conditions under which the Akaike Information Criterion (AIC) is minimized, as shown in Figure 2. Our objective is to reveal how different settings affect clustering quality, mainly through the lens of AIC optimization.

The second graph illustrates the time evolution for four methods across varying dataset sizes, as depicted in Figure 3. Each method's processing time, represented in seconds, is plotted against the dataset size. The logarithmic scale on the y-axis facilitates comparing relative changes in processing time. This visualization offers insights into how each method's processing time scales with dataset size, providing valuable information on computational efficiency and scalability without delving into specific numeral evaluations.

The third graph demonstrates the evolution of the α and β parameters, as depicted in Figure 4. This analysis segment is essential for comprehending how NMI and AIC adjust to parameter changes and for identifying the configurations that lead to the best results. The study focuses on the methods EINDM, OSDM, and OSGM while omitting EStream since it does not involve parametric techniques. This focused examination aids in identifying the optimal parameters for enhanced clustering efficiency.

5 Discussion

AIC analysis highlights CB-TMCOH and TMCOH's superior efficiency compared to MSTREAM, indicating better optimization of cluster numbers without sacrificing model complexity. This underscores their ability to capture data

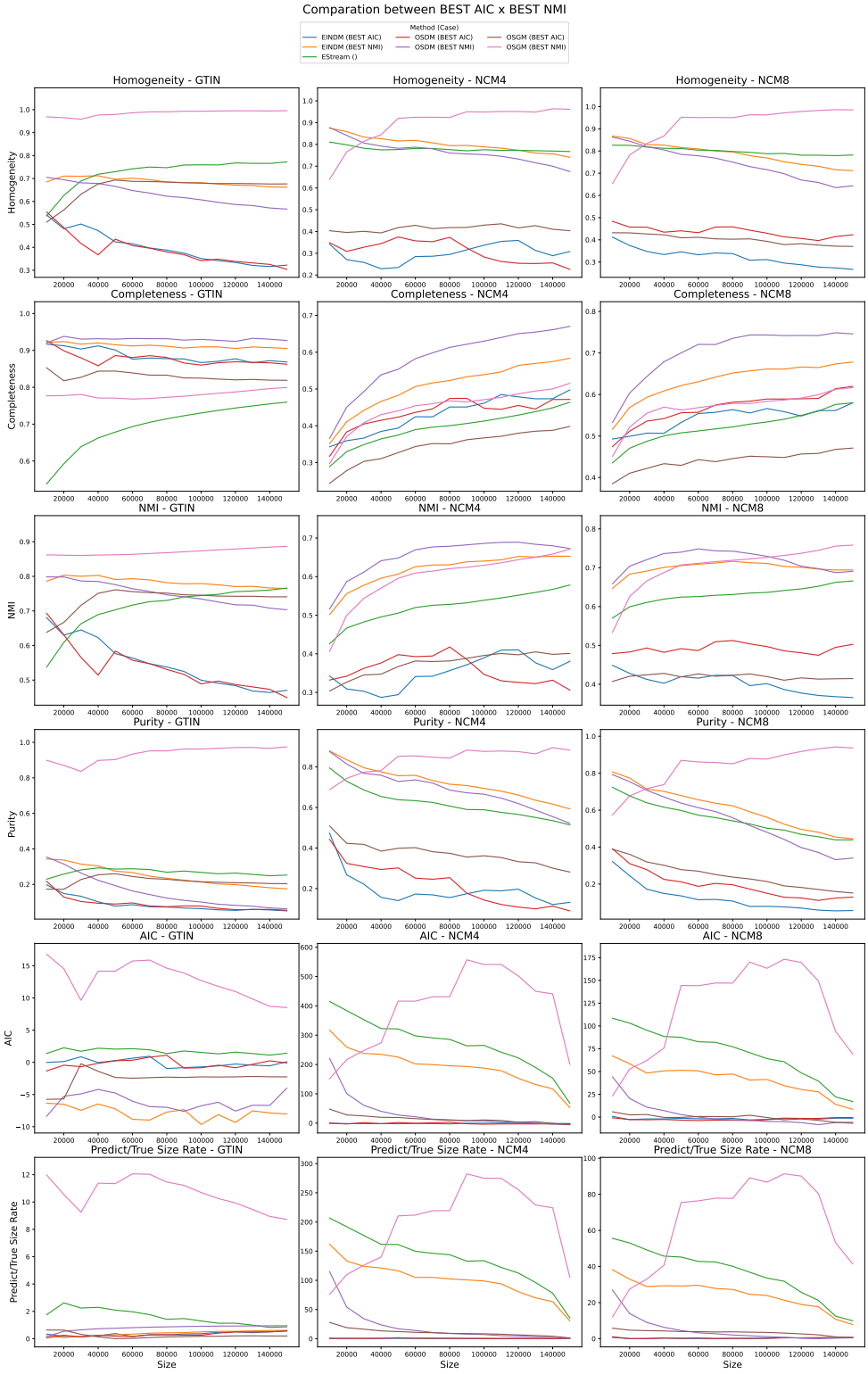

Fig. 2: Result of homogeneity, completeness, purity, NMI, predict/true size rate, and AIC. For GTIN, NCM4, and NCM8 datasets in BEST NMI analyses context.

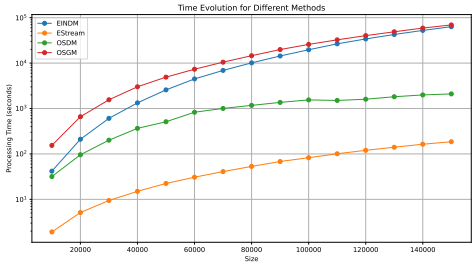

Fig. 3: Time evolution (processing time in seconds) versus the size of the GTIN dataset for four different methods: EINDM, OSGM, OSDM, and EStream.

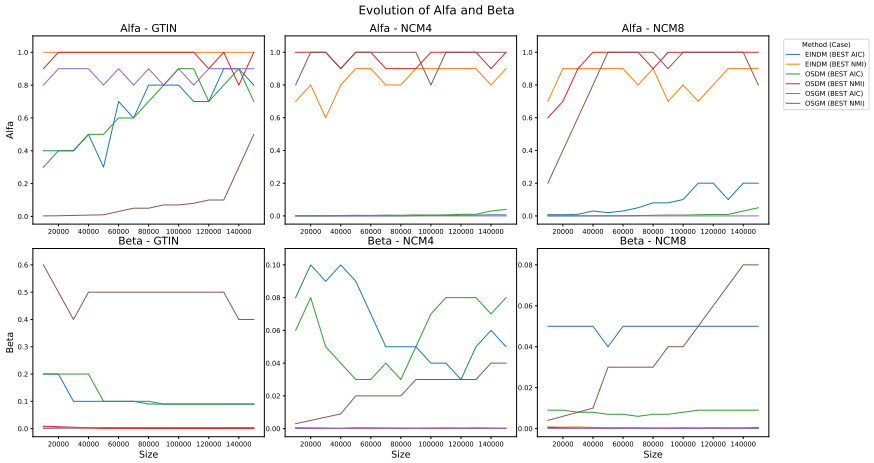

Fig. 4: α and β evolution for the methods EINDM, OSDM and OSGM.

structure effectively while avoiding overfitting. Additionally, CB-TMCOH outperforms existing LDA methods like TMCOH and FASTSTREAM, particularly regarding Completeness and Purity. This demonstrates its effectiveness in grouping similar elements while ensuring class exclusivity within clusters.

5.1 Metrics Evaluation

For datasets NCM4, NCM8, and GTIN, *Homogeneity* varies significantly. In NCM4, with fewer clusters, homogeneity is generally higher as clusters are likely more uniform but could be less comprehensive. In contrast, NCM8 and GTIN, with more complex and numerous clusters, show a potential decrease in homogeneity due to the wider variety of data points within each cluster. Comparatively, between the two cases (e.g., BEST AIC vs BEST NMI), BEST NMI

prioritizes achieving higher homogeneity as it directly influences the mutual information score, leading to better-aligning clusters with actual data classes.

NCM4's *Completeness* tends to be high due to simpler data structures, where capturing all class members in a single cluster is easier. However, in NCM8 and GTIN, increasing data complexity and cluster numbers can reduce completeness as classes may spread over multiple clusters. Between cases, BEST AIC might show reduced completeness compared to BEST NMI because AIC focuses more on model simplicity and goodness of fit rather than ensuring all class members are together, directly influencing NMI.

Purity is generally higher in datasets with fewer clusters like NCM4, where the likelihood of mixing different classes in one cluster is lower. In datasets like NCM8 and GTIN, despite a higher number of clusters, the complexity of data might lead to lower purity as clusters can contain a more diverse set of class labels. In terms of cases, BEST NMI likely shows better purity than BEST AIC, as maximizing mutual information would inadvertently align clusters more cleanly with true class labels.

NMI is particularly insightful for comparing clustering performance across datasets. NCM4 might show stable NMI values due to less complexity, while NCM8 and GTIN could exhibit more variability in NMI due to their increased complexity and cluster counts. Comparing cases, BEST NMI is explicitly optimized to maximize this metric, likely resulting in higher NMI scores across all datasets than BEST AIC, which might sacrifice some mutual information for simpler model structures.

The PTR measures the clustering algorithm's efficiency in controlling cluster sizes, ideally approaching 1. NCM4 usually shows higher PTR, indicating over-clustering. NCM8 and GTIN show lower PTR due to their complex structures, leading to many small clusters. BEST AIC, favoring simpler models, typically has PTR closer to 1 than BEST NMI. BEST NMI might increase the number of clusters to boost mutual information, resulting in lower PTR.

AIC values can help gauge the efficacy of clustering concerning model simplicity. Lower AIC values in NCM4 suggest a better fit with simpler models. In contrast, higher AIC in NCM8 and GTIN might indicate that the models, despite being more complex, are necessary to fit the data adequately. Comparatively, BEST AIC consistently focuses on minimizing AIC, potentially leading to simpler, more adaptable models across datasets than BEST NMI, which might accept more complex models if they yield better information sharing.

5.2 Assessment of Methods

EINDM is likely effective in datasets like GTIN, where data may be diverse and voluminous. EINDM can adapt to complex structures by combining multiple models to improve clustering accuracy. In NCM4, which may be simpler, the ensemble approach might not significantly outperform simpler methods but could still provide robustness against noise. In BEST AIC, EINDM might be configured between the two cases to limit the model complexity and avoid unnecessary computation. In contrast, in BEST NMI, the ensemble could be tuned

to maximize information captured from the data, potentially at the expense of increased computational load.

EStream tends to form more clusters than exist, performing better only than OSGM. It achieves worse results than the BEST NMI cases but better than the BEST AIC cases. However, it offers an excellent cost-benefit ratio as it is the most computationally efficient method, especially for the GTIN dataset.

ODSM achieves its best performance with smaller and simpler datasets like NCM4. As the complexity of the dataset increases, its performance degrades. ODSM does not tend to form many clusters for datasets with many classes, maintaining a Predict/True size rate close to ideal.

OSGM has the worst case for AIC values due to its tendency to form many clusters. Even in the BEST AIC case, this method generates more clusters than others. The best scenario to use OSGM is with the GTIN dataset and with large datasets. Between cases, BEST AIC would prefer setting stringent criteria for model updates to avoid complexity. At the same time, BEST NMI might leverage the flexibility of Gaussian mixtures to refine cluster definitions continuously, aiming for higher NMI scores.

Each method exhibits a distinct pattern of processing time as dataset size increases. EINDM demonstrates a notable increase in processing time with larger datasets, while OSGM and OSDM also show an upward trend, albeit with different slopes. In contrast, EStream exhibits significantly lower processing times across all dataset sizes, indicating its efficiency in handling larger datasets.

5.3 Hyperparameters

The parameter α shows a varying pattern across the dataset sizes for the EINDM method. At a size of 10,000, the α values are 0.4 and 1.0 for cases BEST AIC and BEST NMI, respectively. As the dataset size increases to 20,000 and 30,000, the α values remain consistent at 0.4 for the BEST AIC case. Still, the values adjust slightly to 1.0 for the BEST NMI case, indicating a stable but distinct configuration for optimal clustering performance. This evolution suggests that the α parameter is finely tuned to balance, preserving cluster structure and enhancing cluster homogeneity.

The parameter β exhibits a steady behaviour, particularly for the BEST AIC cases, where it remains constant at 0.200 for the initial dataset sizes of 10,000 and 20,000. Then, it reduces to 0.100 for a dataset size of 30,000. β values are significantly lower for the BEST NMI cases, starting at 0.006 and 0.004 for dataset sizes of 10,000 and 20,000, respectively. This trend continues as the dataset size increases, reflecting the parameter's role in optimizing cluster quality. The reduction in β values for the BEST NMI case suggests a focus on enhancing the granularity of clustering results to achieve a higher normalized mutual information score.

6 Conclusions

Our work demonstrates that LDA-based methods face significant difficulties when applied to large datasets. This is evidenced by the increasing processing time proportional to the dataset size, highlighting the scalability challenges inherent in traditional LDA approaches. It was confirmed that high NMI and homogeneity scores are often associated with forming single-element or small clusters. This is evident from the elevated Predict/True Size Rate observed. However, an exception was noted in the GTIN dataset, where methods such as OSDM, EINDM, and EStream did not conform to this premise, indicating some variability in performance across different datasets.

Evaluating clustering based on the best AIC demonstrated effectiveness in controlling single-element clusters. However, this came at the cost of reduced NMI, homogeneity, purity, and completeness. This trade-off suggests that while AIC can mitigate overly simplistic clustering solutions, it may also impact the overall quality of the clustering results.

Acknowledgments The authors wish to clarify that the first author received support from Amazonas State Government/Brazil for this research project.

References

1. X. Cheng, X. Yan, Y. Lan, and J. Guo: BTM: Topic modeling over short texts. In: *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928-2941, 2014. doi: 10.1109/TKDE.2014.2313872.
2. Pan Zhang. Evaluating accuracy of community detection using the relative normalized mutual information. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(11):P11006, 2015.
3. Maximilian Jerdee, Alec Kirkley, and MEJ Newman. Normalized mutual information is a biased measure for classification and community detection. *arXiv preprint arXiv:2307.01282*, 2023.
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
5. Campello, R. J. G. B., Moulavi, D., & Sander, J.: Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 160–172. Springer (2013)
6. Zhuyun Dai and Jamie Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, (2019).
7. Asyaky, M. S., & Mandala, R.: Improving the performance of HDBSCAN on short text clustering by using word embedding and UMAP. In: *8th Int. Conf. on Advanced Informatics: Concepts, Theory, and Applications (ICAICTA)*, pages 1–6. IEEE (2021)
8. Eklund, A., Forsman, M.: Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 635–643 (2022)

9. Schulte, J. P., Giuntini, F. T., Nobre, R. A., Nascimento, K. C. d., Meneguette, R. I., Li, W., Gonçalves, V. P., & Rocha Filho, G. P.: ELINAC: Autoencoder approach for electronic invoices data clustering. *Applied Sciences* **12**(6), 3008 (2022)
10. Andrade, C., Ribeiro, R. P., & Gama, J.: Topic Model with Contextual Outlier Handling: a Study on Electronic Invoice Product Descriptions. In: *EPIA Conference on Artificial Intelligence*, pages 365–377. Springer (2023)
11. Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. *Journal of Machine Learning Research* **3**(Jan), 993–1022 (2003)
12. Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas. Dynamic Clustering of Streaming Short Documents. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 995–1004, 2016.
13. Yin, J., Wang, J., Xu, W., & Gao, M.: Model-based clustering of short text streams. In: *27th ACM Int. Conf. on Information and Knowledge Management*, pages 697–706. ACM (2018)
14. Junyang Chen, Zhiguo Gong, and Weiwen Liu. A Nonparametric Model for Online Topic Discovery with Word Embeddings. *Information Sciences*, 504:32–47, 2019, Elsevier.
15. Junyang Chen, Zhiguo Gong, and Weiwen Liu. A Dirichlet Process Biterm-Based Mixture Model for Short Text Stream Clustering. *Applied Intelligence*, 50:1609–1619, 2020, Springer.
16. Xu, Y., Wang, S., Zhang, S., & Wang, F.: Dynamic clustering for short text stream based on Dirichlet process. *IEEE Access* **10**, 22852–22865 (2022)
17. Rakib, M. R. H., & Asaduzzaman, M.: Fast clustering of short text streams using efficient cluster indexing and dynamic similarity thresholds. *CoRR* **abs/2101.08595** (2021)
18. M. R. H. Rakib, N. Zeh, M. Jankowska, and E. Milios, "Enhancement of short text clustering by iterative classification," in *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25*, pages 105–117. Springer, 2020.
19. Kumar, J., Shao, J., Uddin, S., & Ali, W.: An online semantic-enhanced Dirichlet model for short text stream clustering. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 766–776. Association for Computational Linguistics (2020)
20. Jay Kumar, Salah Ud Din, Qinli Yang, Rajesh Kumar, and Junming Shao. An Online Semantic-Enhanced Graphical Model for Evolving Short Text Stream Clustering. *IEEE Transactions on Cybernetics*, 52(12):13809–13820, 2021.
21. J. Kumar, J. Shao, R. Kumar, S. Ud Din, C. B. Mawuli, and Q. Yang: A context-enhanced Dirichlet model for online clustering in short text streams. In: *Expert Systems with Applications*, vol. 228, p. 120262, 2023.
22. Md Rashadul Hasan Rakib, Norbert Zeh, and Evangelos Milios. Efficient clustering of short text streams using online-offline clustering. *Proceedings of the 21st ACM Symposium on Document Engineering*, pages 1–10, 2021.
23. Belal Abdullah Hezam Murshed, Suresha Mallappa, Jemal Abawajy, Mufeed Ahmed Naji Saif, Hasib Daowd Esmail Al-Ariki, and Hudhaifa Mohammed Abdulwahab. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*, 56(6):5133–5260, 2023.