Modeling and Semantic Clustering in Large-scale Text Data: A Review of Machine Learning Techniques and Applications

Danish Khan, Prof. Arpana Jaiswal

Abstract

With the exponential growth of textual data across diverse domains, the task of efficiently modelling and clustering large-scale text has emerged as a key challenge in natural language processing (NLP). Conventional text representation approaches, such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW), often fall short in capturing semantic nuances. This limitation has encouraged the adoption of more advanced techniques, including word embeddings (e.g., Word2Vec, GloVe) and transformer-based models like BERT and GPT. Similarly, traditional clustering algorithms such as K-Means and Hierarchical Clustering often struggle with the high dimensionality and sparsity inherent in text data. Consequently, models like Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and deep learning-based clustering frameworks have gained popularity. This review paper presents a comprehensive overview of recent machine learning-based text representation and semantic clustering techniques, examining their performance, scalability, and relevance across applications. It also outlines persisting challenges such as interpretability, noise handling, and computational overhead, while identifying potential research directions to enhance semantic clustering in large-scale text environments.

Keywords: Semantic Clustering, Text Representation, Word Embeddings, Transformer Models, Deep Learning in NLP, Text Mining.

1. Introduction

The growing availability of digital text data has posed both challenges and opportunities for machine learning and NLP practitioners. The capacity to organise and extract meaning from massive textual datasets is vital for various real-world applications such as document summarisation, sentiment analysis, information retrieval, and automated literature review. In this context, the modelling and semantic clustering of large-scale text data has gained significant research attention.

Traditionally, text was processed using techniques like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). While these methods are computationally simple and useful for basic tasks, they often fail to represent the semantic relationships between words, especially in unstructured and sparse data. As a result, more advanced representation techniques have been introduced, such as Word2Vec, GloVe, and FastText, followed by transformer-based models like BERT and GPT, which provide deeper and more contextual understanding of text.

Similarly, standard clustering methods like K-Means and Hierarchical Clustering, though widely used, are often inadequate for managing the complexity of textual data due to scalability issues and the curse of dimensionality. More robust models such as LDA, NMF, and various deep learning-based clustering techniques have shown promising results in identifying hidden patterns and grouping semantically similar texts.

2. Text Representation Techniques

Semantic clustering is primarily concerned with grouping text documents not only based on surface-level features but also based on their underlying meaning. This has become increasingly important in the era of big data, where textual information is abundant and varied. Several modern tools and frameworks have been developed to assist with large-scale semantic organisation of texts.

For instance, ASReview (Schoot et al., 2020) employs machine learning and active learning to facilitate systematic literature reviews. Similarly, the Empath tool (Fast et al., 2016) utilises deep learning to build and

validate lexical categories from a small set of seed terms. Both tools exemplify practical applications of semantic clustering in real-world scenarios.

2.1 Traditional Text Representation Techniques

Bag-of-Words (BoW):

This approach represents each document as a vector of word frequencies, disregarding grammar and word order. Despite its ease of implementation, BoW suffers from data sparsity and lacks the ability to capture semantic relationships.

Term Frequency–Inverse Document Frequency (TF-IDF):
TF-IDF improves upon BoW by giving higher importance to unique words within a document. Although it helps in identifying important terms, it still does not preserve the contextual or syntactic information of words within sentences.

2.2 Word Embedding-Based Methods

To overcome the shortcomings of traditional approaches, word embedding techniques have been developed to generate dense, low-dimensional representations that capture both syntactic and semantic features.

- Word2Vec:
  Developed by Google, Word2Vec employs two architectures—Continuous Bag-of-Words (CBOW) and Skip-Gram—to learn vector representations based on the surrounding words. It is efficient but does not handle out-of-vocabulary terms or word ambiguity effectively.
- GloVe (Global Vectors):
  GloVe constructs embeddings using matrix factorisation of co-occurrence matrices. It captures global statistical information but does not provide context-sensitive word meanings.

2.3 Transformer-Based Contextual Embeddings

Fixed word embeddings, although effective to some extent, fail to account for the dynamic nature of language where the same word may carry different meanings depending on the context. Transformer-based models address this limitation by learning contextual embeddings.

- BERT (Bidirectional Encoder Representations from Transformers):
  BERT processes text bidirectionally, understanding a word based on both its left and right context. It has achieved state-of-the-art performance in numerous NLP tasks including clustering, classification, and topic modelling.
- GPT (Generative Pre-trained Transformer):
  GPT uses a unidirectional transformer architecture, which makes it particularly effective for text generation. When combined with unsupervised learning techniques, GPT-based embeddings can also be utilised for semantic clustering.
- T5 (Text-to-Text Transfer Transformer):
  T5 offers a unified framework by converting all NLP tasks into a text-to-text format. This versatility makes it suitable for diverse tasks such as summarisation, question-answering, and clustering.

Table 1: Comparison of Text Representation Techniques

Technique	**Strengths**	**Limitations**
Bag of Words (BoW)	Simple, interpretable	Ignores word order and meaning, high sparsity
TF-IDF	Highlights important words, widely used	Does not capture semantic meaning
Word2Vec	Captures word similarity, efficient	Fixed word representations, lacks context
GloVe	Effective for capturing global co-occurrence	Limited contextual understanding
FastText	Handles rare words and misspellings well	Higher computational cost

BERT	Context-aware, bidirectional learning	Requires large computational resources
GPT	Strong generative capabilities	Unidirectional, context limitations
T5	Versatile across NLP tasks	Computationally expensive

3. Semantic Clustering Techniques

Semantic clustering refers to the process of grouping textual data based on its underlying meaning, rather than relying solely on surface-level or syntactic similarities. Unlike conventional keyword-based methods, semantic clustering takes into account the context and conceptual associations among words, phrases, and documents. As textual datasets become increasingly large and complex, more sophisticated machine learning techniques are required to enhance the accuracy and efficiency of clustering.

3.1 Traditional Clustering Approaches

Conventional clustering methods have long served as a foundation in text mining due to their ease of implementation and computational efficiency. However, these algorithms often face limitations when dealing with sparse, high-dimensional, and semantically complex text data.

3.1.1 K-Means Clustering

K-Means is a popular and widely adopted algorithm that divides data into *K* clusters based on similarity. It operates in an iterative manner through the following steps:

1. Random selection of *K* initial centroids.
2. Assigning each data point to the nearest centroid using similarity metrics such as cosine similarity or Euclidean distance.
3. Recomputing the centroid of each cluster.
4. Repeating steps 2 and 3 until the centroids stabilise.

Advantages:

- Scalable to large datasets.
- Simple and easy to implement.
- Efficient with structured and dense data.

3.1.2 Hierarchical Clustering

Hierarchical clustering forms a nested structure of clusters (called a dendrogram) by following one of two approaches:

- Agglomerative (bottom-up): Begins with individual points and merges them into clusters.
- Divisive (top-down): Starts with a single cluster that is split recursively.

Advantages:

- Does not require specifying the number of clusters in advance.
- Provides interpretable relationships between clusters.

3.1.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN clusters data by identifying dense regions and labelling low-density regions as noise. It uses two key parameters:

- ε (Epsilon): Radius for neighbourhood consideration.
- MinPts: Minimum number of points to define a dense region.

Advantages:

- Capable of detecting clusters with arbitrary shapes.
- Robust in handling noisy data.

3.2 Topic Modelling-Based Clustering

Topic modelling refers to techniques that uncover hidden thematic structures within documents. These methods are often used for clustering documents based on underlying topics rather than lexical similarity.

3.2.1 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that assumes:

- A document is composed of multiple latent topics.
- Each topic is represented as a distribution over words.

LDA assigns topic probabilities to words and documents, enabling cluster formation based on shared themes.

Advantages:

- Efficient in discovering abstract topics.
- Useful for exploratory data analysis.

3.2.2 Non-Negative Matrix Factorization (NMF)

NMF is an algebraic technique that decomposes a document-term matrix into two non-negative matrices:

- One links words to topics.
- The other links documents to topics.

Unlike LDA, NMF does not operate on a probabilistic framework.

Advantages:

- Generates interpretable and coherent topics.
- Comparatively faster to compute.

3.2.3 BERTopic

BERTopic is a contemporary approach that combines **transformer-based embeddings** with **HDBSCAN**, a hierarchical density-based clustering algorithm. It uses contextual embeddings (e.g., BERT or Sentence-BERT) to group semantically similar documents.

Advantages:

- Automatically determines the number of clusters.
- Offers better coherence and accuracy compared to traditional topic models.

3.3 Deep Learning-Based Clustering

Deep learning approaches have significantly advanced semantic clustering by enabling models to learn rich, hierarchical, and context-aware representations of text.

3.3.1 Autoencoder-Based Clustering

Autoencoders are neural architectures that compress data into a lower-dimensional latent space and then reconstruct the input from this compressed form. Once trained, clustering can be applied to the encoded representations.

Process:

1. **Encoder:** Compresses high-dimensional embeddings.
2. **Decoder:** Attempts to reconstruct original inputs.
3. **Clustering:** Performed on latent representations using algorithms like K-Means.

Advantages:

- Reduces dimensionality while preserving structure.
- Discovers hidden patterns in text.

3.3.2 Transformer-Based Clustering

Models like **BERT** and **Sentence-BERT (SBERT)** are commonly used to generate high-quality embeddings suitable for clustering tasks. SBERT modifies BERT using Siamese architecture to produce sentence-level vector representations.

Advantages:

- Captures deep contextual semantics.
- Scales effectively to large datasets.

3.3.3 Contrastive and Self-Supervised Learning

Recent research has explored self-supervised approaches, particularly **contrastive learning**, where models learn to differentiate between similar and dissimilar text pairs.

Examples include:

- **SimCSE:** Trains models using contrastive loss for sentence embeddings.
- **DeepCluster:** Simultaneously learns feature representations and cluster assignments.

Advantages:

- Does not require labelled data.
- Delivers high-quality clustering in large-scale settings.

Table 2: Comparison of Clustering Techniques

Clustering Method	Strengths	Limitations
K-Means	Fast, scalable	Requires predefined clusters, struggles with high-dimensional data
Hierarchical	Interpretable, no need for predefined clusters	Computationally expensive
DBSCAN	Handles noise, detects arbitrarily shaped clusters	Struggles with varying density
LDA	Identifies latent topics	Requires predefined topic count
NMF	Efficient, produces coherent topics	Sensitive to hyperparameter tuning
BERTopic	Captures deep semantic structures	Computationally intensive
Autoencoders	Learns compact representations	Requires large training data
BERT/SBERT	High accuracy in semantic clustering	Computationally expensive
Contrastive Learning	No labeled data required	Requires large-scale datasets

3.5 Selecting the Right Clustering Technique

The choice of clustering technique depends on:

- Dataset size: Large datasets benefit from transformer-based methods.
- Computational resources: Traditional methods like K-Means work well with limited resources.
- Semantic complexity: Deep learning-based approaches provide better results for complex semantic relationships.

4. Evaluation Metrics and Benchmarking for Semantic Clustering

Evaluating the performance of semantic clustering techniques is crucial to ensure the effectiveness and reliability of clustering models in large-scale text data applications. Unlike traditional clustering, which relies heavily on numerical distance metrics, semantic clustering requires specialized evaluation metrics that assess the coherence, separation, and real-world relevance of clusters.

4.1 Intrinsic Evaluation Metrics

Intrinsic evaluation assesses clustering quality without external reference labels. It primarily focuses on cluster compactness, cohesion, and separation.

4.1.1 Silhouette Score

The Silhouette Score measures how well samples are clustered by comparing intra-cluster similarity to inter-cluster separation. It is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where:

- a(i) = Average intra-cluster distance (distance to other points in the same cluster).
- b(i) = Average nearest-cluster distance (distance to points in the closest other cluster).
- S(i) ranges from **-1 to 1**, where higher values indicate better clustering.

4.1.2 Davies-Bouldin Index (DBI)

DBI measures the similarity between clusters by computing the ratio of intra-cluster scatter to inter-cluster separation. A lower DBI indicates better clustering.

$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\frac{\sigma_i + \sigma_j}{d_{ij}}$$

Where:

- σ_i and σ_j are the average distances of points in clusters i and j to their respective centroids.
- d_{ij} is the distance between cluster centroids i and j.

4.1.3 Calinski-Harabasz Index (CHI)

CHI measures clustering compactness and separation, where higher values indicate better-defined clusters. It is computed as:

$$CHI = \frac{T_r(B_k)}{T_r(W_k)} \times \frac{N-K}{K-1}$$

Where:
Tr (B_k)) = Between-cluster dispersion.
Tr (W_k) = Within-cluster dispersion.
K = Number of clusters.

4.2 Extrinsic Evaluation Metrics

Extrinsic evaluation compares clustering results against a ground-truth labeled dataset.

4.2.1 Adjusted Rand Index (ARI)

ARI measures the agreement between predicted clusters and true labels, correcting for randomness. It is computed as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}/\binom{N}{2}\right]}{0.5\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}/\binom{N}{2}\right]}$$

Where: n_{ij} is the number of common elements between predicted and actual clusters. a_i and b_j are sums over cluster elements.

4.2.2 Normalized Mutual Information (NMI)

NMI measures the shared information between predicted and actual clusters. It is given by:

$$NMI = \frac{2 \times I(X,Y)}{H(X) + H(Y)}$$

Where: I(X, Y) = Mutual information between clusters.

- H(X)H(Y) = Entropies of the cluster distributions.

Strength: Handles varying cluster numbers well.
Limitation: Assumes discrete labels.

4.2.3 Fowlkes-Mallows Index (FMI)

FMI measures the similarity between predicted and true clusters using precision and recall:

$$FM\ I = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}$$

Where:

- TP, FP, and FN represent true positives, false positives, and false negatives.

4.3 Benchmark Datasets for Evaluating Semantic Clustering

To conduct fair and comparative assessments of semantic clustering techniques, researchers typically rely on standard benchmark datasets that come with pre-defined ground-truth labels. These datasets are instrumental in validating the effectiveness, accuracy, and scalability of clustering models.

- **20 Newsgroups Dataset:** Consists of over 18,000 newsgroup documents

classified into 20 categories. Widely used for evaluating unsupervised topic modelling and text categorisation tasks.

- **Reuters-21578 Dataset:** Contains 21,578 newswire articles tagged with 135 economic and financial topics. Particularly suitable for evaluating hierarchical and multi-label clustering methods.
- **Wikipedia Concept Dataset:** Comprises structured textual data extracted from Wikipedia pages, grouped according to semantic concepts. Ideal for benchmarking deep learning-based clustering models.
- **Amazon Product Reviews Dataset:** A large-scale corpus featuring millions of product reviews across multiple categories. Commonly used for assessing sentiment-based semantic clustering and recommendation systems.
- **SQuAD (Stanford Question Answering Dataset):** Encompasses over 100,000 question-answer pairs derived from Wikipedia. Useful in testing models that require contextual and semantic understanding for clustering question–answer texts.

5. Challenges and Identified Knowledge Gaps

Despite commendable progress in the field of semantic clustering and text modelling, several open challenges continue to hinder its widespread applicability:

- **Multilingual and Cross-Domain Data Quality:** Existing multilingual corpora often suffer from inconsistent formatting, translation errors, or limited language coverage. Studies such as Caswell et al. (2021) have drawn attention to these quality issues in large-scale multilingual datasets, which can adversely affect clustering outcomes.
- **Domain-Specific Limitations:** Much of the current research is tailored to specific domains, such as biomedical or genomic data (e.g., Babelomics). However, techniques that perform well on structured data may not generalise effectively to unstructured text.
- **Inadequate Evaluation Frameworks:** The need for domain-adaptable evaluation systems—particularly in summarisation tasks using datasets like WikiHow—underscores the limitations of existing metrics. Improved tools are required to assess clustering relevance in such scenarios accurately.

6. Conclusion

This review paper presents a comprehensive overview of recent advances in text modelling and semantic clustering, emphasising their applications across various domains and use cases. The shift from traditional representations to contextual embeddings, and from simple clustering algorithms to deep learning-based frameworks, marks a significant transformation in the field.

However, challenges related to dataset quality, scalability, evaluation, and multilingual compatibility continue to persist. Addressing these concerns through innovative models, refined evaluation techniques, and inclusive datasets will play a pivotal role in enhancing the reliability and utility of semantic clustering systems in large-scale text processing.

References:

[1] Dou, Wenwen., Wang, Xiaoyu., Skau, Drew., Ribarsky, W.., & Zhou, Michelle X.. (2012). LeadLine: Interactive visual analysis of text data through event identification and exploration. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), 93-102. http://doi.org/10.1109/VAST.2012.6400485

[2] Shokri, R.., & Shmatikov, Vitaly. (2015). Privacy-preserving deep learning. 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 909-910. http://doi.org/10.1145/2810103.2813687

[3] Medina, Ignacio., Carbonell, J.., Pulido, Luis., Madeira, S.., Götz, Stefan., Conesa, A.., Tárraga, Joaquín., Pascual-Montano, A.., Nogales-Cadenas, Rubén., Santoyo, J.., García-García, F.., Marbà, Martina., Montaner, D.., & Dopazo, J.. (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Research, 38 , W210 - W213. http://doi.org/10.1093/nar/gkq388

[4] Borisyuk, Fedor., Gordo, Albert., & Sivakumar, V.. (2018). Rosetta: Large Scale System for Text Detection and Recognition in Images. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. http://doi.org/10.1145/3219819.3219861

[5] Fast, Ethan., Chen, Binbin., & Bernstein, Michael S.. (2016). Empath: Understanding Topic Signals in Large-Scale Text. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. http://doi.org/10.1145/2858036.2858535

[6] Nickel, Maximilian., Murphy, K.., Tresp, Volker., & Gabrilovich, E.. (2015). A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE, 104, 11-33. http://doi.org/10.1109/JPROC.2015.2483592

[7] Schoot, R.., Bruin, J. D.., Schram, Raoul., Zahedi, Parisa., Boer, J. D.., Weijdema, F.., Kramer, Bianca., Huijts, M.., Hoogerwerf, M.., Ferdinands, Gerbrich., Harkema, Albert., Willemsen, Joukje E.., Ma, Yongchao., Fang, Qixiang., Hindriks, Sybren., Tummers, L.., & Oberski, D.. (2020). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence , 3 , 125 - 133 . http://doi.org/10.1038/s42256-020-00287-7

[8] Glaz, A. Le., Haralambous, Y.., Kim-Dufor, Deok-Hee., Lenca, P.., Billot, R.., Ryan, Taylor C.., Marsh, Jonathan J.., Devylder, J.., Walter, M.., Berrouiguet, S.., & Lemey, C.. (2019). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. Journal of Medical Internet Researc , 23 . http://doi.org/10.2196/15708

[9] Yoo, Kang Min., Park, Dongju., Kang, Jaewook., Lee, Sang-Woo., & Park, Woomyeong. (2021). GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation., 2225-2239. http://doi.org/10.18653/v1/2021.findings-emnlp.192

[10] https://www.semanticscholar.org/paper/afc2850945a871e72c245818f9bc141bd659b453

[11] Absalom, Ezugwu E.., Ikotun, A. M.., Oyelade, O. N.., Abualigah, L.., Agushaka, J.., Eke, C.., & Akinyelu, A. A.. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Eng. Appl. Artif. Intell. , 110 , 104743 . http://doi.org/10.1016/j.engappai.2022.104743

[12] Schuhmann, Christoph., Beaumont, Romain., Vencu, Richard., Gordon, Cade., Wightman, Ross., Cherti, Mehdi., Coombes, Theo., Katta, Aarush., Mullis, Clayton., Wortsman, Mitchell., Schramowski, P.., Kundurthy, Srivatsa., Crowson, Katherine., Schmidt, Ludwig., Kaczmarczyk, R.., & Jitsev, J.. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. ArXiv, abs/2210.08402. http://doi.org/10.48550/arXiv.2210.08402

[13] Albalawi, Rania., Yeap, T.., & Benyoucef, Morad. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Frontiers in Artificial Intelligence, 3. http://doi.org/10.3389/frai.2020.00042

[14] Althoff, Tim., Clark, Kevin., & Leskovec, J.. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. Transactions of the Association for Computational Linguistics, 4, 463 - 476. http://doi.org/10.1162/tacl_a_00111

[15] Zhou, Xiaokang., Li, Y.., & Liang, Wei. (2020). CNN-RNN Based Intelligent Recommendation for Online Medical Pre-Diagnosis Support. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18, 912-921. http://doi.org/10.1109/TCBB.2020.2994780