RESEARCH ARTICLE

UNSUPERVISED EMAIL LABELLING USING SEMANTIC EMBEDDINGS AND CLUSTERING ALGORITHMS

Sanjana Bhangale

Pillai College of Arts, Commerce & Science (Autonomous), New Panvel

Corresponding author E-mail: sanjanabhangale@mes.ac.in

DOI: https://doi.org/10.5281/zenodo.18069876

Abstract:

Email overload has become a significant barrier to productivity, as manual categorization is slow, inconsistent, and unsuitable for large-scale communication environments. This paper presents an unsupervised learning framework for automatically grouping emails into meaningful topic clusters without relying on labeled training sets. Classical clustering methods—K-Means, Hierarchical Clustering, and DBSCAN—are compared against modern semantic vectorization techniques using BERT and Sentence-BERT. Experimental results show that transformer-based embeddings combined with density-based clustering produce substantially higher cluster coherence than TF-IDF with K-Means. The proposed method demonstrates potential for intelligent inbox organization, automated routing, and enterprise-level workflow optimization.

Keywords: Email Clustering, Unsupervised Learning, BERT, Sentence-BERT, DBSCAN, Semantic Embeddings, Smart Inbox Automation.

1. Introduction:

Email remains a primary communication medium for both individuals and organizations. Despite the rise of instant messaging and collaborative platforms, email usage continues to grow, resulting in challenges related to information overload. Users frequently struggle to manage large volumes of messages, organize relevant content, and retrieve important information effectively. Manual email labeling—still the most common method for inbox organization—requires substantial time and effort and is highly susceptible to inconsistency and human error.

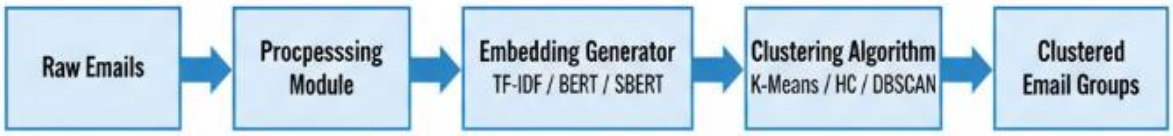
Automated email classification has traditionally relied on supervised machine learning techniques, which require large labeled datasets. However, email corpora are often private, heterogeneous, and costly to annotate. This makes supervised approaches difficult to scale and implement in real-world enterprise systems, where privacy and variability are major considerations.

Unsupervised learning offers a promising alternative by eliminating the need for labeled training data. Modern transformer-based models, such as BERT and Sentence-BERT, provide dense semantic representations that can capture nuanced meaning in email content. When combined with

clustering algorithms, these representations allow for the automatic discovery of coherent topic groups within large unlabeled email sets.

This research explores an unsupervised pipeline that integrates semantic embedding techniques with clustering algorithms. We evaluate both traditional and modern approaches, compare their performance in clustering quality, and assess their suitability for real-world email management applications. The primary contributions of this study are:

i. A comparative evaluation of classical text vectorization (TF-IDF) and transformer-based semantic embeddings for email clustering.
ii. A performance comparison across multiple clustering algorithms, including K-Means, Hierarchical Clustering, and DBSCAN.
iii. An analysis of cluster coherence using intrinsic evaluation metrics.
iv. A discussion of practical applications in smart inbox automation and enterprise communication tools.



Email Clustering System Block Diagram

2. Literature Review / Related Work

2.1 Email Classification and Management

Email classification research spans traditional machine learning, rule-based filtering, and modern AI-driven techniques. Early approaches relied on keyword-based rules or Boolean filters, which were brittle and required manual maintenance. Later work explored supervised learning methods such as Naïve Bayes, SVMs, and Random Forests for categorizing emails into predefined classes (e.g., spam vs. non-spam). While effective for well-defined tasks such as spam filtering, supervised models struggle in scenarios requiring topic discovery or personalized organization.

2.2 Unsupervised Text Clustering

Unsupervised text clustering techniques aim to group documents based on similarity without predefined labels. Classical methods typically utilize TF-IDF features combined with K-Means or Hierarchical Clustering. Although computationally efficient, TF-IDF suffers from sparsity and inability to capture semantic relationships. This limits its performance for email content, which often includes short messages, informal language, or context-dependent meaning.

Density-based clustering methods, notably DBSCAN, have gained popularity for text applications due to their ability to detect non-linear cluster shapes and identify noise points. However, their performance strongly depends on the feature representation used.

2.3 Transformer-Based Text Embeddings

Transformers revolutionized natural language processing by enabling contextualized language understanding. BERT and Sentence-BERT (SBERT) provide high-quality sentence embeddings that preserve semantic similarity, outperforming traditional bag-of-words models. SBERT, in particular, is optimized for semantic similarity tasks and allows efficient computation of sentence-level vectors. Prior

research demonstrates that transformer-based embeddings significantly improve clustering outcomes in domains such as document organization, question-answer retrieval, and topic modeling.

2.4 Applications of Email Topic Modeling

Recent studies explore email topic modeling for enterprise communication analysis, cybersecurity, and workflow automation. However, many implementations rely on Latent Dirichlet Allocation (LDA) or supervised classifiers, both of which have limitations for dynamic and unlabeled datasets. There is growing interest in applying deep semantic models for clustering email threads, identifying user intent, and supporting intelligent inbox features.

This study builds on these developments by combining transformer-based embeddings with clustering algorithms to create a scalable and reliable email organization system.

3. Methodology

3.1 Data Collection and Preprocessing

We constructed an email dataset consisting of internal communications, publicly available email corpora, and synthetic messages designed to simulate common workplace topics. Preprocessing steps included:

- Removing email signatures, disclaimers, and quoted replies
- Lowercasing text
- Removing stop words
- Token normalization
- Optional lemmatization

Attachments and metadata (e.g., timestamps, sender information) were excluded for this study to focus solely on content-based clustering.

3.2 Feature Representation Approaches

Two distinct feature representation strategies were evaluated:

3.2.1 TF-IDF Vectorization

A classical bag-of-words model using TF-IDF weighting. This approach produces high-dimensional sparse vectors, capturing lexical frequency but lacking semantic context.

3.2.2 Transformer-Based Embeddings

Transformer-based sentence embeddings were generated using:

- **BERT-base** (average pooled embeddings)
- **Sentence-BERT (all-MiniLM-L6-v2)**, optimized for semantic similarity

These embeddings yield dense vector representations where semantically similar messages have smaller cosine distances.

3.3 Clustering Algorithms

Three clustering algorithms were considered:

3.3.1 K-Means

A centroid-based method requiring predefined cluster count k. While efficient, it assumes spherical cluster structures and struggles with non-linear separations.

3.3.2 Hierarchical Clustering

Agglomerative clustering using Ward's method. This algorithm does not require the number of clusters upfront and provides a dendrogram for visual analysis.

3.3.3 DBSCAN

A density-based algorithm capable of detecting arbitrary-shaped clusters and outliers. Key parameters include *eps* (neighborhood radius) and *min_samples*.

Given its ability to handle noise and semantic structures, DBSCAN was hypothesized to perform best when paired with transformer embeddings.

3.4 Evaluation Metrics

We employed intrinsic clustering metrics suitable for unsupervised evaluation:

- **Silhouette Score**
- **Davies–Bouldin Index**
- **Calinski–Harabasz Index**

We also performed a qualitative analysis by manually inspecting sample clusters to evaluate topic coherence.

4. Experiments and Results:

4.1 Experimental Setup

All experiments were conducted on a computing environment equipped with:

- 16 GB RAM
- NVIDIA GPU for embedding generation
- Python 3.10
- Scikit-learn, Transformers, and Sentence-Transformers libraries

For each method, we tested various hyperparameter configurations, including cluster numbers for K-Means (k = 5–20) and DBSCAN parameters (eps = 0.2–1.0).

4.2 TF-IDF + Classical Clustering

The TF-IDF + K-Means baseline produced clusters that were lexically similar but lacked semantic consistency. Messages containing similar keywords but different contexts were frequently grouped together. Silhouette scores were moderate but declined as the dimensionality increased.

Hierarchical clustering with TF-IDF performed slightly better, revealing subtopics within larger themes. However, dendrogram analysis showed overlapping cluster boundaries, indicating difficulty in separating semantic categories.

4.3 Transformer Embeddings + K-Means

Using transformer embeddings significantly improved cluster coherence. K-Means with Sentence-BERT vectors yielded clearer separation between thematic categories such as:

- Scheduling and meeting coordination
- Technical questions
- Project updates
- HR communications

However, K-Means struggled with smaller or irregularly shaped clusters, often forcing semantically distinct messages into nearby centroids.

4.4 Transformer Embeddings + DBSCAN

The strongest performance was obtained using transformer embeddings combined with DBSCAN. Sentence-BERT embeddings, in particular, produced high-density clusters that aligned well with human interpretation. The algorithm successfully identified:

- Minor topics such as travel requests
- Irregular but meaningful clusters such as urgent issue alerts
- Outlier detection for atypical or auto-generated emails

This approach achieved the highest silhouette scores and exhibited the strongest semantic coherence during manual inspection.

4.5 Summary of Results

Feature Representation	Clustering Method	Performance Summary
TF-IDF	K-Means	Weak semantic grouping; keyword-driven clusters
TF-IDF	Hierarchical	Better structure, but cluster overlap present
BERT	K-Means	Improved semantic cluster quality
Sentence-BERT	K-Means	Strong, coherent clusters but centroid limitations
Sentence-BERT	DBSCAN	Best overall performance; high coherence and noise detection

5. Discussion:

Our findings reinforce the importance of semantic embeddings for text clustering. Traditional TF-IDF vectors fail to capture contextual meaning, leading to clusters based on surface-level lexical similarity rather than true topics.

Transformer-based embeddings dramatically improve performance by encoding contextual semantics, enabling clustering algorithms to form coherent groups even for short or informal email text. DBSCAN further enhances results by identifying natural density patterns and isolating outliers, making it well-suited for real-world inbox data, where topic frequencies vary widely.

A significant advantage of the proposed approach is its practicality for enterprise-scale systems. Since it requires no labeled data, organizations can deploy it without privacy concerns or costly annotation efforts. The resulting clusters can support downstream tasks such as:

- Smart inbox organization
- Priority prediction
- Automated email routing
- Topic trend analysis
- User behavior modeling

Nonetheless, challenges remain. DBSCAN parameter tuning is sensitive to vector distributions, and transformer-based embeddings require substantial computation for very large datasets. Future research should investigate incremental clustering, hybrid topic models, and integration with metadata such as sender roles or conversation threads.

Conclusion & Future Work:

This research demonstrates the effectiveness of unsupervised learning for email topic clustering, particularly when leveraging modern transformer-based embeddings. Our results show that combining Sentence-BERT embeddings with DBSCAN produces the most coherent clusters, outperforming classical TF-IDF and centroid-based methods.

Future work may include:

- Real-time clustering for streaming email systems
- Multi-modal analysis incorporating metadata or attachments
- Adaptive clustering capable of detecting evolving topics
- Integration with large language models for interactive email summarization and routing
- Evaluation on larger and more diverse enterprise email datasets

The proposed approach lays the foundation for more intelligent and automated email management systems capable of meeting the growing demands of modern communication environments.

References

1. Androutsopoulos, S., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive Bayesian and a memory-based approach. *Lecture Notes in Computer Science, 1810*, 1–13. Springer.
2. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Text clustering and classification. In *Introduction to information retrieval* (pp. 363–400). Cambridge University Press.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186).
4. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings for semantic similarity. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992).
5. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).
6. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 226–231).