Article

Not peer-reviewed version

An Innovative Approach to Topic Clustering for Social Media and Web Data Using AI

Ioannis Kapadaidakis * , Emmanouil Perakakis , George Mastorakis , Ioannis Kopanakis

Posted Date: 31 March 2025

doi: 10.20944/preprints202503.2272.v1

Keywords: social media monitoring; social listening; topic clustering; data analysis; AI-powered analytics; intelligent insights; LLM

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025

doi:10.20944/preprints202503.2272.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

An Innovative Approach to Topic Clustering for Social Media & Web Data Using AI

Ioannis Kapadaidakis [1,*], Emmanouil Perakakis[1,2], Georgios Mastorakis [1,2] and Ioannis Kopanakis [1,2]

[1] Department of Management Science and Technology, Hellenic Mediterranean University; 72100 Agios Nikolaos, Greece

[2] Mentionlytics LTD; 20-22 Wenlock Road, London, N1 7GU, info@mentionlytics.com

* Correspondence: jkapad@hmu.gr

Abstract: The vast amount of social media and web data offers valuable insights for purposes such as brand reputation management, topic research, competitive analysis, product development, and public opinion surveys. However, analysing this data to identify patterns and extract valuable insights is challenging due to the vast number of posts, which can number in the thousands within a single day. One practical approach is topic clustering, which creates clusters of mentions that refer to a specific topic. Following this process will create several manageable clusters, each containing hundreds or thousands of posts. These clusters offer a more meaningful overview of the discussed topics, eliminating the need to categorise each post manually. Several topic detection algorithms can achieve clustering of posts, such as LDA, NMF, BERTopic, etc. The existing algorithms, however, have several important drawbacks, including language constraints and slow or resource-intensive data processing. Moreover, the labels for the clusters typically consist of a few keywords that may not make sense unless one explores the mentions within the cluster. Recently, with the introduction of AI Large Language Models, such as GPT-4, new techniques can be realised for topic clustering, to address the aforementioned issues. Our novel approach (*AI Mention Clustering*) employs LLMs at its core to produce an algorithm for efficient and accurate topic clustering of web and social data. Our solution was tested on social and web data and compared to the popular existing algorithm of BERTopic, demonstrating superior resource efficiency and absolute accuracy of clustered documents. Furthermore, it produces summaries of the clusters that are easily understood by humans instead of just representative keywords. This approach enhances the productivity of social and web data researchers by providing more meaningful and interpretable results.

Keywords: social media monitoring; social listening; topic clustering; data analysis; AI-powered analytics; intelligent insights; LLM

1. Introduction

Gathering data from social media and the web has become essential for businesses and researchers, offering various applications and benefits.

Brand monitoring has become increasingly important in the digital age, with social media platforms providing a wealth of data for enterprises to analyse and improve their reputation among consumers. Cloud-based big data sentiment analysis applications can be used for brand monitoring and analysis of social media streams, allowing enterprises to detect sentiment in social posts and their influence on consumers [1]. AI-powered social media monitoring platforms can provide intelligent insights for effective online reputation management and competitor monitoring, helping digital marketers better understand customers and improve their brand's web and social presence [2]. By

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

leveraging these tools, companies can enhance their competitiveness and better meet consumer needs and expectations in the digital landscape.

Moreover, social media monitoring extends beyond business applications. In the healthcare sector, it has been used to track public responses to health threats, such as the COVID-19 pandemic. For example, a study in Poland used social listening tools to analyse coronavirus discussions across various social media platforms [3].

Social media listening platforms have become increasingly popular for product research and development, offering valuable insights into customer preferences, market trends, and product feedback. These AI-driven tools extract actionable information from large amounts of social media data, addressing research questions and helping develop data-backed brand strategies [4].

While the many uses of large social media and web data make them valuable, data volume and velocity pose major obstacles to social media monitoring. The massive amount of user-generated content produced daily across platforms like Facebook, X, Instagram and YouTube creates difficulties in data storage, processing, and analysis [5]. The high dynamics and real-time aspects make effective capture and analysis difficult. [6]. Additionally, new social media are rising (e.g. TikTok, Threads, Bluesky, etc.), making it even more difficult to acquire and process all this heterogeneous data from all the different sources. Media Monitoring and Social Listening tools help greatly with collecting this data; however, they often lack advanced functionality for efficiently processing and analysing large amounts of data [7].

Topic detection refers to the clustering of different pieces of content based on the similarity of the topic they discuss. It focuses on identifying and extracting meaningful topics from large volumes of textual data, particularly news streams and social media content [8,9]. An example of this, applied in news articles, is how Google clusters multiple news sources under a news topic in Google News, so that the reader can see a list of today's topics easily. Google News employs sophisticated topic clustering algorithms to effectively organise and present news articles [10]. If the reader is interested in more coverage of a particular news topic, they can easily see the different sources, with the news pieces about the topic, and visit the different websites to see more. This makes Google News very easy to read, allowing users to get an overview of today's news in just a few seconds. This approach helps avoid repetitive browsing through similar materials and visiting multiple news sites' home pages. Therefore, the clustering process is crucial for assisting users in navigating, summarising, and organising the vast amounts of textual documents available on the internet and news sources. [11].

In this paper, the same principle is applied to social and web data gathered from social media listening tools. By clustering this data, users and data analysts will find it much easier to extract the information they seek more quickly and meaningfully.

2. Literature Review

Businesses and researchers often utilize brand monitoring and social listening to retrieve posts from multiple online sources. This is usually triggered by a keyword or a query related to their interests, which could include the name of a brand or specific product, a particular event, a public figure's name, or a location, among others. These tools typically leverage APIs (Application Programming Interfaces) provided by social media platforms to access and collect publicly available data. [12]. The posts and comments collected are typically referred to as “mentions. " Depending on the popularity of the keyword, the retrieved mentions can range from just a few to even millions. Analyzing social media mentions presents significant challenges due to the vast volume and dynamic nature of the data. The complexity of social media content requires human interpretation; however, the growing scale necessitates automated analysis techniques. [13]. Topic detection algorithms could be very helpful in clustering mentions that refer to the same or similar topics, even from multiple social media sources (e.g., X, Instagram, Facebook, YouTube, etc.), thereby consolidating multiple posts on the same topic into a single cluster. This could save an enormous amount of time for the users of such a system, as they would not need to go through each mention separately; instead, they

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

can quickly get an overview of the topics of mentions easily. They could then focus on the topic clusters they are most interested in for further analysis, cutting through the noise and clutter.

There are many different approaches to Topic detection and clustering. The next chapters outline the main categories of these algorithms.

2.1. "Traditional" Topic Detection Algorithms

2.1.1. Bag-of-Words Based

In this category, prominent topic modelling algorithms such as Latent Dirichlet Allocation (LDA) [14], Non-negative Matrix Factorization (NMF) [15], and Latent Semantic Analysis (LSA) [16] assume a bag-of-words representation of text, thereby, disregarding word order and semantic relationships. As a result, they provide topics that are less comprehensible and lack interpretability [29,30]. Furthermore, they encounter difficulties in distinguishing words that might have the same meaning (synonymy) or different meanings of the same word (polysemy), which results in mixed or inaccurate topic extraction [31]. Additionally, these algorithms perform ineffectively when processing short texts, such as social media posts, owing to the limited word availability, thereby hindering the discernment of underlying patterns [29].

2.1.2. Embedding-Based

Recent approaches in natural language processing, including BERTopic [17] and Top2Vec [18], use embeddings for text representation that offer enhanced coherence relative to Bag-of-Words based methodologies. Nevertheless, the actual representation of topics is based on Bag-of-Words and does not directly account for context, which might lead to redundancy in the words used to represent each topic. Moreover, resulting topics are presented as keyword lists that frequently lack clarity in interpretation, while certain mathematical inconsistencies within their formulations render them ineffective at eliminating stop words [32].

2.2. Using Large Language Models (LLMs)

In recent years, new methods incorporating LLMs in several ways into text clustering and topic analysis have emerged due to their current explosion. Some studies demonstrate that LLMs can serve as an intelligent guide to improve clustering outcomes, essentially injecting domain knowledge or preferences into the process [43,44]. It is also shown that LLMs, with appropriate prompting, can serve as an alternative to traditional topic modelling [41]. Furthermore, Miller et al [42], used LLMs to interpret clusters generated by other methods. Their results showed that an LLM-inclusive clustering approach produced more distinctive and interpretable clusters than LDA or doc2vec, as confirmed by human review.

However, Large Language Models (LLMs) present several challenges when applied to topic detection, particularly for large document collections. A key limitation is the restricted contextual limit. The contextual limit or context length in an LLM refers to the number of tokens that a model can process. Each model has its context length, also known as max tokens or token limit. For instance, a standard GPT-4 model has a context length of 128,000 tokens [33]. As a result, LLMs can only process a limited amount of text at once, meaning long documents must be split into chunks [40]. This approach to chunking, however, potentially compromises the prevailing context, resulting in incorrect topic detection.

Subsequently, using LLMs for large-scale text processing can be computationally expensive. Processing large corpora of data requires significant computational resources that incur high costs. For example, the costs of using the GPT4 model to analyse large datasets, like a corpus of 10K social media posts, will exceed $10 for input and output tokens. This cost can be prohibitive for many applications, especially when dealing with continuously updated datasets or real-time processing requirements.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

Ongoing academic work into novel techniques, including hierarchical summarisation and memory-augmented LLMs [19,20], aims to moderate these obstacles. However, these emerging methodologies remain under development and do not eliminate the challenges associated with processing sizable amounts of documents using LLMs.

This work is distinguished from previous research by combining the strengths of traditional clustering and LLMs while moderating their weaknesses. Unlike existing methods that attempt to prompt an LLM with an entire corpus [41], we first employ a classical unsupervised clustering to split data into coherent groups. We then apply the LLM exclusively to a small subset of representative documents from each cluster. This minimises the LLM's context requirements and reduces costs to a fraction of what they would be for processing the full dataset. Yet, it still harnesses its powerful language understanding to generate interpretable summaries. In the next section, we detail the methodology of AI Mention Clustering, which embodies these theoretical innovations.

3. Proposed Solution

We propose a novel approach for topic detection in social media corpora that exploits the power of Large Language Models (LLMs) while minimising the computational cost. Our method applies a clustering-based approach to group semantically similar social media documents (posts) together and then uses LLMs to analyse the clusters but only by sending a small subset of representative documents from each cluster. This allows us to efficiently and cost-effectively process large datasets of social media posts while benefiting from the LLMs' advanced language understanding capabilities. Specifically, our approach consists of the following steps (Figure 1):

1. Extract embeddings from Web & Social Media documents (posts)
2. Cluster embeddings
3. Specify the representative social media documents from each cluster
4. Send the representatives to LLM to extract topics and summarise.

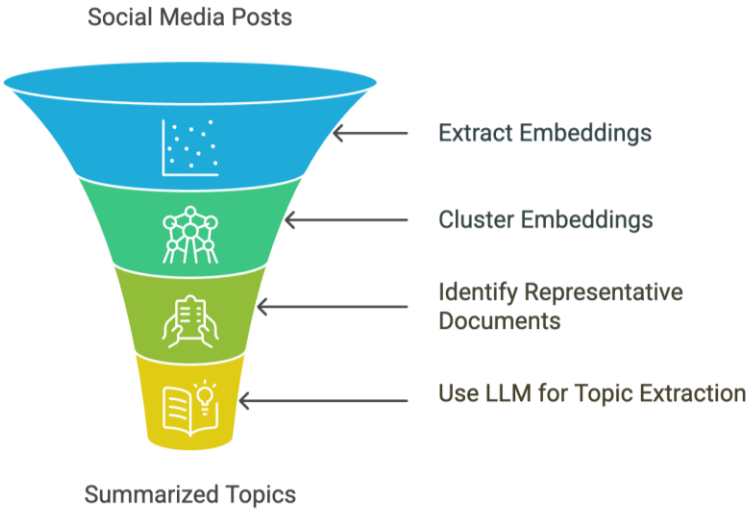


Figure 1. Overview of the proposed solution - *AI Mention Clustering*.

1. *Extract Embeddings*

In this step we transform the text of social media documents into a sequence of numbers (i.e. vectors) called embeddings. These embeddings capture the semantic meaning of the text, with similar documents having similar vector representations (i.e. close to each other in the vector space). Various techniques and models can be used for embedding generation. From traditional methods like

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

Word2Vec[21] and TF-IDF[22] to more advanced (transformer-based models). In this category, we can find both free (Sentence-BERT [23]) and commercial (OpenAI's Ada) models.

2. *Cluster Embeddings*

Once we generate the social media document embeddings, we can employ a clustering algorithm like DBScan [24], K-means [25] or Optics [26]. This step aims to group semantically similar documents together, assuming that documents within the same cluster discuss related topics and to identify outliers in order to exclude them from further processing.

3. *Specify the Representative Documents*

Rather than sending all documents within a cluster to the LLM as input, which can be computationally expensive and cost-inefficient, in this step, we select a small percentage of a few representative documents from each cluster. These representatives should ideally capture the core themes and discussions within the cluster. Various methodologies can be employed to determine the representatives, including identifying documents proximal to the cluster centroid or determining medoids [27].

The concept of a medoid refers to a representative point within a cluster that minimises the average dissimilarity (or distance) to all other points within the cluster. Medoids are similar in concept to means or centroids, but medoids are always restricted to be members of the data set. The formal definition of a medoid is the following:

Let $X \coloneqq \{x_1, x_2, \dots x_n\}$ be a set of n points in a space with a distance function d. A medoid is defined as [27]:

$$x_{medoid} = \arg\min_{y \in X} \sum_{i=1}^{n} d(y, x_i)$$

This step of specifying the medoid representatives for each cluster will significantly reduce the total amount of input data that the LLM will finally process.

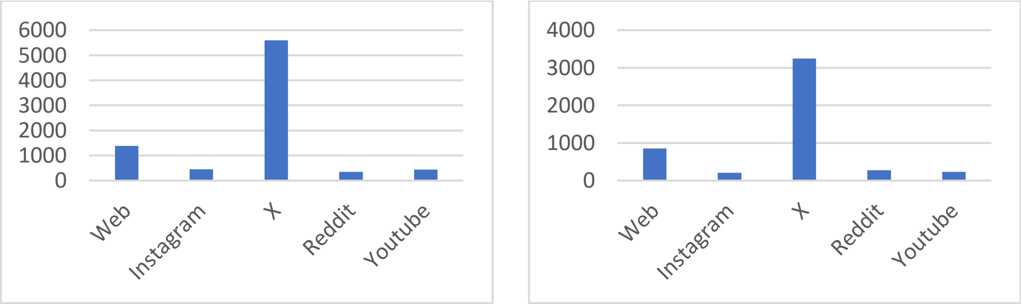
4. *Send Cluster Representatives to LLM*

In this final step, each cluster's representative documents are sent as input to the LLM. The LLM is then prompted to generate a summary of the overall discussion within each cluster, thus providing a cohesive overview of each topic. This process exploits the LLM's text analysis and synthesis capabilities to produce topic summaries that are both meaningful and comprehensible to humans.

4. Evaluation

4.1. Qualitative Evaluation

In order to evaluate our AI Mention Clustering, we created a dataset of approximately 10K multilingual posts about Ryanair from various websites and social media platforms and we compared our approach with the BERTopic algorithm. Furthermore, a secondary dataset, consisting of about 5,000 exclusively English posts, was used for the evaluation (Figures 2 and 3).



(**a**) (**b**)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

Figure 2. Source distribution for: (**a**) Multilingual and (**b**) English Dataset.

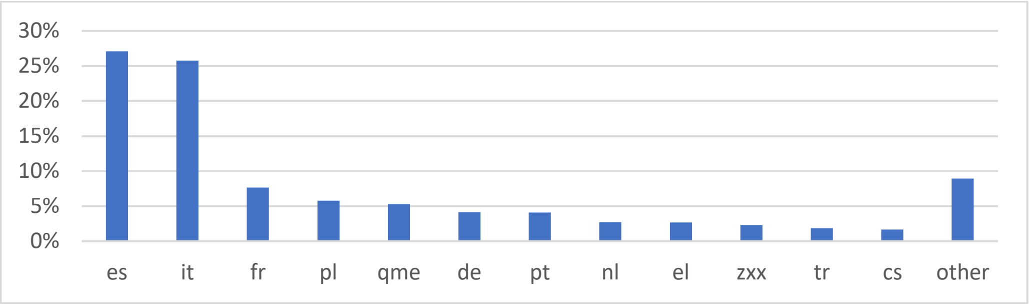
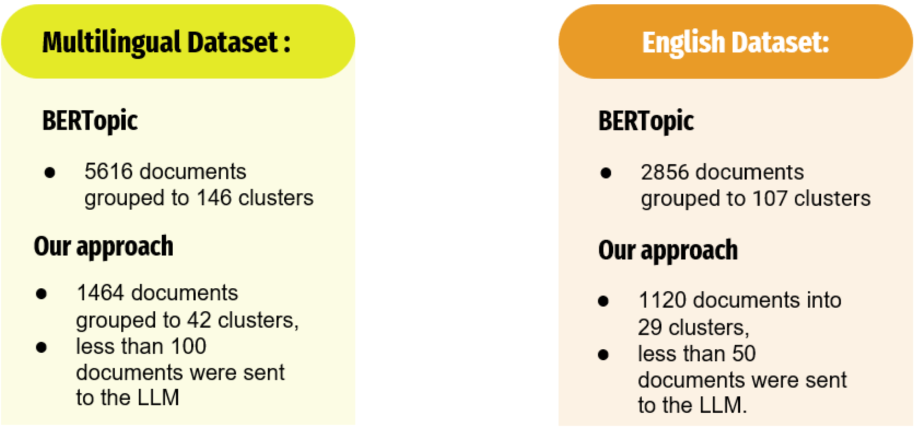


Figure 3. Language distribution for the Multilingual dataset (English Language excluded).

To create vector representations of the data, our evaluation used OpenAI's text-embedding-ada-002 embeddings (dimension 1536). We used the density-based spatial clustering of applications with noise (DBSCAN) approach, with parameters (min_samples=5, epsilon=0.24), because the number of clusters in the dataset was unknown beforehand. Medoids, which offer a reliable indicator of central tendency, were chosen as cluster representatives. In practice, we found that using 3 representative posts per cluster worked well for large clusters, and just 1 or 2 for smaller clusters Lastly, ChatGPT-4 large language model was used to summarize each cluster using the following prompt: ”Write a summary up to 30 words for the following list of news titles and social media posts”.

In Figure 4 we present a summary of the resulting clusters in both datasets. One important difference that someone could easily detect is the variation between our AI Mention Clustering and BERTopic in the proportion of documents assigned to clusters and the granularity of the clustering itself. BERTopic clustered more than 50% of the total posts in both datasets, resulting in a larger number of clusters. This indicates an over clustering strategy that potentially contains noise and fragmenting topics. On the contrary, our approach clustered a percentage between 15% to 20% of the total posts, indicating a more refined clustering approach while we achieved accurate topics, without irrelevant documents within each of the generated clusters. Another crucial point to note is that our approach achieves significant performance efficiency through cautious use of Large Language Models, utilizing less than 1% of the posts, by only sending representative posts, thereby reducing costs and computational demands. This efficiency makes our method a more practical and scalable option for large datasets.



Multilingual Dataset :

BERTopic

- 5616 documents grouped to 146 clusters

Our approach

- 1464 documents grouped to 42 clusters,
- less than 100 documents were sent to the LLM

English Dataset:

BERTopic

- 2856 documents grouped to 107 clusters

Our approach

- 1120 documents into 29 clusters,
- less than 50 documents were sent to the LLM.

(**a**) (**b**)

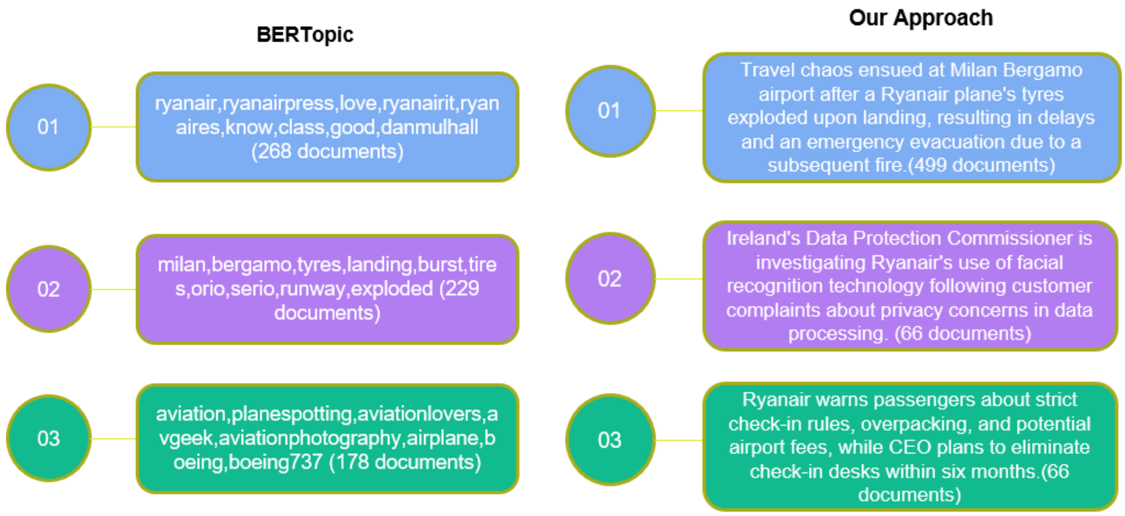
Figure 4. Summary of resulting clusters in: (**a**) Multilingual and (**b**) English Dataset.

Starting with the English dataset, in Figure 5 we can see from the top 3 clusters that while both methods concur that a major incident at a Milan airport was important, they differ greatly in how effectively they capture it. Our method recognized this event as the main topic, grouping 499

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

documents into it. On the other hand, BERTopic also noted this event but not as the main one, and only assigned 229 documents to it, which is less than half of what our approach did. This difference suggests that BERTopic may have missed or misclassified many relevant posts related to this event and incorrectly placed them in less relevant clusters.

Additionally, our method created more precise and easily interpretable topics than BERTopic's clustering, which produced clusters with generic key words (e.g. "love","know","good","airplane" etc.) that reduce interpretability and the identification of underlying topics. Thus, it was really hard for a human reader to understand what each cluster is about, based only on these few keywords. A user will need to look at a number of mentions from within the cluster to understand the actual topic. On the contrary, our approach describes each cluster as a textual summary of it's inner mentions. This description is very accurate, and a user can simply understand the full context of the cluster without any need to read the actual mentions, thus making it very efficient for analysts to understand the results.



ryanair,ryanairpress,love,ryanairit,ryanaires,know,class,good,danmulhall (268 documents)

milan,bergamo,tyres,landing,burst,tires,orio,serio,runway,exploded (229 documents)

aviation,planespotting,aviationlovers,avgeek,aviationphotography,airplane,boeing,boeing737 (178 documents)

Travel chaos ensued at Milan Bergamo airport after a Ryanair plane's tyres exploded upon landing, resulting in delays and an emergency evacuation due to a subsequent fire.(499 documents)

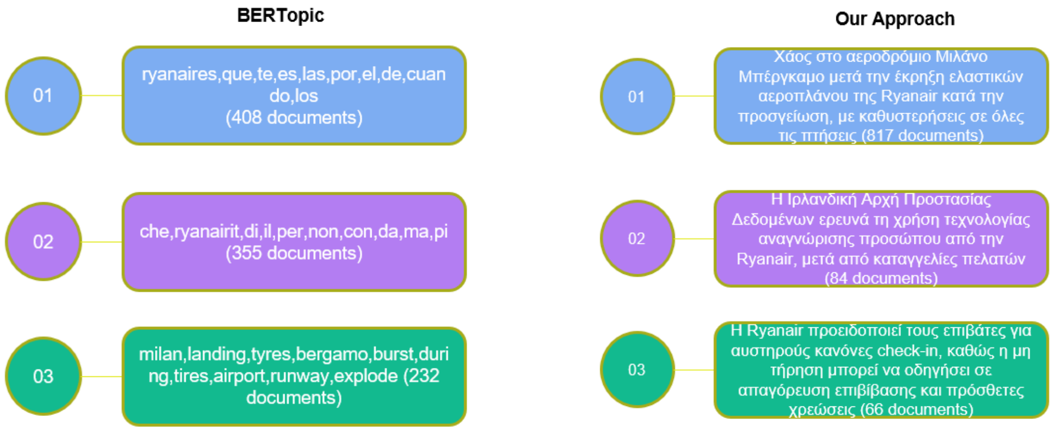
Ireland's Data Protection Commissioner is investigating Ryanair's use of facial recognition technology following customer complaints about privacy concerns in data processing. (66 documents)

Ryanair warns passengers about strict check-in rules, overpacking, and potential airport fees, while CEO plans to eliminate check-in desks within six months.(66 documents)

(**a**) (**b**)

Figure 5. Top 3 clusters in the English dataset for (a) BERTopic and (b) AI Mention Clustering.

The performance gap wasn't just a problem with the English language as it worsened significantly with the multilingual dataset (Figure 6). On the Milan airport incident topic (which is clearly the major topic in the dataset), our method identified 819 relevant documents, showcasing its robust multilingual skills. BERTopic, in contrast, only found 232 related documents, a notably smaller portion that placed it in the third position. This significant difference suggests that BERTopic may not be able to capture key information across multilingual data.



ryanaires,que,te,es,las,por,el,de,cuando,los (408 documents)

che,ryanairit,di,il,per,non,con,da,ma,pi (355 documents)

milan,landing,tyres,bergamo,burst,during,tires,airport,runway,explode (232 documents)

Χάος στο αεροδρόμιο Μιλάνο Μπέργκαμο μετά την έκρηξη ελαστικών αεροπλάνου της Ryanair κατά την προσγείωση, με καθυστερήσεις σε όλες τις πτήσεις (817 documents)

Η Ιρλανδική Αρχή Προστασίας Δεδομένων ερευνά τη χρήση τεχνολογίας αναγνώρισης προσώπου από την Ryanair, μετά από καταγγελίες πελατών (84 documents)

Η Ryanair προειδοποιεί τους επιβάτες για αυστηρούς κανόνες check-in, καθώς η μη τήρηση μπορεί να οδηγήσει σε απαγόρευση επιβίβασης και πρόσθετες χρεώσεις (66 documents)

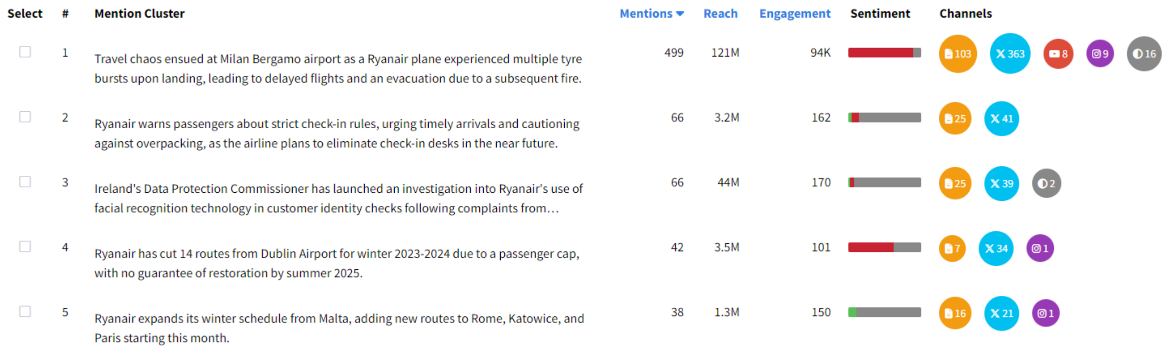
(**a**) (**b**)

Figure 6. Top 3 clusters in Multilingual dataset for (a) BERTopic and (b) AI Mention Clustering.

Furthermore, another limitation BERTopic exhibited in processing the multilingual dataset is that dominant clusters contain high-frequency words such as "que," "te," "por," "el," "da," "ma," and "pi." These words lack semantic significance to the underlying topics and are considered as stop words, meaning that they should have been excluded. This shows a weakness in the model's ability to effectively filter noise from multilingual data.

In contrast, our approach produced the same topics as in the English dataset and only the number of the assigned documents was altered. However, we requested the summary in Greek to demonstrate the potential to leverage topic summaries across a diverse range of languages despite the actual languages that appear in the multilingual dataset.

The screenshot provided in Figure 7 illustrates a commercial implementation of our methodology, utilized by the Social Listening tool Mentionlytics [28], which depicts a ranked cluster ordering based on the number of documents in each cluster. Additionally, informative data and key metrics, such as accumulated engagement, overall reach, sentiment, and corresponding channel sources for documents within each cluster, are also presented.



Select	#	Mention Cluster	Mentions ▾	Reach	Engagement	Sentiment	Channels
□	1	Travel chaos ensued at Milan Bergamo airport as a Ryanair plane experienced multiple tyre bursts upon landing, leading to delayed flights and an evacuation due to a subsequent fire.	499	121M	94K		103 X 363 8 9 16
□	2	Ryanair warns passengers about strict check-in rules, urging timely arrivals and cautioning against overpacking, as the airline plans to eliminate check-in desks in the near future.	66	3.2M	162		25 X 41
□	3	Ireland's Data Protection Commissioner has launched an investigation into Ryanair's use of facial recognition technology in customer identity checks following complaints from…	66	44M	170		25 X 39 2
□	4	Ryanair has cut 14 routes from Dublin Airport for winter 2023-2024 due to a passenger cap, with no guarantee of restoration by summer 2025.	42	3.5M	101		7 X 34 1
□	5	Ryanair expands its winter schedule from Malta, adding new routes to Rome, Katowice, and Paris starting this month.	38	1.3M	150		16 X 21 1

Figure 7. An implementation of our AI Mention Clustering applied in the Social Listening tool Mentionlytics [28].

4.2. Quantitative Evaluation

Besides qualitative evaluation, a quantitative assessment was required to compare our method with BERTopic. For this purpose, we used the previously described English Ryanair dataset, and we added three more datasets: Easyjet (another aviation company), Trello (a computer software) and Asana (another computer software). To offer a more comprehensive review scope and reduce the inherent bias of relying on one source, these datasets differed in size, chronological range, and span in two very different industries (Aviation and Computer Software).

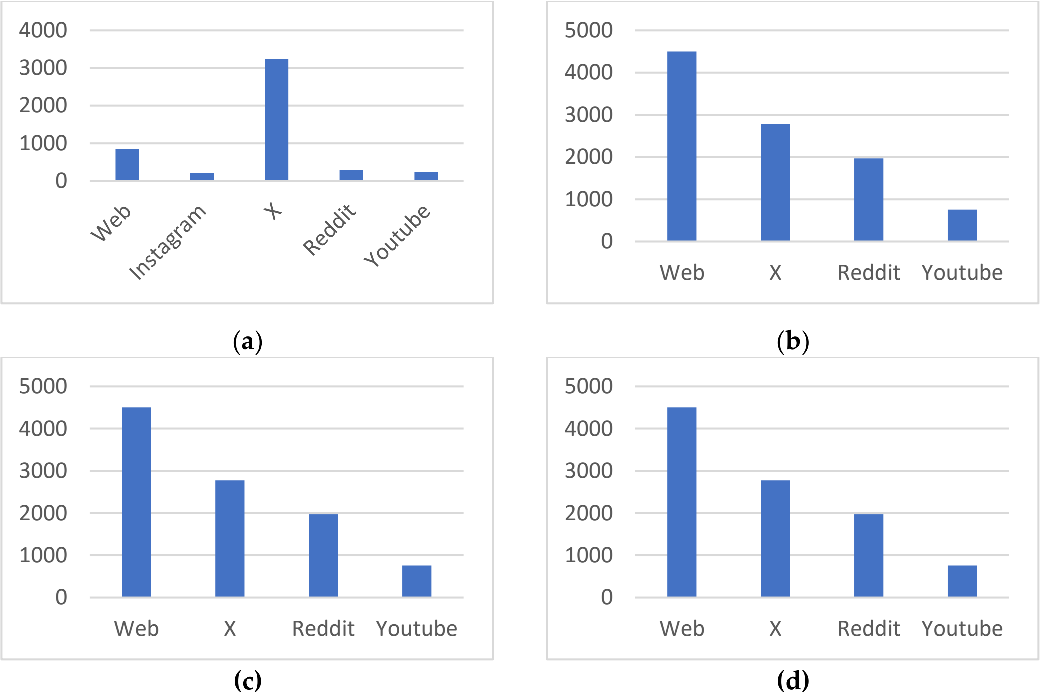
We selected these datasets to represent typical social listening scenarios: two from the airline industry and two from the tech industry, encompassing different time spans and dataset sizes. This variety ensures that our evaluation includes cases of relatively focused conversation (software communities) and broad, sometimes volatile discussions (airline customers and news). It also enables us to test how the approach scales from approximately 5,000 to 10,000 documents**.** All datasets consist of public posts collected through a social listening tool (Mentionlytics) by querying specific keywords, primarily their brand names**.** Duplicate posts (exact repeats or retweets) were eliminated. We conducted light preprocessing, which involved removing URLs, emojis, and Twitter handles (usernames) to minimize noise in topic modeling.

All four datasets were used for our evaluation. The clustering results from both approaches are described in Table 1, while Figure 8 depicts their source distribution.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

Table 1. Dataset description and clustering result summarization.

Dataset	Ryanair	Easyjet	Trello	Asana
Size	5600	5216	9470	10004
Date Range	28/9 – 4/10	1/12 – 18/12	1/12 – 15/1	1/11 -10/1
Language	English	English	English	English
AI Mention Clustering (% total document)	29 (20%)	57 (19%)	68 (18%)	75 (18%)
BERTopic Clustering (% total document)	107 (51%)	125 (74%)	156 (60%)	159 (63%)



(a) **(b)**

(c) **(d)**

Figure 8. Source distribution for datasets: (a) Ryanair, (b) Easyjet, (c) Trello and (d) Asana.

As we already noted in the previous chapter, for the Ryanair dataset, over half of the posts in each of the three new datasets were clustered by BERTopic, indicating the possibility of producing noise and fragmenting topics. Our method, in comparison, keeps grouping a much smaller percentage of mentions of the dataset (15–20%), suggesting a more focused and refined clustering technique.

Since our approach outputs human-readable summaries, instead of keywords for each cluster, as a first step we used the TF-IDF technique [22] to identify the top-10 most important keywords per cluster (the main keywords from each dataset i.e. *Ryanair*, *Easyjet*, *Trello* and *Asana* respectively were excluded). The TF-IDF score for a term in a document is obtained by multiplying its TF and IDF scores.

$$TF - IDF(t, d, D) = \mathrm{TF}(\mathrm{t}, \mathrm{d}) \times IDF(t, D)$$

where:

$$\boldsymbol{TF(t,d)} = \frac{\boldsymbol{Number\ of\ times\ term\ t\ appears\ in\ document\ d}}{\boldsymbol{Total\ number\ of\ terms\ in\ document\ d}}$$

$$\boldsymbol{IDF(t,D)} = \boldsymbol{log}\left(\frac{\boldsymbol{Total\ number\ of\ documents\ in\ the\ corpus\ N}}{\boldsymbol{Number\ of\ documents\ containing\ term\ t}}\right)$$

Using the keyword sets found using the TF-IDF technique, we calculated two important metrics: topic coherence and topic diversity. These metrics would allow us to quantitively measure our method’s performance against BERTopic across the four datasets. Also, to evaluate the clustering structure itself we calculated Davies-Bouldin Index metric [45].

Topic Coherence

The topic coherence metric [37] assesses the semantic similarity of words within a given cluster identified by clustering (or topic modelling) algorithm. Assuming $T = \{w1, w2, \dots, wn\}$ as a generated topic which is represented by its top-n most important words and given a similarity measure $Sim(wi, wj)$ topic coherence is defined as follows:

$$TopicCoherence = \frac{\sum_{\substack{1\leq i \leq n-1 \\ i+1 \leq j \leq n}} Sim(wi, wj)}{\binom{n}{2}}$$

A high coherence score suggests a well-defined and relevant topic since it shows that the words within the topic are closely connected and make intuitive sense together. On the other hand, a low coherence score suggests that the topic is poorly defined or meaningless and that the words are mostly unrelated. For our evaluation, we used the Cv method, which, as described in [34], was found to correlate the highest with human interpretation.

Table 2 represents the resulting coherence scores of our approach compared to BERTopic across all four datasets. Our approach achieved higher coherence scores in each case, from 5% to 12%. This suggests that our approach produces more semantically coherent topics compared to BERTopic.

Table 2. Topic Coherence performance.

Dataset	Ryanair	Easyjet	Trello	Asana
AI Mention Clustering	**0.46**	**0.40**	**0.37**	**0.38**
BERTopic	0.41	0.37	0.35	0.36

Topic Diversity

Topic Diversity metrics measure how distinct the generated topics are, ensuring that a clustering method does not output variations on the same topic. A high diversity score indicates that the clustering method identified distinct topics within the dataset, while low diversity scores suggest redundant and potentially recurrent topics. For evaluating topic diversity, we used two approaches 1) the proportion of the unique keywords to the total number of keywords produced from the computed clusters and 2) the word embedding-based centroid distance [35].

In this approach, we computed the FastText model using the embeddings of the keywords that describe each cluster [36]. Then, the diversity score is calculated as the average cosine distance between the centroids of clusters from all pairs of clusters (see Algorithm 1).

Despite BERTopic clustering a larger number of posts, Table 3 demonstrates that our approach achieves better topic diversity scores than BERTopic in all four datasets (*Ryanair*, *EasyJet*, *Trello*, and *Asana*). Although the two approaches' centroid distances are comparable, our approach's clustering extracts a larger percentage of unique keywords (between 67% and 78%) than BERTopic (48% to 53%). This suggests that our approach produces more unique and varied subjects.

Algorithm 1: *Word Embedding-Based Centroid Distance Calculation*

```
Input: clusters, embedding_model, topk=10
distances_array = [ ]
For each cluster1, cluster2 in combinations(clusters, 2) do:
    centroid1 = [ ]
    centroid2 = [ ]
    For each word1 in cluster1[:topk] do:
        centroid1 = centroid1 + embedding_model[word1]
    For each word2 in cluster2[:topk] do:
        centroid2 = centroid2 + embedding_model[word2]
    centroid1 = centroid1 / length(cluster1[:topk])
    centroid2 = centroid2 / length (cluster2[:topk])
    distances_array.append(distance.cosine(centroid1, centroid2))
```

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

```
return average (distances_array)
```

Table 3. Topic Diversity performance.

Dataset	Ryanair	Easyjet	Trello	Asana
AI Mention Clustering				
Unique Keywords	**78%**	**68%**	**67%**	**68%**
Centroid Distance	0.56	0.55	**0.54**	0.57
BERTopic				
Unique Keywords	52%	52%	48%	53%
Centroid Distance	0.56	0.55	0.53	**0.58**

Davies-Bouldin Index

The Davies-Bouldin Index (DBI) [45] helps us understand how good a clustering algorithm is. It looks at how similar items are within the same cluster and how different clusters are from each other. Lower values of the Davies-Bouldin Index indicate better clustering quality. Assuming that there is a dataset of k clusters $X = \{X_1, X_2, \dots, X_k\}$, the Davies-Bouldin Index can be calculated as:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max\left(\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)}\right)$$

where $\Delta(X_k)$ is the intracluster distance (compactness) within the cluster X_k and $\delta(X_i, X_j)$ is the intercluster distance (separation) between the clusters X_i and X_j.

Table 4 represents the DBI scores of our approach compared to BERTopic across all four datasets. Our approach exhibits significantly lower DBI scores than BERTopic for each dataset tested. This suggests AI Mention Clustering creates more distinct and well-defined clusters compared to BERTopic for these datasets.

Table 4. Davies-Bouldin Index scores.

Dataset	Ryanair	Easyjet	Trello	Asana
AI Mention Clustering	**1.8460**	**1.7529**	**1.6616**	**1.6689**
BERTopic	3.0871	3.1058	3.4994	3.4099

5. Discussion & Future Work

This work presents an efficient approach for extracting easily interpretable topics from large social media data. By leveraging the power of large language models (LLMs) for natural language processing, we achieve effective topic modelling compared to BERTopic for both multilingual and language-specific datasets while maintaining cost-effectiveness since only 1% of the posts were sent to the LLM for processing.

The demonstrated methodology generates meaningful interpretations of topics from noisy social media data and could offer valuable insights for various applications, including social trend analysis, market research, social media crisis identification and public opinion monitoring. Additionally, the underlying framework's adaptability raises the possibility that it may be used for NLP tasks other than topic extraction, like knowledge graph generation, sentiment analysis, and named entity recognition (NER).

Future research includes optimisations in the clustering step of our methodology. Techniques such as dimensionality reduction on embedding representations. Dimensionality reduction techniques are crucial in improving the efficiency and effectiveness of embedding representations. These methods aim to preserve essential information while reducing the dimensionality of high-dimensional data, which is particularly useful for word embeddings and other types of vector representations [38]. Also, parallelisation within clustering algorithms will further enhance the methodology's capability to process larger volumes of social data rapidly. Parallel clustering algorithms distribute the workload across multiple processors, allowing for simultaneous computation of different parts of the clustering process [39].

Additionally, a comprehensive evaluation of different embedding models, LLMs, and alternative methods for selecting the most representative documents for each cluster could further improve the interpretability and accuracy of the extracted topics. We can also compare our approach to other topic modelling algorithms besides BERTopic, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Top2Vec.

In conclusion, while the current approach demonstrates considerable effectiveness and efficiency, ongoing improvements and comparisons with other methodologies will ensure that the solution remains at the forefront of topic modelling in social media and web data analytics. The continued evolution of these techniques promises even greater scalability and adaptability in the future, opening up new possibilities for effective social data analysis.

Data Availability Statement: The data analyzed in this study is public posts data available on Social Media and the Web as described in the corresponding section. It can be derived from the Social Media APIs of the providers or by using a Social Media Monitoring tool using the described keywords and dates.

Conflicts of Interest: The authors declare that E. Perakakis, G. Mastorakis & I. Kopanakis are cofounders of Mentionlytics Ltd. The data collection for this study was conducted using Mentionlytics, developed by Mentionlytics Ltd. The company had no role in the study design, data analysis, interpretation of results, manuscript preparation, or the decision to publish.

References

1. Tedeschi, A.; Benedetto, F. A cloud-based big data sentiment analysis application for enterprises' brand monitoring in social media streams. *Proc. IEEE RSI Conf. on Robotics and Mechatronics* **2015**, 2, 186–191. https://doi.org/10.1109/rtsi.2015.7325096
2. Perakakis, E.; Mastorakis, G.; Kopanakis, I. Social Media Monitoring: An Innovative Intelligent Approach. *Designs* **2019**, *3*(2), 24. https://doi.org/10.3390/designs3020024
3. Burzyńska, J.; Bartosiewicz, A.; Rękas, M. The social life of COVID-19: Early insights from social media monitoring data collected in Poland. *Health Informatics Journal* **2020**, *26*(4), 3056–3065. https://doi.org/10.1177/1460458220962652
4. Hayes, J.L.; Britt, B.C.; Evans, W.; Rush, S.W.; Towery, N.A.; Adamson, A.C. Can Social Media Listening Platforms' Artificial Intelligence Be Trusted? Examining the Accuracy of Crimson Hexagon's (Now Brandwatch Consumer Research's) AI-Driven Analyses. *J. Advert.* **2020**, *50*(1), 81–91. https://doi.org/10.1080/00913367.2020.1809576
5. Hussain, Z.; Hussain, M.; Zaheer, K.; Bhutto, Z. A.; Rai, G. Statistical Analysis of Network-Based Issues and Their Impact on Social Computing Practices in Pakistan. *J. Comput. Commun.* **2016**, *4*(13), 23–39. https://doi.org/10.4236/jcc.2016.413003
6. Shi, L.; Luo, J.; Zhu, C.; Kou, F.; Cheng, G.; Liu, X. A survey on cross-media search based on user intention understanding in social networks. *Inf. Fusion* **2022**, *91*, 566–581. https://doi.org/10.1016/j.inffus.2022.11.017
7. Kitchens, B.; Abbasi, A.; Claggett, J. L. Timely, Granular, and Actionable: Designing a Social Listening Platform for Public Health 3.0. *MIS Q.* **2024**, *48*(3), 899–930. https://doi.org/10.25300/misq/2023/17381
8. He, Q.; Lim, E.-P.; Banerjee, A.; Chang, K. Keep It Simple with Time: A Reexamination of Probabilistic Topic Detection Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*(10), 1795–1808. https://doi.org/10.1109/tpami.2009.203
9. Li, C.; Liu, M.; Yu, Y.; Wang, H.; Cai, J. Topic Detection and Tracking Based on Windowed DBSCAN and Parallel KNN. *IEEE Access* **2020**, *9*, 3858–3870. https://doi.org/10.1109/access.2020.3047458
10. Ahmed, A.; Ho, Q.; Smola, A. J.; Teo, C. H.; Xing, E.; Eisenstein, J. Unified analysis of streaming news. In *Proceedings of the 2011 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, August 21–24, **2011**; ACM: New York, NY, USA, 2011; pp. 1–9. https://doi.org/10.1145/1963405.1963445

11. Lu, Q.; Conrad, J. G.; Al-Kofahi, K.; Keenan, W. Legal document clustering with built-in topic segmentation. *Proceedings of the Fifth International Conference on Statistical Data Analysis Based on the L1-Norm and Related Methods*, Shanghai, China, July 5–8, 2011; Elsevier: Amsterdam, The Netherlands, **2011**; pp. 383–392.
12. Davis, C. A.; Serrette, B.; Hong, K.; Rudnick, A.; Pentchev, V.; Menczer, F.; Gonçalves, B.; Grabowicz, P. A.; Mckelvey, K.; Chung, K.; Ciampaglia, G. L.; Ratkiewicz, J.; Ferrara, E.; Peli Kankanamalage, C.; Wu, T.-L.; Flammini, A.; Meiss, M. R.; Shiralkar, P.; Aiello, L. M.; Weng, L. OSoMe: The IUNI Observatory on Social Media. *PeerJ Comput. Sci.* **2016**, *2*, e87. https://doi.org/10.7717/peerj-cs.87
13. Chen, X.; Vorvoreanu, M.; Madhavan, K. P. C. Mining Social Media Data for Understanding Students' Learning Experiences. *IEEE Trans. Learn. Technol.* **2014**, *7*(3), 246–259. https://doi.org/10.1109/tlt.2013.2296520
14. Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. https://doi.org/10.1145/945138.945145
15. Lee, D. D.; Seung, H. S. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* **1999**, *401*(6755), 788–791. https://doi.org/10.1038/44565
16. Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*(6), 417–428.
17. Grootendorst, M. BERTopic: Neural topic modeling with a class-based embedding model. *arXiv* **2022**, arXiv:2203.05794. Available online: https://arxiv.org/abs/2203.05794
18. Angelov, D. Top2Vec: Distributed Representations of Topics. *arXiv* **2020**, arXiv:2008.09470. Available online: https://arxiv.org/abs/2008.09470
19. Li, P.; et al. Hierarchical Summarization with Reusable Abstractive Units. *arXiv* **2023**, arXiv:2305.14546. Available online: https://arxiv.org/abs/2305.14546
20. Shao, Y.; et al. Long-Range Summarization with Memory-Augmented Transformers. In *Proceedings of the 2023 Association for Computational Linguistics Conference*, Toronto, ON, Canada, 23–28 July 2023; Association for Computational Linguistics: Toronto, Canada, **2023**. Available online: https://arxiv.org/abs/2305.14546
21. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems* (NeurIPS 2013), Lake Tahoe, NV, USA, December 5–10, **2013**; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119. Available online: https://arxiv.org/abs/1310.4546
22. Salton, G.; Buckley, C. Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
23. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2019), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Hong Kong, China, **2019**; pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410
24. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Ester, M.; Kriegel, H. P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (KDD-96), Portland, OR, USA, August 2–4, **1996**; AAAI Press: Portland, OR, USA, 1996; pp. 226–231. Available online: https://doi.org/10.5555/3001460.3001507
25. MacQueen, J. K. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, June 21–23, **1967**; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
26. Ankerst, M.; Breunig, M. M.; Kriegel, H. P.; Sander, J. OPTICS: Ordering Points to Identify the Clustering Structure. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA, USA, June 1–3, **1999**; ACM: New York, NY, USA, 1999; pp. 49–60. https://doi.org/10.1145/304182.304187
27. Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: Hoboken, NJ, USA, 2005.
28. Mentionlytics [Computer Software]. Available online: https://www.mentionlytics.com (accessed on 15 March 2025).

29. Zhou, K.; Yang, Q. LDA-PSTR: A Topic Modeling Method for Short Text. In *Proceedings of the 2018 International Conference on Big Data Analysis*, Beijing, China, July 25–27, 2018; Springer: Singapore, **2018**; pp. 339–352. https://doi.org/10.1007/978-3-319-98643-9_34
30. Kim, H. D.; Zhai, C.; Park, D. H.; Lu, Y. Enriching Text Representation with Frequent Pattern Mining for Probabilistic Topic Modeling. *Proceedings of the American Society for Information Science and Technology* **2012**, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901062
31. Sriurai, W. Improving Text Categorization By Using A Topic Model. *Adv. Comput. Int. J.* **2011**, 2(6), 21–27.
32. Milios, E.; Zhang, X. MPTopic: Improving Topic Modeling via Masked Permuted Pre-training. *arXiv* **2023**, arXiv:2309.01015. Available online: https://arxiv.org/abs/2309.01015
33. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. Available online: https://arxiv.org/abs/2303.08774
34. Röder, M.; Both, A.; Hinneburg, A. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (WSDM 2015), Shanghai, China, February 2–6, **2015**; ACM: New York, NY, USA, 2015; pp. 399–408. https://doi.org/10.1145/2684822.2685324
35. Bianchi, F.; Terragni, S.; Hovy, D.; Nozza, D.; Fersini, E. Cross-lingual Contextualized Topic Models with Zero-shot Learning. *Proceedings of the 2021 European Chapter of the Association for Computational Linguistics (EACL 2021)*, Online, April 19–23, 2021; Association for Computational Linguistics: Online, **2021**; pp. 84–96. https://aclanthology.org/2021.eacl-main.9/
36. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *CoRR* **2016**, arXiv:1607.04606. Available online: http://arxiv.org/abs/1607.04606
37. David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic Evaluation of Topic Coherence. *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-HLT 2010), Los Angeles, CA, USA, June 1–6, **2010**; Association for Computational Linguistics: Los Angeles, CA, USA, 2010; pp. 100–108. https://doi.org/10.3115/1860610.1860645
38. Allaoui, M.; Kherfi, M. L.; Cheriet, A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In *Proceedings of the 2020 International Conference on Machine Learning and Data Science*, Singapore, September 6–8, **2020**; Springer: Cham, Switzerland, 2020; pp. 317–325. https://doi.org/10.1007/978-3-030-51935-3_34
39. Luo, G.; Luo, X.; Tian, L.; Gooch, T. F.; Qin, K. A Parallel DBSCAN Algorithm Based on Spark. *Proceedings of the 2016 IEEE International Conference on Big Data and Cloud Computing*, Beijing, China, November 4–6, **2016**; IEEE: Piscataway, NJ, USA, 2016; pp. 548–553. https://doi.org/10.1109/bdcloud-socialcom-sustaincom.2016.85
40. Borgeaud, S.; et al. Improving Language Models by Retrieving from Trillions of Tokens. *Proceedings of the International Conference on Machine Learning (ICML 2022)*, Baltimore, MD, USA, July 17–23, **2022**; PMLR: Baltimore, MD, USA, 2022. Available online: https://arxiv.org/abs/2208.10157
41. Mu, Y.; Dong, C.; Bontcheva, K.; Song, X. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. *arXiv* **2024**, arXiv:2403.16248. Available online: https://arxiv.org/abs/2403.16248
42. Miller, J. K.; Alexander, T. J. Human-Interpretable Clustering of Short-Text Using Large Language Models. *arXiv* **2024**, arXiv:2405.07278. Available online: https://arxiv.org/abs/2405.07278
43. Zhang, Y.; Wang, Z.; Shang, J. ClusterLLM: Large Language Models as a Guide for Text Clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2023), Singapore, November 7–11, **2023**; Association for Computational Linguistics: Singapore, **2023**; pp. 13903–13920. https://doi.org/10.18653/v1/2023.emnlp-main.223
44. Viswanathan, V.; Gashteovski, K.; Lawrence, C.; Wu, T.; Neubig, G. Large Language Models Enable Few-Shot Clustering. *arXiv* **2023**, arXiv:2307.00524. Available online: https://arxiv.org/abs/2307.00524
45. Davies, D.; Bouldin, D. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1979**, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2025 doi:10.20944/preprints202503.2272.v1

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.