Event Detection in News Articles: A Hybrid Approach Combining Topic Modeling, Clustering, and Named Entity Recognition

Nikos Kapellas[a] and Sarantos Kapidakis[b]
Department of Archival, Library and Information Studies, University of West Attica, Ag. Spyridonos 28 12243, Athens, Greece

Keywords: Event Detection, News Articles, Topic Modeling, Natural Language Processing, Unsupervised Clustering, Named Entity Recognition.

Abstract: This research presents a comprehensive analysis of news articles with the primary objectives of exploring the underlying structure of the data and detecting events contained within news articles. The study collects articles from Greek online newspapers and focuses on analyzing a sub-set of this data, related to a predefined news topic. To achieve this, a hybrid approach that combines topic modeling, feature extraction, clustering, and named entity recognition, is employed. The obtained results prove to be satisfactory, as they demonstrate the effectiveness of the proposed methodology in news event detection and extracting relevant contextual information. This research provides valuable insights for multiple parties, including news organizations, researchers, news readers, and decision-making systems, as it contributes to the fields of event detection and clustering. Moreover, it deepens the understanding of applying solutions that do not require explicit human intervention, to real-world language challenges.

1 INTRODUCTION

1.1 Background and Motivation

The growing significance of online news has led to a massive flood of news articles, demanding in-depth analysis and comprehension. Extracting meaningful knowledge from this endless amount of data is a challenging task due to the complexity of the data and the diverse range of sources. In this context, the application and integration of various algorithms, that focus on language analysis and computational learning techniques, becomes vital for exploring hidden patterns and identifying events within news articles.

Motivated by the need to evaluate the reliability of news articles and news media, this research aims to progressively analyze news articles through distinct stages of the methodology, employing diverse techniques at each step. The objective is to process the textual content of articles, cluster them, and ultimately identify events, utilizing advanced methods such as unsupervised clustering and language models, among others. By detecting events and extracting contextual information, this research endeavors to enhance the awareness of expert news readers, casual news consumers, and organizations, and potentially support the decision-making process of an intelligent system that aggregates, organizes, filters, and presents news based on specific user criteria.

The research methodology encompasses several stages, starting with the creation of a dataset through web scraping of online newspapers. The dataset of nearly 300.000 entries, comprises a diverse collection of news articles from various domains, including politics, economy, social life, sports, and entertainment. The dataset is then processed using a pre-trained language model and subjected to topic modeling, which uncovers the underlying thematic structure. Subsequently, feature extraction and clustering algorithms are employed to group the articles within the specified dataset topic. The clustering process is guided by the evaluation of various metrics, such as the Silhouette score, Davies-Bouldin index, and Calinski-Harabasz index, to ensure the quality of the separation of clusters. The resulting clusters serve as the basis for uncovering the underlying news events structure. The process of identifying news events is performed at the last stage of the process by leveraging the named entity recognition capabilities of language

[a] https://orcid.org/0000-0002-2767-3956
[b] https://orcid.org/0000-0002-8723-0276

Kapellas, N. and Kapidakis, S.
Event Detection in News Articles: A Hybrid Approach Combining Topic Modeling, Clustering, and Named Entity Recognition.
DOI: 10.5220/0012234300003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 2: KEOD*, pages 272-279
ISBN: 978-989-758-671-2; ISSN: 2184-3228

models to identify specific entities, including geopolitical entities, organizations, and persons. By exploring the connections between the document contents and the identified named entities, events are detected and associated with multiple entities.

Overall, this research presents a novel approach that combines several techniques such as NLP processing, topic modeling, clustering, and named entity recognition to analyze news articles. The proposed methodology showcases the potential of applying unsupervised-automated techniques in discovering patterns and inspecting textual data. Also, it offers a solid framework for detecting events and extracting contextual information from news articles and provides a broad interpretation of news events, how they are tied to news articles, and which elements are prominent.

1.2 Relevant Work

Recent advances in computer science have led researchers to leverage the structure and capacity of Large Language Models (LLMs). In 2017, the Transformers architecture (Vaswani et al., 2017) was proposed, which introduced the concept of an attention function that eliminated the need for recurrent and convolution layers previously used in sequence modeling. Soon after various models based on the attention mechanism were born that outperformed the previous state-of-the-art models in various NLP tasks.

These models, such as BERT (Devlin et al., 2019) are essentially language representation models that are used to pre-train deep bidirectional vectors of unlabeled text. BERT is able to read the entire input sequence at once, incorporating both left and right context, and produces a contextualized representation for each word in the sequence. BERT-like models were released for other languages, including Greek. In (Koutsikakis et al., 2020) authors introduce GREEK-BERT, a language model specifically designed for the Greek language, aiming to address the lack of resources in the specific language.

In event detection, there are three main categories of methods used for topic detection. First, the document-pivot approach clusters documents based on similarity, second, the feature-pivot approach that extracts topic-representing terms using statistical methods, and lastly, the probabilistic approach involves topic modeling. These techniques aim to improve topic detection and uncover hidden structures in text corpora (Rafea and Gaballah, 2018).

Other approaches revolve around clustering. For example, (PATEL and PATEL, 2018) explores the problem of topic detection in news articles and introduces the task of topic tracking. The authors propose a combined approach that utilizes various learning techniques, including agglomerative clustering. Similar topics are also examined in (Shah and ElBahesh, 2004), where the proposed system addresses the need for news organizations to access related documents easily. Pursuing to solve difficult tasks, researchers have tried complex approaches that combine various worlds, for example, (Bouras and Tsogkas, 2013) tried to enhance the clustering process by incorporating a weights, clustering and n-grams.

There are many different clustering methodologies to approach event detection and while there is no unanimous definition for clustering, a common one involves maximizing similarity within clusters and maximizing dissimilarity between clusters. In general, clustering aims to extract representative features, design or apply an effective clustering algorithm, evaluate the clustering results, and provide practical interpretations of the outcomes (Xu and Tian, 2015).

On the other hand, traditionally event detection research focuses on two lines of investigation: theme detection, which identifies major themes with distinct semantics, and action extraction, which extracts fine-grained mention-level actions. (Zhang et al., 2022) introduced a new task called key event detection, aiming to identify key events within a news corpus. They define key events as specific events occurring at a particular time/location, focusing on the same topic. Others in the field researched the use of position vectors as time-embedding feature representations, combined with semantic features, to distinguish different nodes in a graph-based clustering approach (Liu et al., 2023).

The news domain is of particular interest to many researchers. In (Flynn and Dunnion, 2004) a system is presented that utilizes domain-informed techniques to group news reports into clusters that capture the narrative of events in the news domain. News can be also treated as events where a real-time and multilingual news event extraction system could accurately extract violent and natural disaster events from online news. (Piskorski et al., 2008) utilizes a linguistically lightweight approach to do so, that mostly relies on clustered news throughout the processing stages.

A more recent approach is that of topic modeling. Topic models are computer algorithms that identify patterns of word occurrence by analyzing the distribution of words across a collection of documents. The output of topic modeling is a set of topics, which are clusters of co-occurring words in the documents. The interpretation of these topics determines the usefulness of topic models in social science research (Jacobi et al., 2016).

In the event detection context, a new topic is defined as a significant event or activity along with directly related events and activities. For example, an earthquake and its associated discussions would form a topic. The first story reporting the occurrence of the earthquake is considered the initial story for that topic. Similarly, new event detection presents challenges, particularly in distinguishing between stories about similar events in different locations. Traditional vector space models rely on TF-IDF weighting schemes to highlight location differences. However, these methods often fail as the shared terms between earthquake-related stories overshadow location-specific distinctions (Kumaran and Allan, 2005).

In (Budiarto et al., 2021) authors propose an unsupervised model for news topic modeling, eliminating the need for manual labeling. The model utilizes Doc2Vec to generate word vectors for each article and applies a spectral clustering algorithm to group the data based on similarity. A recent topic modeling methodology, that integrates various techniques is Top2Vec (Angelov, 2020). Top2Vec leverages distributed representations of documents and words to capture the semantics and ordering of words. Topic modeling has also been applied for propaganda identification. (Kirill et al., 2020) introduces a method for identifying texts with propagandistic content using topic modeling.

Regarding the Greek news media landscape there is relevant research, each exploring news articles from a different perspective. For example, a recent study focused on analyzing the text similarity between news articles from different Greek news media. To do so, the authors examined three small distinct datasets, with different degrees of similarity, aiming to understand how differently Greek news media describe news events and the variations in their reporting (Kapellas and Kapidakis, 2022). Other researchers (Papadopoulou et al., 2021), explore the influence of global media on Greek news outlets in the context of intermedia agenda-setting. Another study explores the organizational structures of alternative media in Turkey and Greece. In Greece, the media landscape has been influenced by interconnected interests among private corporations, media organizations, and political powers, affecting the independence of journalism (Aslan Ozgul and Veneti, 2021).

2 METHODOLOGY

2.1 Dataset Creation

The creation of a news articles dataset was accomplished using an automated web scraping technique. Within the wider scope of our research area, the dataset creation process plays a pivotal role in any further analysis and in fact for smaller-audience languages, datasets or other linguistic resources are limited. Also, each language has its own inherent characteristics, thus applications that have given a solution to a challenging problem in a given language, might not apply to another one.

The process of web scraping the articles was initialized by carefully selecting online newspapers, considering factors such as popularity, readership, credibility, and topic coverage. This selection aimed to ensure a representative sample of articles from various domains, including politics, economy, social life, sports, and entertainment. During that stage of the research, a small-scale web scraper was utilized to extract links from a starting URL. These links include references to both internal and external web pages. Only internal links were used, to extract text from news articles. This approach ensured that the extracted text belongs to each newspaper's own domain and external links were not used in the web scraping process. In many cases, such external links point to social media websites, advertisements, that have no use to this study.

The process of collecting news articles (the term documents is also used interchangeably) involved three stages that occurred on different dates. These steps led to the extraction and storage of the text from approximately 300.000 news articles. Although this amount of data can be considered relatively small compared to large-scale decision or recommendation systems, it is sufficient to validate the methodology. Here a large amount of data is not a prerequisite for the event detection methodology.

These articles were obtained from popular Hreek news sources, including the following: in.gr, dikaiologitika.gr, news247.gr, flashnews.gr, newsbeast.gr, newsbomb.gr, efsyn.gr, gr.euronews.com, ieidiseis.gr, newsit.gr, thetoc.gr, protothema.gr, dimokratia.gr, cnn.gr, skai.gr, enikos.gr, kathimerini.gr, reporter.gr, ertnews.gr, capital.gr, dw.com, huffingtonpost.gr, tvxs.gr, imerisia.gr, ethnos.gr, and naftemporiki.gr.

The extracted text is basically high-dimensional unlabeled text, that cannot be used for classification purposes. This characteristic presents a significant challenge, as in the absence of labels learning algorithms cannot rely on prior knowledge. Instead, we must employ unsupervised learning techniques to discover meaningful patterns and relationships directly from the raw text.

2.2 Data Processing

In the first steps of the analysis, a filtering criterion was used to include quality and informative articles, by applying a size-related restriction. Specifically, news articles that had a content length of 700 characters or more were taken into account, while smaller articles were excluded. This approach allowed the elimination of document-level noise from the dataset, ensuring that the focus will be given to articles that provide sufficient information while avoiding incomplete, click-bait, or spam articles. After applying this filtering process, a total of 236,029 text articles met the criteria and were retained for subsequent analysis.

In the next step, several NLP techniques were applied to clean the corpus text, using a pre-trained language model:

1. Tokenization: Each document was tokenized, breaking it down into individual words or tokens.

2. Removal of Stop Words and Punctuation: Stop words, which are commonly occurring words with little semantic value, as well as punctuation marks, were removed from the tokens. This way, mostly meaningful content was retained.

3. Lemmatization: The remaining tokens were lemmatized, reducing words to their base or dictionary form. This normalization step potentially increases the accuracy of the analysis.

4. Elimination of Empty Spaces and Lines: Any remaining empty spaces or lines resulting from previous steps were removed, ensuring a clean and coherent text corpus.

5. Document Reconstruction: The processed tokens were reassembled into a single string, representing the cleaned version of each document.

The NLP data processing described plays a pivotal role in this study. It enables the following stages of the methodology, improves the accuracy of the algorithms, and ensures that the results will describe meaningful content.

2.3 Topic Modeling

After creating and cleaning the news dataset a topic modeling algorithm was applied, to uncover the underlying themes and subjects present in the text. For this purpose, a method that leverages the strengths of advanced text representation and hierarchical clustering was used. This method, Top2Vec is able to capture the semantic relationships between words and these word embeddings serve as the basis for constructing topic vectors for individual documents. Out of the processed dataset of 236,029 news articles, there were 1,271 distinct topics identified. These topics represent different thematic categories found within the corpus. The distribution of documents across these topics varied, suggesting a less prominent theme within the dataset. Figure 1 displays the distribution of the documents per topic, providing an overview of the document-topic associations.

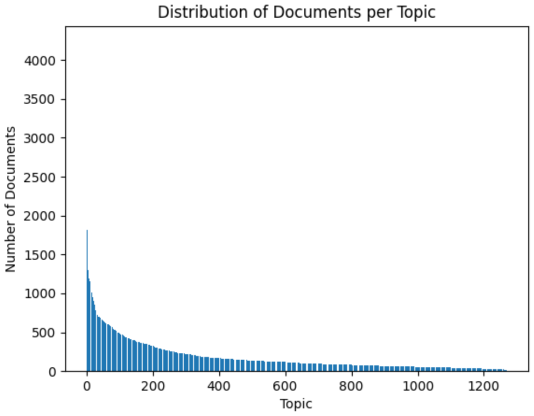


Figure 1: Distribution of the dataset documents in the topics as identified by Top2Vec algorithm.

Additionally, words that are strongly associated with each topic are extracted. These terms are identified based on the vectors learned by the model and capture semantic relationships between words. For instance, let's consider the third largest topic, which encompasses 1,813 documents. Some representative words for this topic, translated to English from Greek, are ”coronavirus”, ”epidemiologist”, ”transmissibility”, ”infectious disease specialist”, ”mutation”, ”co-infection”, ”super-spreader”, ”infectious disease specialists”, ”immunity”, ”unvaccinated”, ”lockdowns”, ”mandatory vaccination”.

These terms basically provide an abstract summary of the topic and present the themes of the document collection. This method has the benefit of analyzing large-scale corpora, without any prior knowledge for the news articles. Further on, we focus on analyzing articles from a single topic. Detecting events within documents on the same topic presents a challenging task, but it is a good starting point to test our research.

2.4 Features Extraction

The topic selected is the 5th largest topic of the dataset, containing 1,509 documents and it centers around motor vehicle accidents. Based on the articles related to this topic, a feature set of key terms is constructed using a combination of Named Entity Recognition (NER) and keyword extraction techniques. NER is essential for event detection as it helps in extracting relevant entities, such as people or places, that are important for interpreting the nature of events. As an example, consider a document's feature set: ”motorcycle,” ”survived,” ”unidentified,” ”Alyki Thasos,” ”park,” ”road,” ”driver,” ”EKAV,” and ”condition.” These terms showcase the key aspects of the event presented in this specific news article. Overall, the goal of developing a feature set is to strengthen the event detection process by using a broader context.

2.5 Clustering

In the next step, the news articles are clustered using the previously constructed feature sets for each document. This process involves several stages, starting with processing the feature sets using two pre-trained LLMs, the Bidirectional Transformer Model (BERT), and a language model from the spaCy library. Furthermore, the feature sets were given as input to a Variational Auto-Encoder (VAE) neural network. This way the feature sets are reconstructed by learning a lower-dimensional latent representation.

Finally, the reconstructed feature sets were clustered using HDBSCAN, K-means, and Agglomerative Hierarchical Clustering. These algorithms operate on unlabeled text data and use distance similarity metrics to determine the progression of clustering. They are capable of handling varying cluster densities and accommodating different resulting structures. The objective here is to see how these news articles, within the same topic, can be further grouped, leveraging the semantic and contextual information encoded in the feature sets. This process brings closer news that use similar terms, while separate articles with dissimilar terms. Thus, it is expected that each cluster will contain documents that describe similar events, similar accidents.

In evaluating the quality of the clusters, the Davies-Bouldin index, the Silhouette score, and the Calinski-Harabasz index were used. K-means and Agglomerative Clustering, were initialized to search for the optimal number of clusters. Both methods searched for the optimal number of clusters (in range of 2-600), that provide the highest Silhouette score. Once the highest Silhouette score is found, the documents are separated using the optimal clusters number. In the following section, the findings of the above process are presented and discussed.

Table 1: spaCy: Numbers of clusters and evaluation metrics per algorithm.

-	HDBSCAN	Kmeans	Agglomerative
Clusters	229	599	599
Silhouette	0.71	0.68	0.74
Davies-Bouldin	19.54	0.14	0.04
Calinski-Harabasz	752 K	7 bil	85 bil

Table 2: BERT: Numbers of clusters and evaluation metrics per algorithm.

-	HDBSCAN	Kmeans	Agglomerative
Clusters	226	598	553
Silhouette	0.57	0.72	0.74
Davies-Bouldin	1.66	0.05	0.07
Calinski-Harabasz	416	92 bil	237 bil

2.6 Evaluation of Clustering

As shown in Table 1. and 2., results vary in terms of number of clusters and performance.

Table 1., shows that the spaCy model applied to K-means and Agglomerative Clustering found 599 clusters, while HDBSCAN found a much smaller number of 229 clusters. Interestingly, despite K-means separating news articles into more clusters, it achieved the lowest Silhouette score compared to the other methods. On the other hand, HDBSCAN's performance was weaker, based on the Davies-Bouldin and Calinski-Harabasz criteria, suggesting that the separation between clusters may not be significant. In contrast, the Agglomerative Hierarchical algorithm demonstrated superior performance across all three metrics, achieving the highest scores. With 599 distinct clusters, it provided better cluster quality and separation. The evaluation scores it acquired show that it can potentially provide more detailed insights into the topic under inspection.

Table 2., shows that BERT combined with HDBSCAN identified 226 clusters, while Agglomerative clustering stopped at 553 and K-means at 598 clusters. Similar to the previous results Agglomerative clustering performed better across all three evaluation metrics, with BERT model. K-means followed Agglomerative clustering in terms of performance, achieving relatively good scores in all three categories. HDBSCAN, on the other hand, obtained lower scores in comparison to the other two algorithms. While it identified a smaller number of clusters, results indicate that article groups are less distinct.

By comparing Table 1 and Table 2, we can observe that the choice of pre-trained models did not significantly impact the clustering results. The number of clusters and the scores achieved across all metrics are generally close in both cases. Regarding HDBSCAN, there is a notable difference in performance when using the spaCy model compared to the BERT model. HDBSCAN achieved higher scores in terms of Silhouette and Calinski-Harabasz when using the spaCy model. In contrast, K-means achieved better performance when using BERT. It obtained higher scores across all metrics compared to spaCy model. Agglomerative clustering presents more complex results. The Silhouette score remains the same regardless of the pre-trained model. However, Davies-Bouldin index and Calinski-Harabasz index present some differences. When using the BERT model, Davies-Bouldin index is higher, indicating potentially less optimal clusters, while the Calinski-Harabasz score is higher when using the spaCy model, suggesting better cluster separation.

Next, news article are clustered using a single method and an event detection algorithm is applied to each cluster to extract named entities and relevant context. Figure 2., illustrates how clusters of the above process are drawn in a two-dimensional space.

2.7 Event Detection

The spaCy-based Agglomerative clustering is used, this time with more degrees of freedom as it searches for the optimal number of clusters (ranging from 1-1509), that provides the best Silhouette score. It identified 725 unique clusters, for a Silhouette score of 0.65. Based on these results, documents were parsed to recognize the named entities that are contained in them. For the documents of each cluster, it is examined whether their text is associated with named entities of the following three categories: geopolitical entity (GPE), organization (ORG), and person (PERSON). Thus, each event is tied with one or multiple entities, the entity type, and the event's context. For example:

Cluster 1 with 12 events, in 12 documents:

Event no.1
Event entities (4):
Entity: Fire department EMS, Type: ORG
Entity: Korydallos, Type: GPE
Entity: Saturday, Type: GPE
Entity: Syggrou Avenue, Type: GPE
Event content:
car accident, Fire Department EMS, Saturday, 60-year-old, occurred, 14/1, ending, tragedy, Syggrou Avenue, collision, guardrail, Korydallos, morning hours.

Event no.2
Event entities (2):
Entity: Chalkida, Type: GPE
Entity: Eretria, Type: GPE
Event content: Saturday, noting, tile, eviathema.gr, ESSDE, Saturday, red, Eretria, traffic accident, according to, severe, Eretria, evening, Chalkida.

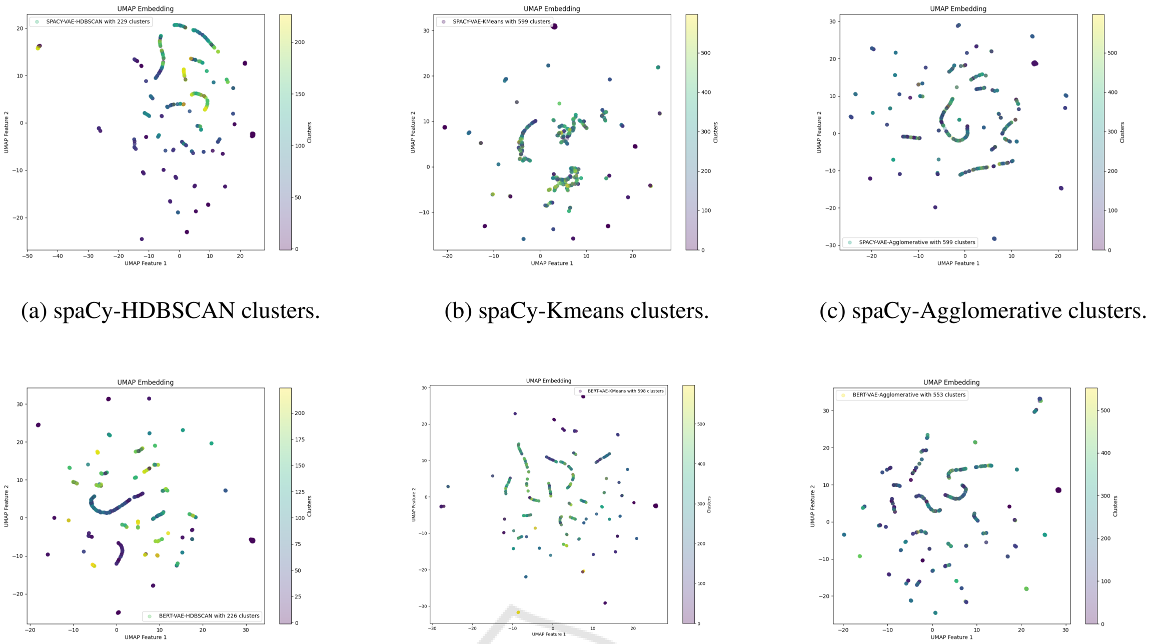
In total, the above process identified 1.455 events, in 1510 documents. Practically each article describes a news event, without knowing how many of these events are distinct. We have only a vague idea of the similarity between events by looking at the clusters in which they appear. The count of entities associated with each event, as well as their type, seems to be random, depending on the keywords used to describe the event, for example, places, or medical services.

Looking at the provided results we can make some observations. Usually, the identity of the people associated with a tragic event is not disclosed in the text. Thus, there are no references to any First or Last name information. Instead other characteristics are used to describe the conditions under which the event occurred, such as the age of the people involved (28 years old, 80 years old), the name, and the type of the street (Kifisias Avenue). Also, the time that the event happened is often described with terms such as ”Saturday night”, ”morning”, ”yesterday afternoon”, and so on. NER misidentified entity categories, placing for example street names in the PERSON category. In total there are 5.375 entities identified, while only 978 entities are of type PERSON, 3.240 entities are of type GPE, and 1.159 entities are of type ORG. This potentially highlights the granularity of the analysis conducted and the challenges faced in identifying entities, and at times, distinguishing between different events. Examining the uniqueness of each event is out of the scope of this research, as we are mostly interested to explore the core elements that constitute news events, such as the keywords used to describe them, and entities associated with each.

3 RESULTS

3.1 Discussion of Findings

From the early stages of the data exploration process, we managed to find the underlying topics that govern the news articles. The 236,029 documents were separated into 1,271 topics and we glanced at the keywords related to each topic. A specific topic, which contains 1,510 news articles was chosen to be investigated further. This topic revolves around the theme of motor vehicle accidents. By following a features extraction technique, we focused on those terms that



(a) spaCy-HDBSCAN clusters. (b) spaCy-Kmeans clusters. (c) spaCy-Agglomerative clusters.

(d) BERT-HDBSCAN clusters. (e) BERT-Kmeans clusters.. (f) BERT-Agglomerative clusters

Figure 2: Clustering distribution comparison.

best describe the article's contents. The feature documents were grouped using a combination of processing and clustering algorithms. The application of the spaCy-based Agglomerative clustering resulted in the identification of 725 unique clusters. These clusters represent distinct groups of articles that share common thematic content within the broader domain.

Grouping related articles into clusters, events can be inferred based on the shared content and contextual information. Additionally, articles clustering can improve the overall information retrieval within the dataset. We can explore specific subtopics of interest by accessing the relevant clusters and examining the articles within them. By capitalizing on the organized representation of the news articles we further analyzed data, detected named entities, and associated them with events. These events can be utilized by an intelligent system, for example a recommendation system or a focused-news application, to track and deliver specific news articles to a wide audience.

3.2 Limitations and Future Work

While our research presents a promising methodology for event detection in news articles, there are several limitations to consider. The accuracy of named entity recognition can be challenging, leading to inaccuracies in event representation. Additionally, the experimentation on automated event detection methods requires a significant amount of data to be parsed and analyzed. Optimizing this process requires a considerable amount of resources. There is not an easy way to evaluate the methodology followed since we are working with unlabeled data, unsupervised learning and there is no golden, ground-truth dataset to compare the performance of our framework.

To address these limitations future work can focus on improving entity recognition accuracy. We would also like to be in a position to properly evaluate our methodology so it can be generalized and expanded in a broader range of domains/topics. Fine-tuning clustering and selecting the optimal number of clusters also needs improvement. It would be interesting to dive even deeper into event detection examining more challenging tasks, identifying unique events, or trying to track event progression in terms of time and developments. Last but not least, we aim to ameliorate the data collection process to include articles metadata, an expansion that will enable us to structure and answer more sophisticated research questions. By pursuing these avenues of future work, our research can make significant contributions to automatically processing news articles, both in Greek and other languages as well.

3.3 Conclusion

In this research, we developed a methodology for pattern exploration and event detection in news articles. We used topic modeling to infer 1,271 topics from a total of 236,029 articles and identified 725 unique clusters for the chosen dataset topic. By examining shared content and contextual information within each cluster, 1,455 events were detected. The proposed methodology demonstrated promising potential for analysis and event detection across various domains, offering valuable insights for researchers and practitioners. Further refinement and enhancements will advance the field of event detection in news articles and the application of automated methods to solve relevant everyday issues.

ACKNOWLEDGEMENTS

This research is supported by the University of West Attica.

REFERENCES

Angelov, D. (2020). Top2vec: Distributed representations of topics. *CoRR*, abs/2008.09470.

Aslan Ozgul, B. and Veneti, A. (2021). The Different Organizational Structures of Alternative Media: Through the Perspective of Alternative Media Journalists in Turkey and Greece. *Digital Journalism*, 0(0):1–20.

Bouras, C. and Tsogkas, V. (2013). Enhancing news articles clustering using word N-grams. *DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications*, (1994):53–60.

Budiarto, A., Rahutomo, R., Putra, H. N., Cenggoro, T. W., Kacamarga, M. F., and Pardamean, B. (2021). Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering. *Procedia Computer Science*, 179(2020):40–46.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Flynn, C. and Dunnion, J. (2004). Event clustering in the news domain. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 3206:65–72.

Jacobi, C., Van Atteveldt, W., and Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1):89–106.

Kapellas, N. and Kapidakis, S. (2022). A Text Similarity Study: Understanding How Differently Greek News Media Describe News Events. *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings*, 2(Ic3k):245–252.

Kirill, Y., Mihail, I. G., Sanzhar, M., Rustam, M., Olga, F., and Ravil, M. (2020). Propaganda Identification Using Topic Modelling. *Procedia Computer Science*, 178(2019):205–212.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., and Androutsopoulos, I. (2020). GREEK-BERT: The greeks visiting sesame street. *ACM International Conference Proceeding Series*, pages 110–117.

Kumaran, G. and Allan, J. (2005). Using names and topics for new event detection. *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 121–128.

Liu, Z., Zhang, Y., Li, Y., and Chaomurilige (2023). Key News Event Detection and Event Context Using Graphic Convolution, Clustering, and Summarizing Methods. *Applied Sciences (Switzerland)*, 13(9).

Papadopoulou, L., Kavoulakos, K., and Avramidis, C. (2021). Intermedia Agenda Setting and Grassroots Collectives: Assessing Global Media's Influence on Greek News Outlets. *Studies in Media and Communication*, 9(2):12.

PATEL, V. and PATEL, A. (2018). Clustering News Articles for Topic Detection. *Iconic Research And Engineering Journals*, 1(11):57–61.

Piskorski, J., Tanev, H., Atkinson, M., and Van Der Goot, E. (2008). Cluster-centric approach to news event extraction. *Frontiers in Artificial Intelligence and Applications*, 181(1):276–290.

Rafea, A. and Gaballah, N. A. (2018). Topic Detection Approaches in Identifying Topics and Events from Arabic Corpora. *Procedia Computer Science*, 142:270–277.

Shah, N. A. and ElBahesh, E. M. (2004). Topic-based clustering of news articles. *Proceedings of the Annual Southeast Conference*, pages 412–413.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Xu, D. and Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2(2):165–193.

Zhang, Y., Guo, F., Shen, J., and Han, J. (2022). Unsupervised Key Event Detection from Massive Text Corpora. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2535–2544.