Creating an Agglomerative Clustering Approach Using GDELT

Oskar Emilius Buserud Jahr

Supervisor: Andreas L. Opdahl

Master's Thesis
Department of Information Science and Media Studies
University of Bergen

June 1, 2023

Acknowledgements

I want to thank Prof. Andreas L. Opdahl, my supervisor, for the crucial help he has provided as I have taken my first steps in academic research and writing. He has helped immensely with his knowledge of the field, and interest in the task at hand. I also want to thank Sara, my dear friend, for all her support through this journey.

Oskar E. B. Jahr
Bergen, 01.06.2023

Abstract

GDELT is a project with a large scale, continuously updated databank that provides a real-time image of the global news picture by outputting these as files that can be downloaded and used by anyone. However, this data is of low granularity, and each source of data does not provide much information on its own. This thesis attempts to leverage the large amount of data available by utilizing a Hierarchical Agglomerative Cluster method to identify news articles that report about the same real life event. To do this, the thesis also explores if the GDELT data is granular enough to be used without extensive preprocessing, and if a distance metric for the cluster algorithm can be created. The findings show promising results when regarded with qualitative measures, but the quantitative measures are not yet optimized. Inherent flaws in GDELT and clustering algorithms are a hurdle to be overcome before the real potential of GDELT's data can be unleashed, and this thesis will explore some of these difficulties and make recommendations for how to circumvent them in future works.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

The Global Database of Events, Language, and Tone (GDELT) is an expansive and promising platform, providing a repository of news data that is both broad in its geographic scope and deep in its temporal reach. The promise of GDELT lies in its potential to offer a comprehensive perspective on global events, providing insights not only into the events themselves but also the narrative structures that surround them.

Despite the richness of this data, its full potential for broader applications remains largely untapped. Most studies employing GDELT have focused on specific, narrow data extraction, often limiting their scope to specific regions, time periods, or topics. While these studies have undoubtedly contributed valuable insights, they have also somewhat constrained our understanding of the full potential of GDELT.

One of the key hurdles in fully realizing this goal lies in the nature of the data itself. GDELT's data is of low granularity, meaning that individual data points, or even small groups of data points, often provide limited information beyond the raw data. This can make it difficult to extract meaningful insights from individual data lines or small data groups. But here lies an opportunity: while low granularity may limit the utility of individual or small groups of data points, the enormity of GDELT's datasets suggests that there is potential to uncover richer insights when looking at larger collections of data points. This thesis will seek to uncover some of these insights, and create richer, better supported representations of news events than is currently available. Indeed, by aggregating and analyzing data at a higher level, it may be possible to discern patterns, trends, and groupings that are not apparent when looking at individual data points or smaller groups of data. These type of insights are often gained by utilizing machine learning, and such is the case in this thesis too.

This thesis aims to tackle this challenge head-on, by developing a hierarchical clustering algorithm that can segregate issues known to be distinct from each other and from other data points. These clusters will then be analyzed by quantitative and qualitative measures to ensure the results of the clustering. In essence, this thesis attempts to utilize GDELT as a database from which data can be gathered without much fine tuning for a broader spectrum of topics, and then utilize this data to find relevance to other data, possibly highlighting relevance between real life topics.

However, this endeavor is not without its challenges. GDELT's dataset sometimes appears spotty, with occasional sparse data, limited duplication control, and occasional over-reporting of events. These issues can produce outliers that might disrupt a machine learning algorithm to the point of being unusable. Therefore, this thesis will also focus on constructing robust metrics for the algorithm, which can withstand noise in the data, yet flexible enough to incorporate results that do not conform to a strict format.

In conclusion, this thesis aims to unlock the broader potential of GDELT by overcoming its inherent challenges and leveraging its strengths. By doing so, it hopes to demonstrate the full value of GDELT as a tool for understanding our complex, interconnected world.

1.2 Motivation

This project captivates my intellectual curiosity primarily because it offers an intriguing opportunity to delve into what was once considered a set of unconnected data. The process of discovering underlying patterns and connections within such data is indeed a compelling task that fuels my academic interest.

One of the most stimulating aspects of this endeavor will be the chance to develop a product from inception to completion. The idea of taking an initial concept, nurturing it through its embryonic stages, and eventually witnessing it evolve into a fully functioning algorithm, carries a sense of fulfillment and intellectual gratification. This developmental journey, in many ways, mirrors the academic process: beginning with a question, traversing through hypotheses, and culminating in valuable knowledge.

The utilization of the Global Database of Events, Language, and Tone (GDELT) further enhances the allure of the project. GDELT, with its vast range of applications, has been somewhat underexplored to date. Its potential to provide profound insights into a plethora of fields is enormous, yet not fully tapped. This project, therefore, presents an exciting opportunity to unleash the unexploited potential of this vast database, contributing not only to my research but also to the broader academic and technological community.

Furthermore, engaging with GDELT provides a secure environment to explore intricate topics such as machine learning and large data sets. While GDELT does have certain granularity issues, it also possesses what can be referred to as 'low-hanging fruit'. These are easily approachable elements that can be swiftly addressed, facilitating swift progress especially in the preliminary stages of this Master's thesis.

Through this project, I'll not only be able to provide a useful tool derived from the underutilized potential of GDELT but also gain hands-on experience and insights into machine learning and large data set handling. This venture promises to be an enriching journey that will contribute significantly to my professional development and provide valuable outcomes for the larger community.

1.3 Research Questions

1.3.1 RQ1

Is it possible to use GDELTs collected data to accurately identify news articles that report about the same real life event?

1.3.2 RQ2

Does GDELT feature data that is granular enough to be used directly in a Machine Learning Algorithm without extensive preprocessing?

1.3.3 RQ3

Is it possible to create a distance metric for the clustering that can account for differences in region, actors, action codes as well as being robust against outliers?

1.4 Research Contribution

This thesis is focused on a detailed exploration of the GDELT project, aiming to extend its use by developing a novel clustering algorithm. The primary objective is to enhance the utility of GDELT's dataset by identifying more intricate connections between GDELT-events, thereby providing a more thorough and complete picture of the real life events that are being reported on in the news media.

1.4.1 Innovation and Novelty

The proposed clustering algorithm, although inspired by existing techniques, will be adapted and refined for the GDELT dataset's unique attributes. This tailored algorithm design aims to ensure the efficient handling of large-scale data without compromising on the quality of insights generated. While acknowledging the complexities of such a task, this thesis is prepared to undertake this challenge, striving to make a notable contribution in the domain of large-scale data analysis.

1.4.2 Modularity and Flexibility

An integral part of this thesis is to design the clustering algorithm such that it can generate relevant datasets based on user-defined search terms. This approach introduces a new layer of flexibility in how GDELT's data can be utilized, offering researchers a more tailored experience. It is hoped that this aspect of the project will broaden the range of feasible research questions that can be explored using the GDELT dataset.

1.4.3 Differentiation from Previous Works

This thesis aims to distinguish itself from previous studies through its large-scale utilization of the GDELT project. The intention is to embrace this large-scale perspective wholeheartedly and acknowledge the challenges it presents. Consequently, this thesis will place a strong emphasis on the development of robust solutions to ensure the reliability and validity of the insights gained from the extensive application of GDELT.

1.4.4 Potential Contributions and Future Directions

The outcomes of this thesis will shed new light on the potentials and limitations of using GDELT at a larger scale. Beyond developing a novel clustering algorithm, this work will hopefully inspire future research into the large-scale application of GDELT, providing valuable reference points for methodologies and offering a new perspective on how global events can be understood and interpreted through the lens of this rich dataset.

1.5 Thesis outline

After the introduction, the thesis structure is as follows.

Chapter 2: Background

Gives a theoretic overview of the GDELT Project, Python, Machine Learning, Clustering, Word Embedding, as well as conducts a litterature review for an overview of the field.

Chapter 3: Research Method

Discusses the necessary components to complete the clustering pipeline, and this thesis, from start to finish.

Chapter 4: Data Preparation

Describes the data structure of GDELT, as well as taking steps in preparation of the clustering, to ensure that the cluster results will be as reliable as possible.

Chapter 5: Clustering Approach

Builds up the clustering process step by step, to show incremental results. Does so by introducing new variables and processes to the cluster metric, adding nuance and separation of the clusters.

Chapter 6: Results

Analyses internal and external metrics of the clustering that has been performed.

Chapter 7: Discussion and Evaluation

Discusses the results given in the chapter before, considers the research questions and to what degree they have been fulfilled.

Chapter 8: Conclusions and Future Work

Conducts a brief overview of the thesis as a whole, mentions possible improvements to similar projects in the future.

Chapter 2

Background

2.1 GDELT

2.1.1 A Background on GDELT

Incepted in 2011, the Global Database of Events, Language, and Tone (GDELT) program constitutes one of the most comprehensive publicly accessible databases purposed for surveilling and scrutinizing societal behaviors and beliefs on a global scale. The primary objective of GDELT is to meticulously quantify global phenomena in near-real-time, thereby tracking societal patterns worldwide and furnishing a distinctive asset for both social science research and strategic decision-making processes, *Leetaru and Schrodt* (2013).

The scope of the GDELT project is international, surveying broadcast, print, and digital news in excess of 100 languages originating from nearly every country. This endeavor is facilitated by a vast array of multilingual news media sources dispersed globally. The project's function lies in discerning the individuals, geographic locations, organizations, thematic elements, sources, and events that act as catalysts within our global society. This process encapsulates over 300 categories of physical and societal events transpiring globally.

Employing an assortment of advanced natural language processing and machine learning algorithms, GDELT carries out tasks of translation, identification, and categorization of pivotal information extracted from global news coverage. With the aid of these sophisticated methodologies, the program cultivates a comprehensive, multidimensional database embodying the intricacy and interconnectivity of global events, narratives, and perspectives. Consequently, GDELT proves itself as an efficacious instrument for understanding the swiftly evolving sociopolitical fabric of the world.

The sheer magnitude and ambition encapsulated within the GDELT project

set it apart. Its broad scope offers a holistic view of global society, integrating not only salient events but also the contextual nuances that shape them. Owing to its real-time monitoring capabilities, GDELT affords a near-instantaneous panorama of global events, thereby rendering it an invaluable resource for journalists, researchers, policymakers, and all those with a vested interest in deciphering global dynamics.

Despite its monumental potential, the GDELT project is not devoid of challenges. Given its reliance on automated text analysis and translation, it remains vulnerable to the inherent limitations and potential inaccuracies of these technologies. Furthermore, its dependence on media data may inject bias, considering that media representations of events frequently deviate from the actual occurrences. Additionally, the prodigious volume of data generated by GDELT can pose significant challenges in filtering and interpreting the information in a meaningful manner.

2.1.2 GDELTS News Aggregation

The Global Database of Events, Language, and Tone (GDELT) boasts a constantly updated repository of news articles concerning events culled from more than 65 languages and several hundred news sources. These are added to their database every 15 minutes through the operation of their web crawler, which processes each article into one or more GDELT events utilizing a natural language processor. The processor is constructed using Textual Analysis by Augmented Replacement Instructions (TABARI), enabling them to distill articles to their most elemental constituents *Leetaru and Schrodt* (2013). Evidence of this process can be seen in their aggregation pipeline, which initially eliminates textual URLs, phone numbers, email addresses, and non-ASCII characters. Subsequently, they subject the entire article to full-text geocoding, which resolves any ambiguities related to geographic references in the text.

Following this, they generate four iterations of the text that are juxtaposed, designed to be comparable in a manner that mitigates different sources of errors and optimizes processor efficacy. The first iteration is simply the raw text. The second iteration substitutes all mentions of geographic landmarks with the name of the country they reside in. This ensures that, although GDELT may not retain a comprehensive summary of every city and town globally, the event will be geolocalized to the appropriate country. The third iteration substitutes all personal names with the country they are believed to originate from, ensuring that, in the absence of a complete registry of global citizens, articles will be attributed to the correct country. Finally, in the fourth iteration, the strategies of the second and third iterations are merged, substituting every name with "Person of Country", and every city or landmark with "City, Country". As a result, a sentence such as "Støre went to Bergen this weekend" remains unaltered in the first iteration, morphs

into "Støre went to Norway this weekend" in the second, becomes "Norway went to Bergen, Norway this weekend" in the third, and finally turns into "Støre of Norway went to Bergen, Norway this weekend" in the fourth iteration. Each iteration ensures that multiple layers of semantic meaning have been preserved for the processor.

An essential decision in the natural language process is the exclusion of most news related to sports. This decision stems from the inherent challenges associated with the often aggressive language used in sports reporting, leading to statements such as "Messi declares war on Real Madrid" or "Manchester United steals Ronaldo". Such language could potentially result in false positives for a processor not specifically fine-tuned to interpret sports news. Therefore, the team responsible for GDELT opted to remove almost all sports news from the raw data processed by GDELT, to minimize the influx of false positives arising from the sometimes strong language prevalent in sports reporting.

2.2 Python

Python, a high-level, interpreted programming language, was conceptualized by Guido van Rossum and made its initial appearance in 1991, *Zhang* (2015). The language was architectured with an emphasis on the legibility of code, and its syntax enables programmers to use fewer lines of code compared to languages such as C++ or Java. It accommodates procedural, object-oriented, and functional programming paradigms.

The design philosophy underlying Python focuses on code readability through extensive use of indentation. The language constructs and the object-oriented approach are made to aid programmers in making clear, logical code for projects spanning all scales. Python is dynamically typed and garbage-collected, indicating that it conducts type checking at runtime and possesses automatic memory management, thus removing the need for developers to manually allocate and free memory within the code.

Python finds a great score of applications across diverse computing and technology domains, encompassing web and game development, data science, artificial intelligence, machine learning, and scientific computing. It is celebrated for its expansive ecosystem that encompasses a substantial collection of libraries and frameworks employed to expand its functionality. Some of the most renowned Python libraries include NumPy for numerical computations, Pandas for data manipulation and analysis and Matplotlib for plotting and visualization. In addition, the library sklearn will be extensively used for machine learning in the code for this thesis.

Furthermore, Python boasts a large and vibrant user community that contributes to its evolution and maintains an array of open-source libraries and frame-

works. The community also provides ample support via online resources, thereby simplifying the process for developers to resolve issues and learn best practices.

2.3 Machine Learning

In recent years, machine learning has become a valuable tool for information extraction and transformation. These algorithms can learn patterns and relationships in large data sets, allowing for the extraction of useful insights and the transformation of low-level data into higher-level features for various applications.

Machine learning methods can be generally divided into three main types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning, the most common type, requires labeled data (input-output pairs) to train the model. The algorithm learns to associate input features with corresponding outputs, which allows for predictions on new data. Reinforcement learning uses an agent that learns to make decisions based on the cumulative reward it gets from the environment. Unsupervised learning, the focus of this thesis, involves learning from unlabeled data. The algorithm tries to find underlying patterns and structures in the data.

Unsupervised learning has been getting more attention recently due to the availability of unlabeled data and the often high cost of getting labeled data. However, while unlabeled data is less costly to get, it's often more complex to process, and the results from an unsupervised learning algorithm may not be as accurate as those from supervised learning.

A key category in unsupervised learning is clustering. Clustering involves grouping a set of data points into clusters based on their similarities. The main goal is to maximize the similarity within each cluster and minimize the similarity between different clusters. Clustering techniques can be divided into hierarchical and partitioning methods. Hierarchical clustering creates a tree-like structure of nested clusters, while partitioning methods divide the dataset into separate clusters. Common clustering algorithms include K-means, DBSCAN, and hierarchical clustering, such as agglomerative clustering, which is used in this thesis.

Clustering has been used in many areas, such as customer segmentation in marketing, anomaly detection in network security, and gene expression analysis in bioinformatics. These applications show the usefulness and importance of clustering techniques in understanding complex, high-dimensional datasets.

2.3.1 Hierarchical Agglomerative Clustering

Clustering is the unsupervised classification of patterns, of data items, into groups (clusters). It has become an important means of data analytics as manual tagging of data is usually expensive, and offers flexibility in that there are many methods and sub-methods to choose from, offering great flexibility to the user *Bouguettaya et al.* (2015). The clustering type used for this thesis is the hierarchical clustering, which builds a tree-like structure known as an endogram to represent data. The building of the tree can either be started at the top, or the bottom of the structure, and for this thesis it will be built bottom-up. This is knows as agglomerative clustering, in our case, Hierarchical Agglomerative Clustering (HAC).

The concept behind HAC is straightforward: each data point starts as its own cluster, and pairs of clusters are merged iteratively in a way that reflects the structure in the data. This merging process is guided by a linkage criterion, which determines the distance between sets of observations as a function of pairwise distances between observations.

Commonly used linkage criteria include single, complete, average, and Ward's method. A single linkage first places all data in its own cluster, and then constructs a list of all inter-data distances. It will then, from smallest to largest distance, iteratively merge the data that is the closest to each other. Single linkage is sensitive to outliers, and as such will not be used in this thesis, but serves as a base case for linkage criteria. The linkage criterion used in this thesis is the complete linkage. This linkage begins with again placing all data in their own cluster, but then takes the two points that are the farthest apart, and separates all values in between using a graph edge*Jain et al.* (1999).

In figure 2.1 there is a cutoff point, at which all nodes leading to the same branch that is intercepted by the cutoff will be placed in the same cluster. As such, it is possible to define the amount of clusters resulting from a HAC algorithm as any number from n, being the amount of data points, to 1, being the topmost branch. Therefore, choosing the right number of clusters is paramount to the accuracy of the results of such an algorithm. The reason for this will be discussed at a later stage in the thesis.

The appeal of Hierarchical Agglomerative Clustering lies in its simplicity, interpretability, and the fact that it does not require specifying the number of clusters a priori. However, it also has certain limitations. For example, once a decision is made to combine two clusters, it cannot be undone. This can lead to suboptimal clustering results. A significant limitation of the agglomerative clustering algorithm is its computational complexity, especially when dealing with large datasets. Computational complexity refers to the computational resources required for an algorithm to run, typically expressed in terms of time or space. The 'Big O' notation is a standard mathematical notation used to describe the worst-case scenario of an algorithm's time or space complexity.

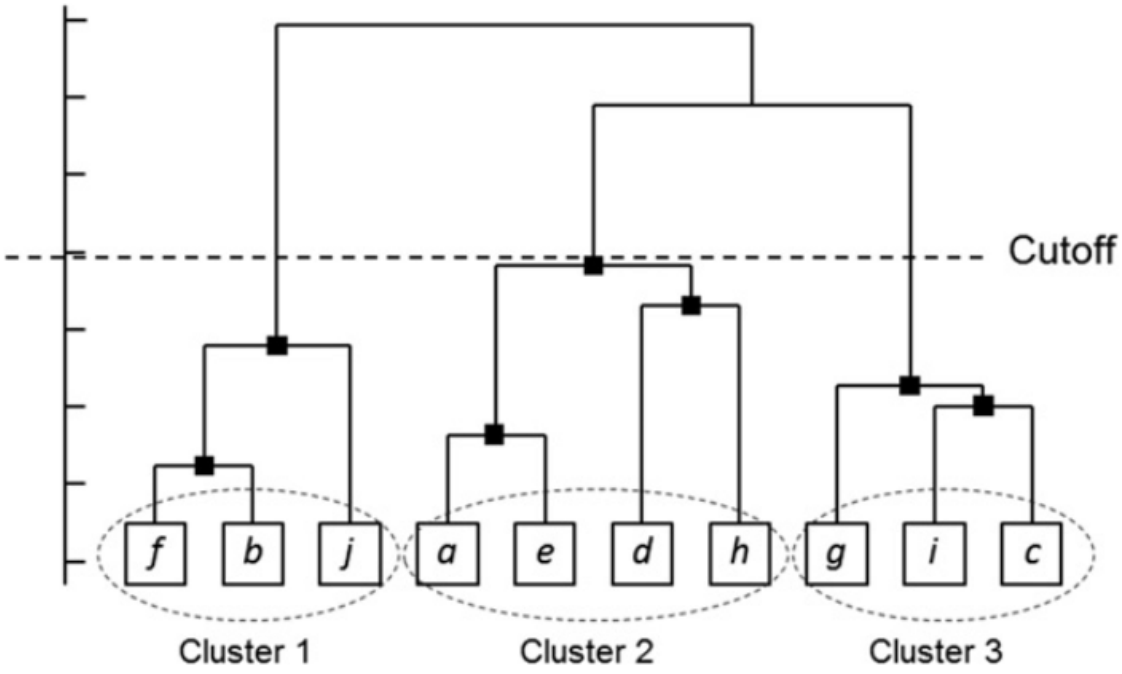

Figure 2.1: Illustration of How the Cutoff on a Dendrogram Decides Clusters, Zhang (2016)

In the case of agglomerative clustering, the time complexity is between $\mathcal{O}(n) = n^2$ and $\mathcal{O}(n) = n^3$. This high computational cost stems from the pairwise nature of the method. The algorithm computes a distance matrix by evaluating the distance between every pair of data points in the dataset. Consequently, the runtime of the algorithm increases dramatically as the size of the dataset increases.

To illustrate this concept, consider two hypothetical datasets. The first dataset contains three events, labeled as event1, event2, and event3. The second dataset expands the first one by adding two more events, labeled as event4 and event5. When the agglomerative clustering algorithm is applied to these datasets, it creates a distance matrix for each dataset. The distance matrix is constructed such that each data point (in this case, each event) is compared to every other data point. More precisely, for each data point 'n', it is compared with each data point 'n+i', where 'i' ranges from 1 to the point where 'n+i' exceeds the size of the dataset.

Importantly, the comparisons are not duplicated. Once the distance between event1 and event2 has been calculated, for instance, there is no need to calculate the distance between event2 and event1 because it is the same. Additionally, the algorithm does not compute the distance between identical data points, such as event1 and event1, because they are identical and thus their distance is zero.

The diagram below illustrates this process. Please note that the calculations avoid redundancy and unnecessary computations, as described above.

Table 2.1: Visualization of Pairwise Function with Three Inputs

Event	1	2	3
1	x	x	x
2	Calculated	x	x
3	Calculated	Calculated	x

This results in 3 calculations for Dataset 1.

Table 2.2: Visualization of Pairwise Function with Five Inputs

Event	1	2	3	4	5
1	x	x	x	x	x
2	Calculated	x	x	x	x
3	Calculated	Calculated	x	x	x
4	Calculated	Calculated	Calculated	x	x
5	Calculated	Calculated	Calculated	Calculated	x

This results in 10 calculations for Dataset 2. Now, while this might not immediately strike one as too much of a computational increase to handle, consider that moving from 3 to 5 inputs resulted in 7 more computations. This trend continues indefinitely, as shown by figure 2.2.

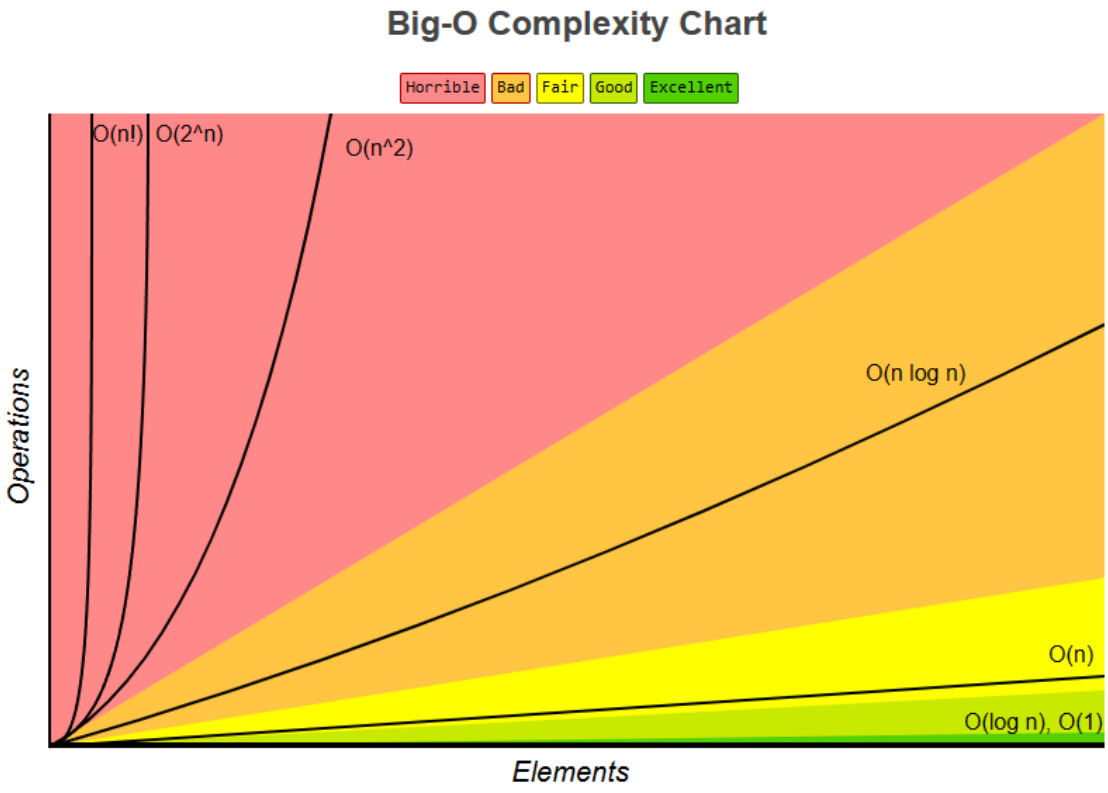

Figure 2.2: Visualization of Some Big O Notations, Rowell (2013)

As shown, the result of this is a heavy, or as Eric Rowell, creator of the graph 2.2 put it, horrible, computational cost when handling large datasets. Despite these drawbacks, the Hierarchical Agglomerative Clustering method remains a popular tool in exploratory data analysis. Its hierarchical nature provides an additional layer of information over other clustering methods, allowing for a multi-resolution analysis of the data. Its versatility and accessibility have ensured its continued relevance in the face of rapid advancements in data analysis techniques.

Distance Metric for HAC

Distance metrics, which are crucial components of clustering algorithms in unsupervised machine learning, provide a means to quantify the similarity between data points, thereby facilitating the creation of clusters. A variety of distance metrics and accompanying heuristics exist, and the selection of a specific metric can notably impact the outcomes of a clustering algorithm.

Euclidean distance, a widely used metric, originates from the Pythagorean theorem. It computes the straight-line distance between two points within Euclidean space. While ideal for numerical and normally distributed data, Euclidean distance may not perform optimally with high-dimensional data due to the "curse of dimensionality." This term describes a phenomenon in which the distances between most pairs of points in high-dimensional space tend to become increasingly similar, resulting in a cluster of data that is not distinct from each other, but forms more of a "blob"*Jain et al.* (1999).

Another distance metric is the Manhattan distance, also referred to as the L1 norm. It calculates the sum of the absolute differences between two points along each dimension. Unlike Euclidean distance, which measures the shortest path, Manhattan distance computes the total path traversed along a grid, proving especially useful in data scenarios where grid-like paths better represent the problem space.

Conversely, Minkowski distance offers a generalization of both Euclidean and Manhattan distances. It incorporates a parameter, typically denoted as 'p', which can be adjusted to produce either of the previously mentioned distances or entirely novel types of distances. This versatility allows for a more personalized approach in defining distances between data points.

Although not strictly a distance metric, cosine similarity frequently finds use in clustering algorithms, particularly in text mining or when dealing with high-dimensional data. It gauges the cosine of the angle between two vectors, effectively capturing the orientation rather than the magnitude of the vectors. This proves valuable in contexts where the direction of data points is more significant than their absolute values, such as in document clustering based on term frequency.

The choice of distance metric should be in harmony with the nature of the data and the specific problem. Various heuristics guide this choice. For example, in high-dimensional data, cosine similarity may be favored over Euclidean distance. For data with outliers, the median absolute deviation might be a superior choice compared to the mean square distance. If clusters are not spherical or have varying sizes and densities, a density-based clustering algorithm like DBSCAN, which utilizes a notion of nearness rather than a strict distance metric, might be more appropriate.

In summation, distance metrics play an integral role in clustering algorithms. They influence the formation of clusters and, consequently, the insights gleaned from the data. Their selection is not a universal decision, but rather requires careful consideration of the data's characteristics, the problem's requirements, and the respective strengths and weaknesses of each metric.

Dimensionality reduction algorithm

Visualizing clusterings in a two-dimensional graph is an effective method to comprehend the structure and relationships within the data intuitively. It can provide a birds-eye view of the distribution of data points and their respective clusters, thereby assisting in understanding the output of a clustering algorithm. It's particularly valuable for high-dimensional data that are challenging to grasp in their raw form.

Transforming high-dimensional data into a two-dimensional representation involves the use of dimensionality reduction techniques. The technique used in this thesis is t-Distributed Stochastic Neighbor Embedding (t-SNE) *van der Maaten and Hinton* (2008). t-SNE is a non-linear method and can therefore capture complex patterns in the data. It reduces dimensions while trying to keep similar instances close and dissimilar instances apart. It is especially well-suited for the visualization of high-dimensional datasets.

The resulting two-dimensional graph allows us to see the clusters as groups of points. Each point corresponds to an instance in the dataset, and the proximity of the points reflects the similarity of the instances. This visual approach can be very informative, providing insights into the number of clusters, their size, and their shape. Color coding can be used to denote different clusters, providing an immediate visual cue about the data's inherent grouping. It's also useful to include the centroid of each cluster in the visual representation, which can be denoted by a distinct marker.

However, care must be taken when interpreting these visualizations as the dimensionality reduction process can sometimes distort the distances between instances. Hence, while 2D visualizations are extremely helpful for gaining an initial understanding, they shouldn't be the sole basis for detailed interpretation of the clusters.

Analyzing clusters

Assessing the performance of a clustering algorithm is essential to confirm that the discovered data structure is significant and not simply a byproduct of the algorithm's inherent randomness or bias. Various metrics and methods are available to gauge the quality of the clusters generated, with different techniques offering distinct insights.

Silhouette score:

The Silhouette Score, a commonly used measure, offers a graphical depiction of how well each object is classified within its cluster. This score is an indicator of how similar an object is to its own cluster relative to other clusters. The Silhouette Score can range between -1 and +1, with a high value suggesting that the object aligns well with its own cluster and poorly with neighboring clusters. If the majority of objects have a high value, the clustering configuration is deemed suitable.

Purity Score:

Purity score is a simple and transparent evaluation metric for clustering, particularly relevant when the ground truth or actual labels of the data points are known. The idea of purity is to assign each cluster to the class which is most frequent in the cluster, then report the percentage of correctly assigned documents.

For each cluster, the class (label) that appears most frequently is considered the correct label for the cluster. The purity score is then the sum of the correctly labeled data points divided by the total number of data points. The purity score ranges between 0 and 1, with 1 indicating that the clusters are perfectly pure. Explained simply, the purity score is what percentage the dominant classification is for a cluster. In this thesis three classifications will be utilized, and as such a purity score of 0.3333 will indicate a completely random clustering, as all three classifications will be equally represented in the cluster.

Although purity is easy to understand and can provide a straightforward interpretation, it has a significant limitation in that it doesn't account for the number of clusters or the distribution of the data points across those clusters, only considering single clusters at a time.

Homogeneity Score:

The homogeneity score is a metric for the evaluation of clustering performance, also applicable when the true labels are known. Homogeneity refers to the extent to which clusters contain only data points which are members of a single class. A homogeneity score of 1 means the clusters are perfectly homogeneous, with each cluster containing data points from only one single class. Conversely, a score of 0 indicates that clusters are randomly assigned without respect to the true labels.

Homogeneity score is a valuable metric for understanding how well a clustering solution respects the distinct categories present in your data. However, it does not measure how completely each class has been assigned to individual clusters. For that, another metric called "completeness" is often used in conjunction. For this thesis completeness will not be utilized, as it performs poorly when there are more clusters than features to consider.

Each of these methods provides a different view on the quality of the clustering results, and the choice of which to use often depends on the specific context and objectives of the analysis. We will further discuss this in chapter 6.

2.4 Word Embedding

In the realm of natural language processing (NLP), word embedding is an essential technique that underpins many machine learning applications. This method is centered around representing textual data by assigning each word or phrase a vector of real numbers. Traditional text representation strategies like Bag-of-Words (BoW) and TF-IDF render words in a high-dimensional space, with the dimensionality being relative to the size of the vocabulary. These strategies treat words as individual atomic symbols, and the vectors that result do not capture the relationships among words effectively.

Contrastingly, word embedding brings the concept of 'similarity' into the domain of word representation. It tackles the limitations of traditional methods by representing words in a dense vector space where words with similar meanings map to proximate points. The dimensionality of this vector space is significantly smaller than the size of the vocabulary, facilitating more efficient computations. A key characteristic of word embeddings is their ability to encode the meaning of a word in relation to other words. For instance, words with similar meanings or contexts have embeddings that are close to each other in the vector space. This feature allows word embeddings to capture semantic and syntactic relationships between words, shown in an example in 2.3.

Various techniques exist for extracting word embeddings from text data. Predictive methods like Google's Word2Vec is popular, but Facebook's FastText is also garnering popularity as a light-weight but quick tool when performing word embeddings. Word2Vec creates word vectors by predicting the context of a word, or conversely, using the word to predict its context. It utilizes both Continuous Bag of Words (CBOW) and Skip-Gram architectures, *Mikolov et al.* (2013a), *Mikolov et al.* (2013b). FastText uses supervised or unsupervised learning, depending on specifications needed, to create their word vectors, *Joulin et al.* (2016)

Word embeddings have significant utility for various NLP tasks, such as text classification, sentiment analysis, machine translation, and information extraction. By transforming words into vectors, we can utilize mathematical opera-

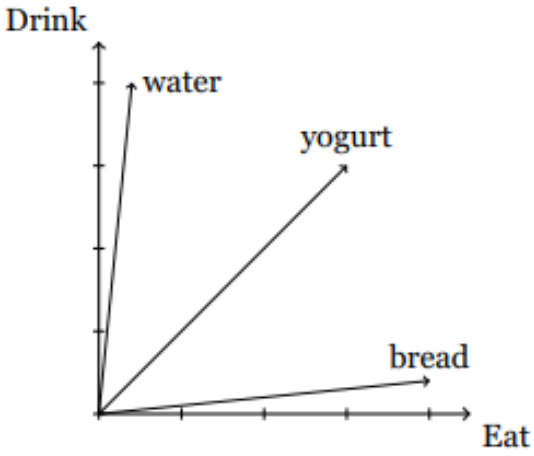

Figure 2.3: A 2D Visualization with 2 dimensions, eat and drink, Winge (2018)

tions to understand and exploit the semantic relationships between words. These vectors can be used as input features for identifying synonyms, grouping similar words, and enhancing the accuracy of machine learning models. In the context of this thesis, word embeddings play a pivotal role in devising a sophisticated method for comparing GDELT events, taking into account the participating actors and their actions. Through the use of word embeddings, we can measure the semantic similarity between actors and actions, thus providing a more accurate and robust assessment of event relatedness.

2.5 Litterature Review

A search was done on the academic articles available through ORIA using the search term GDELT early in the preparative work for this thesis. The findings from this search was that with few exceptions, GDELT has been used as the data for several studies and papers, but in most cases the cases themselves have a narrow lens, which can help disregard the noise that is inherent in GDELT data. In this literature review we will take a look at four articles written between 2016 and 2022, and attempt to highlight how this thesis will fill a gap found in the general literature surrounding GDELT.

2.5.1 Article 1: Predicting Social Unrest Events with Hidden Markov Models Using GDELT

This article, written by *Qiao et al.* (2017), uses the GDELT library as their source for data, and attempts to build a framework that will allow for the prediction of social unrest to allow governments a more effective proactive reaction to such un-

rest. In this paper they use the GDELT provided data as Ground Truths for their model, and provide no additional data to evaluate their findings. In addition, while they show that their Hidden Markov Model outperforms two other models, logistic regression and a baseline model, all their three models use the same dataset, that being GDELT. As such, while the title of the article is interesting for the application of GDELT as a whole, the main focus of their article is to create an effective model, and not necessarily the accuracy of the model put into action in the real world. This is noted in their discussion of their findings Second, we want to add other informative data like Twitter and Facebook to enhance the prediction accuracy. and Third, we also plan to label a Ground Truth dataset for social unrest events in Asia like the Gold Standard Report (GSR) for Latin American[sic] to better evaluate our future methods., meaning that they will attempt in future works to make the model more accurate for a real life scenario, while this is a proof of concept for the model itself.

2.5.2 Article 2: Predicting Social Unrest Using GDELT

While similar in title to the previous article, this paper, written by *Galla and Burke* (2018), instead focuses on using the data they gather from GDELT to produce a model that can be applied to a larger set of circumstances and still prove to be effective. They use an ensemble of supervised learning methods to produce their model, including random forest and neural networks. Differing from both the previous article as well as this thesis is their reliability on the GDELT GKG. The Global Knowledge Graph (GKG) is a large dataset that attempts to grant context beyond what is possible through a normal dataset, by connecting the entirety of its database together as a knowledge graph that will connect all actors, events, locations for each article it analyzes together, before further connecting that article to other articles, creating an interconnected web that spans the entire global media picture.

The paper attempts to predict social unrest on a state level, but also a county level, both of which saw good results from accuracy, precision and recall metric for all methods. When attempting to utilize the model to predict the chance of social unrest a month after its training data ended, the model seemed to accurately predict some protests. However, it is worth noting that the model simply provides a probability of social unrest, and the verification process for the paper to verify if some social unrest took place is that they describe "going through news articles" without further explanation, leaving some room for error in under-reporting as well as missing critical news articles.

As a summary, this paper uses GDELT to produce a model that seems to provide good statistics in its ability to predict social unrest, but the dataset used is still quite small, regarding only social unrest cases in US states, an area of the world that features a large focus from the global news media, which calls into question

the usability of the model on other areas of the world.

2.5.3 Article 3: Analysis of Spatiotemporal Characteristics and Influencing Factors for the Aid Events of COVID-19 Based on GDELT

This article by *Yao et al.* (2022) utilizes GDELT to provide a narrow search of data to be utilized in their research method. They narrow the data in by only selecting data from 4 countries, that being China, the US, the UK and Canada, and then only selecting data that falls within certain action codes, that being the ones focused on aid. After this the process focuses on analyzing the amount of aid these four countries sent to other countries during the COVID-19 pandemic, who their recipients were, and what type of aid was sent over. This was then cross-referenced with other data, like the fact that the countries were more likely to provide more aid to countries with a higher bilateral trade volume, due to their economic interests.

This article uses the data provided by GDELT, but applies it through a narrow lens that will ensure that there is less need for extensive preprocessing and checking for faulty data when applying it to their models and research, which results in them being able to focus on the analysis of the data more than the handling of the data itself.

2.5.4 Article 4: A First Look at Global News Coverage of Disasters by Using the GDELT Dataset

This article, written by *Kwak and An* (2014), focuses on natural disasters, and heavily features the GKG from GDELT to extract these disaster events, along their respective metadata. They cross-reference this data with metrics for each nation like GDP, military expense and population, and attempt to understand what features are the most critical to decide the amount of reporting done on disasters appearing on a global scale. Through a hierarchical multiple regression model they examine what affects global news coverage by gathering a host of different metrics on each country featured in GDELT, then they extract all disaster events from the GKG, and run the model on the amalgamation of these.

Their findings are diverse, and they attempt to neutralize biases of coverage such that the reporting will be as neutral as possible. They then present their findings in the end of the article, but these are not significant for the literature review.

2.5.5 Discussion

While dissimilar in scope, it has proven a challenge to find a large corpus of articles that a. utilize GDELT, b. have a broad scope in terms of event types, and c. feature a global perspective. Either as a result of the scope of the research itself or as a result of difficulties handling GDELT on such a broad application, most articles that utilize GDELT tend to narrow their search in the GDELT database so that the data they are outputting already well suits their need. This is of course a positive, as they do not have to compensate for noise or for statistical outliers, presents the possibility of lost data, as there is little way to ensure that their search terms have not been too narrow, leaving little quality control to make sure they have gathered all relevant information from GDELT. In other words, while they are certainly handling data that is relevant, there is the possibility that they are also missing out on using data that is relevant, but that has been excluded by their search terms.

Chapter 3

Research Method

3.1 Introduction

The primary focus of this study is the exploration, development and application of a clustering algorithm for the analysis of large datasets, specifically, the GDELT project's dataset. This chapter outlines the research methodology designed to guide this process, which integrates computational algorithmic development with quantitative and qualitative data analysis. As this is an exploratory development of an algorithm, it is appropriate to call it experimental design of a cluster algorithm pipeline, including a data-preprocessor and visualizer.

3.2 Literature Review

The study commenced with a small review of existing literature in the field. This review encompassed previous works utilizing GDELT to construct their models. The aim was to understand the current state of research, identify potential gaps, and position this study within the wider academic discourse. The findings from the literature review informed the design and development of the proposed clustering algorithm, as it was found that a only a small subset of articles and research projects utilize the breadth of the GDELT project's dataset.

3.3 Data Preparation

The next step in the research process involves handling the large datasets that GDELT provides to correctly prepare different files for the upcoming cluster algorithm pipeline development. This includes locating relevant news articles to the chosen real life events that will be used as evaluation points, preprocessing

the data in the dataset to a correct format to be utilized in the distancing metric for the cluster algorithm pipeline, as well as sampling the dataset correctly to be used as trial datasets.

3.4 Algorithm Pipeline Development

Subsequent to the data preparation, the research will move into the development phase. Here, a novel clustering algorithm pipeline tailored to handle the GDELT dataset's unique attributes will be crafted. This design process will be iterative, with the algorithm pipeline undergoing regular refinement to ensure efficiency and effectiveness. The focus here will be the management of the distancing metric, as this is a crucial step in designing an unsupervised clustering algorithm pipeline. The stages of the algorithm pipeline's development, from initial conceptualization to eventual coding, will be well documented and explained.

3.5 Data Analysis

With a tested and refined code, the next step will be to evaluate this clustering tool, using both external and internal metrics. The internal metrics will rely on the data inherent to the clusters, while the external metrics will additionally also use data not inherent to the points, such as labeling, will be based on event labels associated with each data point (event) but not used in the clustering process, and a visual inspection of dendrograms and representative graphs will be performed.

3.6 Discussion

The final phase of the research methodology is the discussion of the algorithm and the process overall. This will involve a holistic assessment of the algorithm's performance against its initial objectives. Comparisons may be drawn with traditional clustering algorithms to provide a relative measure of the algorithm's success. Furthermore, a qualitative evaluation of the performance, limitations, and potential improvements will be conducted using graph visualization of clusterings, as well as visual inspections of these.

3.7 Conclusion

By employing a research methodology that combines elements of computational algorithmic development and data analysis, this study aims to make a novel con-

tribution to the field. Through the careful design, testing, and application of a new clustering algorithm, it seeks to enhance the utilization of the GDELT dataset and deepen the understanding of intricate connections between global events.

Chapter 4

Data Preparation

4.1 GDELT News Collection and Data Structure

GDELT collects information from more than 65 languages, and collects these in .csv files that get output between every 15 minutes for GDELT 2.0. These files are run through GDELTs analyzer, and for every article analyzed it gives each event a set of attributes (features). I will include some notable attributes here, as there are 61 attributes for the .export files, as well as 16 attributes for the .mentions files, and writing them all seems superfluous. This system is designed after the CAMEO Code model, *Schrodt* (2012), in which you describe an event as an actor (Actor 1) is doing some action (EventCode) against another actor (Actor 2), *Leetaru and Schrodt* (2015). Then you will have attributes that try to illustrate when this event happened, where it happened, what sort of event it was, and so on.

An example of one event from the GDELT 2.0 database is this

613248743 20170101 201701 2017 2017.0027 ARE
UNITED ARAB EMIRATES ARE
CAN CANADA CAN
0 042 042 04 1 1.9 2 1 2
−5.40045766590389 4 Dubai, Dubayy, United Arab
Emirates AE AE03 28575 25.2522 55.28 −782831 4
Montreal, Quebec, Canada CA CA10 12713 45.5
−73.5833 −569541 4 Dubai, Dubayy, United Arab
Emirates AE AE03 28575 25.2522 55.28 −782831
20170101000000 http://www.princegeorgecitizen.com/news/
local−news/a−list−of−canadians−who−ran−into−trouble−abroad
−in−2016−1.4405000

In the above instance of an event there are some elements that are more eas-

ily understood than others. For instance, certain elements such as 'Actor1' representing the UAE and 'Actor2' representing Canada are straightforward. However, there are instances where information slots in the dataset are left empty, reprsented by underscores, due to the absence of relevant data from the GDELT process. Notably, these gaps are not aberrations, but rather, they highlight the prevalence of incomplete data within the dataset. This inconsistency in data richness across rows could potentially pose challenges for the machine learning algorithm.

One might also notice that while the url, which often mirrors the headline when the article was published, does not specify Canadians in the UAE, but instead Canadian who have ran into trouble abroad generally. This incongruency might stem from the fact that GDELT is not limited to output only one GDELT-event per news article, and in fact frequently outputs multiple GDELT-events per news article, to properly cover all the possible events that the news article has written about.

The dataset employed in this study is bifurcated into two file types, each corresponding to a specific 15 minute period, all throughout the year 2017. The selection of this particular year was driven by the availability of extensive pre-existing data from GDELT and other processors, as well as the future knowledge of events unfurling, leading to a more easily quantifiable analysis of the outcome. Furthermore, the eventful nature of 2017 is expected to facilitate the procurement of event groups A and B.

The 'export' files constitute the primary source of the dataset, encompassing most of the attributes. These files include events captured by GDELT, termed as GDELT-events. A total of 66,327,833 events for the year 2017 are compiled within these files, with March standing out as the month recording the highest number of events, at 6,378,563. In contrast, the 'mentions' files offer supplementary information for the 'export' files, including key attributes such as the 'confidence' tag. This tag reflects the degree of confidence GDELT possesses regarding the accuracy of an event. The 'mentions' files consist of 231,047,178 lines of metadata, each corresponding to a distinct event.

4.2 Choosing News Events

Before the practical development could start, it was necessary to limit the scope of the project. GDELT features over 66 million events in 2017 alone, and GDELT 2.0 has been updating every 15 minutes since 19th February, 2015. A dataset of this scope was considered initially too expansive for this thesis, and as such a process of elimination was conducted on certain criteria to find a sample set of data that would serve as a proof of concept. First of these criteria were the fact that the chosen time span should not be immediately recent. This is because real life events and the reporting on them develop continuously, and in this case being

able to consider most of the direct outcome of a situation was viewed favorably. The year 2017 was chosen for this reason. 2017 is also a good year for a host of practical issues, for example when finding databases for word embeddings.

Next, for the news events themselves. The two events that were picked out were Hurricane Harvey that hit mainland USA in late August of 2017, as well as Brexit, which had been murmuring since 2016 but picked up considerable speed during the summer of 2017. These two real life events were viewed as separate enough that data should not easily favor both cases, which will help keep the data separate from each other in the clustering process. Another facet is that both cases are quite regional, Brexit mainly involving the UK and the EU, and Hurricane Harvey mostly affecting Texas and the US.

To separate data that might be relevant to these two real life events, two techniques were employed. A function was developed that first checked if a GDELT-event had a SOURCEURL that was featured in the references of the main wikipedia page of the corresponding real life event. The wikipedia page for Hurricane Harvey features appx. 300 news articles, while the page for Brexit features appx. 500. Secondarily a word search has been performed in the SOURCEURL itself. For Brexit the only search term was brexit, while for Hurricane Harvey the search needed to identify both hurricane and harvey. Both words were quickly realized to be critical for the function of the process, as if only one was included a large amount of GDELT-events processed from articles regarding either separate hurricanes or individuals named Harvey was included.

4.3 Preprocessing

Preprocessing involves culling the dataset of unrelated data, data that might obscure results, and other elements that will make an accurate clustering algorithm less likely. This involves not only removing columns, but also processing the data featured in the dataset to be more easily managed by the ML-algorithm. Another point mentioned earlier is the need to extract a sample size of the dataset, as the run time of the algorithm would be extensive with the complete dataset. An important aspect of this sampling will also be to provide a balanced amount of events related to Hurricane Harvey, Brexit and otherwise unrelated events in the dataset that will be used for the ML-algorithm. The reason for this is that the algorithm will try to reduce the impact of outliers on the data, and if the amount of unrelated events greatly outnumber the events related to A and B, the possible results will be diluted to the point of no recognition from the software.

Table 4.1: An Overview of GDELT Features

GDELT Feature	Description	Removal Status/Use
GlobalEventId	A unique identifier for each event through the entirety of GDELTs databases.	Stripped after preprocessing, but before clustering. This is to prevent the algorithm from using the integer value of the ID as a feature.
Day, Month, MonthYear, Year	For cataloging events chronologically.	Removed from the data in favor of FractionDate.
FractionDate	Shows the chronology of the event in a given year as a fraction from 0 to 0.9999 computed as (MONTH * 30 + DAY)/365.	Used in the distancing algorithm.
Actor1Code	The complete raw CAMEO code for actor 1, which includes geographic, class, ethnic, religious and type classes.	Removed in favor of Actor1Name, which is more easily handled for word embeddings.
Actor1Name	The actual name of Actor 1.	Used in the distancing algorithm.
For Actor1 and Actor2 CountryCode. KnownGroupCode EthnicCode. Religion1Code. Religion2Code Type1Code Type2Code. Type3Code	Different sub characteristics of Actor 1, these are all deemed too specific to be of special use in this thesis	Removed.
IsRootEvent	Flags if an event is occuring early or late in the source document	Removed.
EventCode	The raw CAMEO action code describing the action that Actor 1 performed on Actor 2.	Used to convert the raw CAMEO code to its description in the preprocessing phase.
EventBaseCode, EventRootCode, QuadClass	Different Data Characteristics of the EventCode	Removed.

Cluster Number	Number of Datapoints	Purity Score
GoldsteinScale	A number from -10 to +10 describing the impact the event will have on the stability of the country, based on the EventCode of the event.	Removed.
NumMentions	Number of mentions of this event across all source documents.	Removed.
NumSources	Total number of information sources containing one or more mentions of this event.	Removed.
NumArticles	Total number of source documents containing one or more mentions of this event	Removed.
AvgTone	The average tone of all documents containing one or more mentions of this event. Ranges from -100 to +100, -100 being extremely negative and +100 being extremely positive, with most events ranging between -10 to +10.	Removed.
Actor1Geo_Type	Specifies the geographic resolution of the match type, and holds one of the following values. 1=COUNTRY 2=USSTATE 3=USCITY 4=WORLDCITY 5=WORLDSTATE	Removed.
For Actor1 and 2 Geo_Fullname Geo_CountryCode Geo_ADM1Code Geo_ADM2Code Geo_Lat Geo_Long Geo_FeatureID	The full name of the location, as well as other subcharacteristics.	Removed.
ActionGeo_Type ActionGeo_Fullname ActionGeo_CountryCode	The same subdescriptors for the Action, i.e. the originator of the EventCode.	

Cluster Number	Number of Datapoints	Purity Score
ActionGeo_ADM1Code ActionGeo_ADM2Code ActionGeo_Lat ActionGeo_Long ActionGeo_FeatureID		Removed.
DATEADDED	The date the event was added to the master database	Removed.
SOURCEURL	The full url of the event	This is used in the distance metric.

Before the data manipulation of specific columns was initiated, several columns were removed from the export.csv file. These have been highlighted in the figure above, together with a brief explanation of the feature(s). These were dropped for a multitude of reasons, most often stemming from the fact that GDELT offered several columns with only small variations. Of these, the most plentiful and easily processable columns were left.

An example of this is in the way dates and date-related data is included. There is the full Date in YYYYMMDD format, MonthYear in YYYYMM format, only Year, then FractionDate, which shows the year as a fraction from 0 to 0.9999 computed as (MONTH * 30 + DAY)/365. Of these, FractionDate was kept as the most relevant statistic for temporality, because it already features a number between 0 and 1, and values closer to each other will represent events that happened closer to each other. This is practical for the purposes of the machine learning algorithm later.

Source URL has been used to categorize data from Hurricane Harvey, Brexit and otherwise unrelated news. This is done by a function checking for two conditions to set an event as related to either Hurricane Harvey or Brexit. First of all it matches the SOURCEURL of the event to a list of URLs that have been scraped from the wikipedia page references. Secondly, it looks for the keywords to the events in the SOURCEURL itself. This is slightly more prone to error as some news sites do not have words in their URLs. Thereafter, the events related to A and B will be identified by adding a column named event_label to the dataset, with a string value of "Hurricane" to signify relevance to Hurrican Harvey, "Brexit" for relevance to Brexit. If the news event is related to neither, it will be represented by the value "Not Relevant" in the column. Strings were chosen instead of integers for easier recognition when inspecting feature values manually.

Thereafter, the dataset is sent over to the preprocessor function. This function first identifies all GDELT-events in the dataset that have the value of N/A to in-

stead feature 0, as the value N/A often causes issues in later functions. It then min max scales the numeric columns. This includes reducing the maximum value of the column to 1, the minimum value to 0, and then, preserving relative distancing, placing all other numeric values in between these. For example the values of 0, 100 and 50 would be reduced to 0, 1 and 0,5. This is done to ensure that one numeric column does not differ in scale from another, reducing the chance of the numeric columns influencing the results differently. This resulted in a dataset with 116 000 GDELT-events, and 10 features, some of which are dropped at various stages of the ML algorithm after no longer serving a function.

The sampling is a simple function that measures an equal amount of events related to Hurricane Harvey, Brexit and not relevant to either, and creates a separate .csv file with these called MLdistributed.csv. This is the dataset that will be used in the ML-algorithm. This reduced dataset has been further stochastically minimized to create several smaller test sets for the distance metric. These test sets range from 500 to 4,000 GDELT-events, thus reducing the required calculations to a more manageable range of 250,000 to 16,000,000 computations.

Chapter 5

Clustering Approach

The clustering approach selected for this study is the Hierarchical Agglomerative Clustering (HAC) algorithm. This particular algorithm was selected owing to several benefits that make it uniquely suited to the task at hand. The HAC algorithm is advantageous due to its simplicity and flexibility. It is not sensitive to the order of data input, consistently yielding identical results regardless of the data sequence. This property is crucial for the current project, as it necessitates a versatile algorithm capable of distinguishing clusters of varying sizes and quantities. Moreover, compared to other partitioning methods such as K-means or k-medoids, the HAC method exhibits superior robustness to outliers.

Another convenience of the HAC approach lies in its straightforward integration with a distance algorithm. Although this study primarily applies this feature at the event level rather than a group level, it can be expanded as needed.

During the initialization of the file used to run the cluster algorithm, there are also two word embedding models that are gathered to be used in later stages. The two models are both trained on a dataset gathered from Wikipedia, with each featuring 300 dimensions, featuring 1 million vectors. This was found to be the most accurate model to be used in this thesis, as other models usually featured too much nuance to give positive effects for the distance metric.

5.1 Characteristics of a Cluster

In approaching the complex matter of clustering, it's essential to define what constitutes a cluster and determine the optimal number of clusters for the dataset. These fundamental aspects have been previously outlined and will now be examined in the context of their practical implementation in this thesis.

For this thesis, obtaining an optimal result in internal or external metrics has not been decided to be critical. The focus of this thesis lies in creating a pipeline

process for a clustering method that proves the possiblity of using GDELT data to visualize connections between news articles on a GDELT-event level, and not necessarily to make such a clustering method as efficient as possible. As such, the silhouette coefficient has been chosen as a metric that will give a value to decide what amount of clusters to use for the rest of the thesis, albeit if this is not the most optimal choice.

This metric provides a graphical representation of how well each object lies within its cluster, it's a measure of how similar an object is to its own cluster compared to other clusters. A high average silhouette score indicates a good clustering solution. The silhouette score is given by using the sklearn.metric function SilhouetteScore, which will do the computation for a silhouette score for the given dendrogram. As seen in figure 5.1, the silhouette scores are quite inconsistent to begin with, but gradually improve as one adds more clusters. This can partly be attributed to the fact that a larger amount of clusters inherently will reduce outliers, an clusters therefore will have a higher degree of centrality. This is both positive and negative for the overall analysis of the clusters, as the silhouette score only improves marginally with a larger number of clusters. This is an issue as the data will provide less granular detail of the dataset at large.

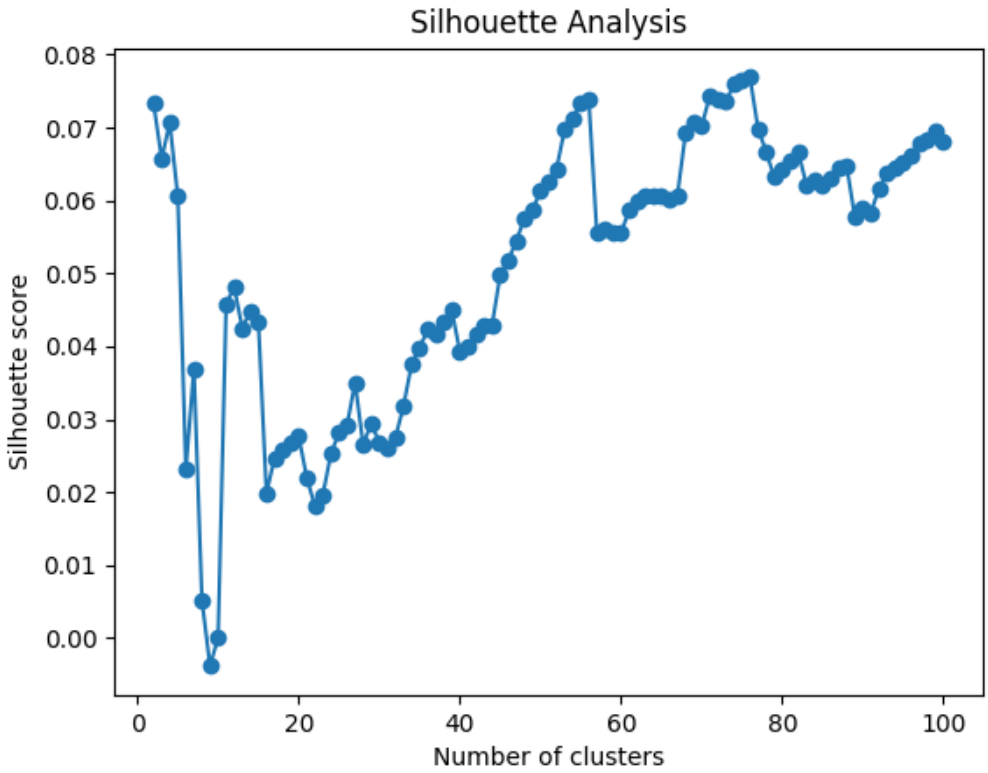

Figure 5.1: A Visualization of the Silhouette Score Increasing with Cluster Amount

5.2 Distance Metric for the Agglomerative Clustering

In developing a distance metric for this thesis, the aim was to satisfy the specific needs of the project while maintaining flexibility to accommodate future, similar applications.

A primary requirement was for the distance metric to be able to differentiate clusters of events with minimal reliance on data completeness. Given that the GDELT dataset occasionally suffers from data incompleteness, a distance metric heavily dependent on all feature values being present for all events could lead to suboptimal outcomes due to the need to exclude incomplete events. As a solution, the distance metric has been designed to depend on a reduced set of features - Actor1Name, Actor2Name, EventCode, FractionDate, and SOURCEURL. These features are most commonly found across all events, making them more likely to yield effective results across a range of use cases.

The second requirement was for the distance metric to operate independently of manual input. Although manual inputs could potentially enhance the performance of the distance metric and machine learning algorithm in the context of this project, the goal is to create a broadly applicable and flexible distance metric. Therefore, all necessary data should be directly sourced from the GDELT events.

The third criterion was to primarily maintain the distance metric at the GDELT event level, rather than elevating it to a grouped level, where one considers grouped events that have been grouped together based on various metrics or premises. This approach not only simplifies the metric's application to other use cases but also explores the potential of deriving high-level insights based solely on individual events. The developed distance metric functions by comparing one event to another, referred to here as event 1 and event 2 for simplicity. The metric involves several steps, each refining its ability to distinguish between different clusters. The following demonstration will illustrate how the distance metric progressively improves in differentiating clusters.

All t-SNE graphs in this chapter not in the red-blue-gray color scheme will have been automatically colored by the visualization function. It applies a color gradient to the clusters, starting with a deep purple at cluster 0, and moving to a bright yellow for cluster 56. This means that when a graph shows nearly all datapoints as one color, and then a few points as a different color, the clustering algorithm has placed nearly all GDELT-events in the same cluster.

5.2.1 Only Identical Source Articles

The foundational assumption of our distance metric is that if two GDELT events are derived from the same SOURCEURL, or news article, they are deemed identical with a distance of 0. Otherwise, they are assigned a distance of 1. This value is provisional and can be adjusted to tailor the robustness of the distance metric, which will be demonstrated later.

This approach solely correlates events based on their common source article, which may not adequately represent the nuances of the task at hand. Importantly, if the SOURCEURL is matching for two given events, the rest of the distancing algorithm will also be disregarded, as they are categorized as a perfect match. As seen in the graph, this results in no tangible clusters, as the clustering will try to simply create small clusters of GDELT-events that have the same SOURCEURL, and otherwise keep the distance to all other clusters at 1.

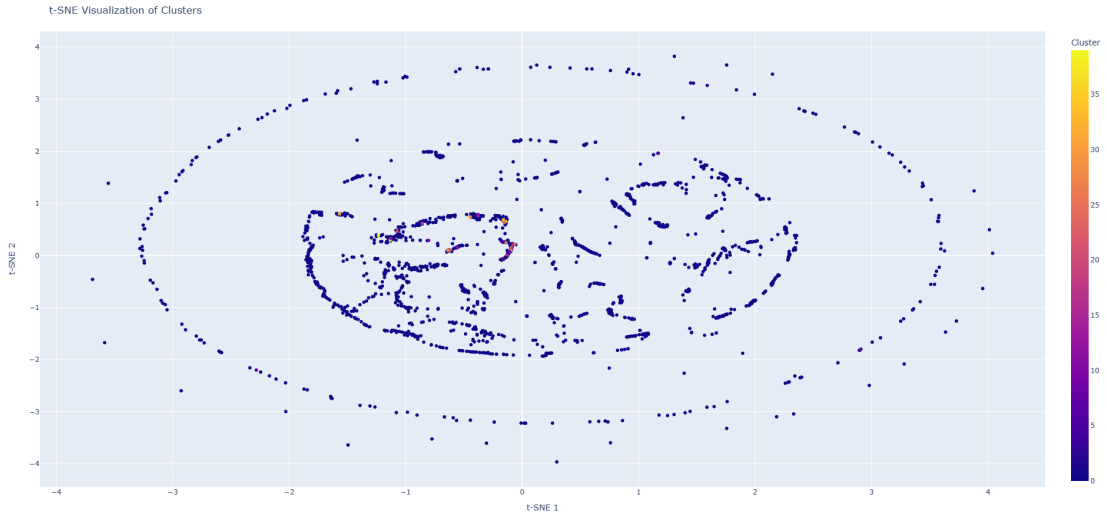

Figure 5.2: t-SNE Graph Only Matching URLs

There are a few characteristics for all graphs following in this section, and later, when visualizing clusterings. First of all is that while this graph is in 2D, the cluster itself is not. As such this is not a true representation of the cluster algorithm, but rather a presentation that is easier to interpret for humans. Secondly, both the shape of the clusters and data points, but also the colors are important. This graph in particular features a clustering algorithm that has put nearly all the GDELT-events in the same cluster, and neither is there any particularly strong collection of clustered data points in the graph. As such we can reasonably argue that only comparing SOURCEURLs is not a valid metric on its own.

5.2.2 Identical Actors

The next level of refinement involves factoring in the actors associated with events. Intuitively, Actor1 and Actor2 seem to be critical parameters to determine event relatedness, which necessitates their incorporation in the distance metric.

Initially, we introduce a negative distancing (indicating closer event grouping) if the same actors are involved. Actor1 from event 1 is compared with Actor1 from event 2, and similarly for Actor2. We then cross-compare Actor2 from event 1 with Actor1 from event 2, and vice versa. The first comparison carries twice the weight of the second, based on the premise that an entity's actions or experiences are more significant than its role reversal. We assign a distance of -10 for sharing the same actor type and -5 for the actor cross examination.

This results in the following graph.

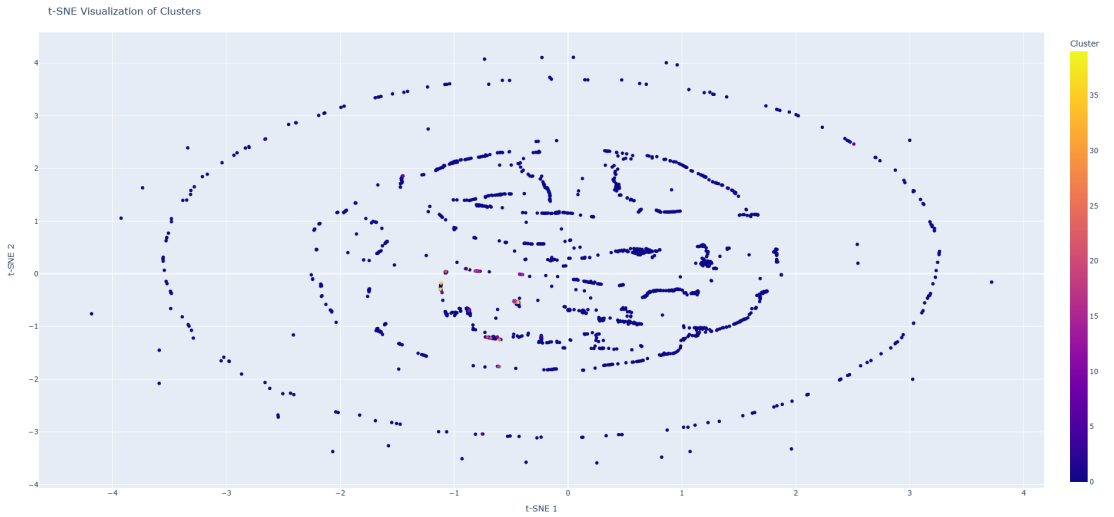

Figure 5.3: t-SNE Graph with Added Comparisons of Actors

Compared to the previous graph there does not appear to be any major changes, neither in location on the graph, nor in coloring/clustering.

5.2.3 Actors, a More Nuanced Approach

A strict measurement of actors still leaves room for better event differentiation. To improve this, we incorporate a measure of actor closeness rather than simply considering absolute identity. We achieve this through word embeddings, generating a vector for each actor and comparing these using cosine similarity. This metric ranges from 1 (complete match) to 0 (no similarity). To align this with our distance scale where 0 indicates proximity, we subtract the cosine similarity from 1. We then apply a multiplier equal to the previous actor similarity measurement. In this case, that multiplier is -10 for the same actor types between events, and -5 for cross-evaluation of actor 1 of event 1 to actor 2 of event 2.

As an example, the final calculation here will then end up looking like 10 - (cosine_similarity(event1_vec, event2_vec) * 10) when evaluating the same actor type between event 1 and 2. The vector model used for this is based on the entirety of Wikipedia in 2017 in the English language, which has provided the overall best results as the thesis also operates in English.

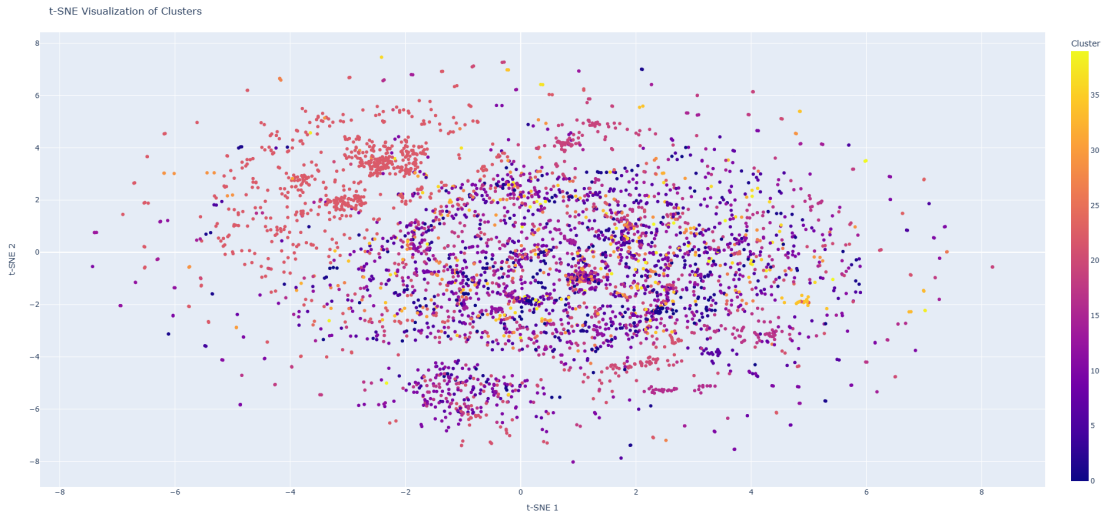

Figure 5.4: t-SNE Graph With Word Embedding Performed on Actors

This graph has started to show some interesting changes from previous graphs. Most noticeable is perhaps that the graph is now a lot more fuzzy, with less distance between points, as well as several groups of data points beginning to emerge. The other noticeable change is in the color patterns. At this point the algorithm is able to cluster together certain data points. However, if one looks for a specific color these can be found in most areas of the graph, indicating a large amount of outliers per cluster.

5.2.4 EventCodes, and How to Measure Them

With the assessment of identical articles and actor nuances completed, we now turn to the integration of event codes, a key component of GDELT's data. During preprocessing, we've converted these numeric codes into their respective descriptions, creating a body of text that can be evaluated using word embeddings for mutual similarity.

However, a complication arises with certain descriptions ending with "..., not specified below," a feature of the CAMEO code structure. As these instances tended to cluster togetherdespite potentially disparate events like "Murder, not specified below" and "Economic loss, not specified below"we decided to trim , not specified below from the respective GDELT event code descriptions. Furthermore, we've eliminated stopwords from the descriptions, a standard practice when employing word embeddings.

The Hungarian or Munkres algorithm plays a crucial role here, its first of two appearances in our approach. The algorithm is instrumental in finding the most compatible similarity between two sentences, rather than settling for the initial match in the similarity matrix. We then implement the same weighting strategy as before, multiplying the resulting average distance between two sentences by a desired weight. For instance, a weight of 2 would signify a weak distancing based on event code similarity, whereas a weight of 10 would indicate a strong distancing. The specific weight used for the following graph and series of graphs is 20.

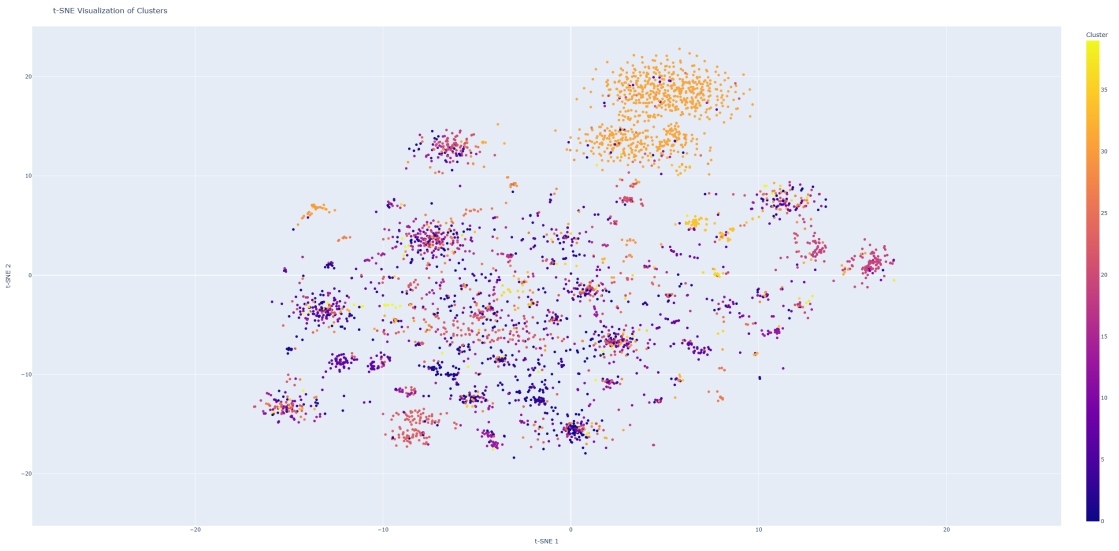

Figure 5.5: t-SNE Graph With Word Embedding over Event Codes

The differences in location on the graph as well as coloring are now starting to take shape, and at this point it was possible to start evaluating the graph itself. While distancing different clusters from each other can be a positive feature, it

might also indicate too heavy of a weight on a few features, in this case, the EventCode has a large impact on the final result.

5.2.5 Some Nuancing Using Dates

To add a certain nuance weighing events more close to each other than events spaced further apart, we use the FractionDate feature from the GDELT dataset in event 1 and event 2. We have min-maxed the column, and as such we simply take the absolute value of the fraction date from event 1, subtracted by the fraction date of event 2. We then multiply this by 2, to give the dating some weight. Adding too much weight results in more noise than nuance, but adding no extra weight is disadvantageous because events that happen closer to each other are more likely to be related.

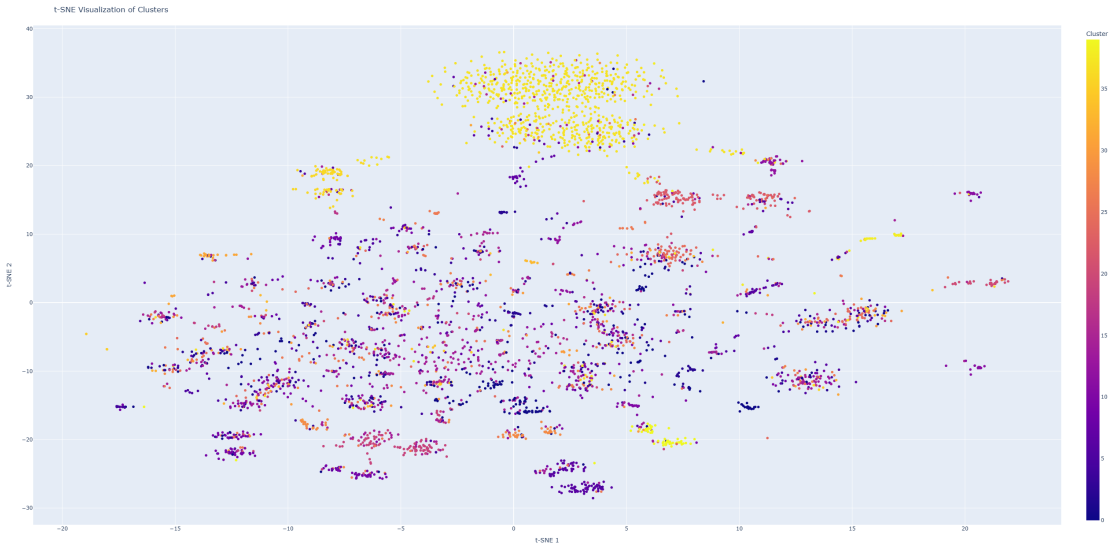

Figure 5.6: t-SNE Graph With Considerations of Dates Events Happened

In accordance with the goals posited when introducing this distance feature, the graph did not drastically change. On the other hand it has generally gained a little sparsity in between data points, giving nuance to the results.

5.2.6 Accounting for Bad Actors

GDELT presents a unique challenge in that each data point can only hold one Actor1 and one Actor2, which occasionally results in an unconventional pairing of actors. Such eccentric choices, if not addressed, could skew the accuracy of our distance metric. For instance, a news article about Jewish businessmen from Israel visiting Texas after Hurricane Harvey was categorized with Actor1 of event 1 as 'JEWS' and Actor1 of event 2 as 'Texas.' Given that 'JEWS' doesn't closely align with 'United States' or 'Texas,' the event could be inaccurately distanced from related GDELT-events.

To mitigate this issue, we've incorporated a non-event-to-event distance metric. Since multiple GDELT-events can be derived from a single news article, we've established a separate CSV file that lists all Actor1Name and Actor2Name entries for each SOURCEURL. This extra layer of comparison extends beyond just the Actor1Name of event 1 and 2, allowing us to also consider their associated Actor1Name and Actor2Name, if present. We have applied only a minor weight to this aspect given that the information doesn't directly originate from event1 or event2, thus providing less assurance about the distancing between these two individual events. The weight given to these similarity measures is a negative weight (closer distancing) of -2 when matching actor 1 in event 1 with actor 1 in event 2, and similarly for actor 2. The weight given is -1 when comparing actor 1 in event 1 with actor 2 in event 2, and similarly for event 1 actor 2 to event 2 actor 1.

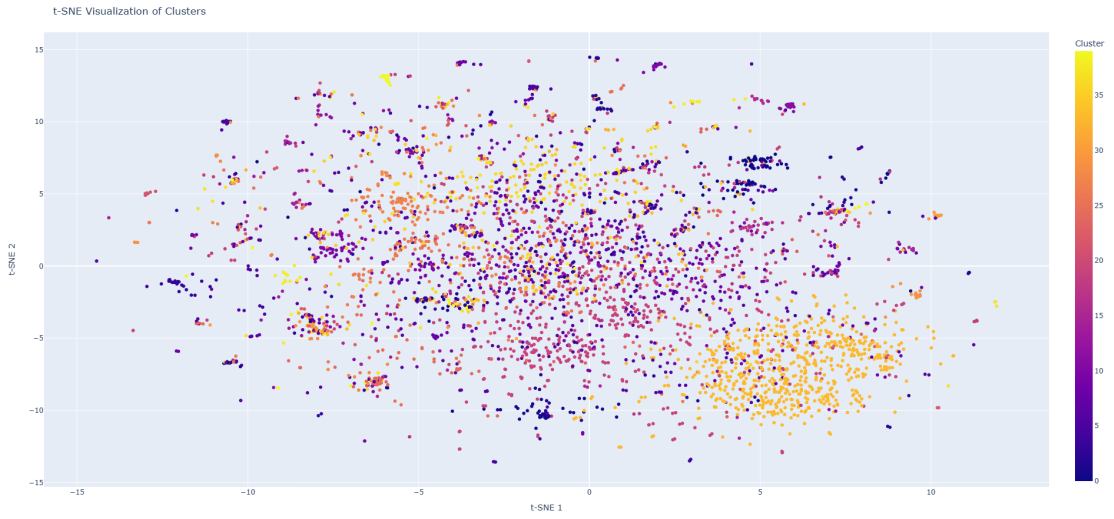

Figure 5.7: t-SNE Graph Accounting for Faulty GDELT Reporting

The graph again at this point looks to be more compact. This possibly stems from GDELTs at times very specific actor classification. To serve as an example, a given news article might have 50 GDELT events extracted from it, and if 49 of these have Actor 1 classified as United States and 1 as Texas, the collection of Actor 1s for this SOURCEURL will finally be (United States, Texas). This can

be mitigated using either the mean Actor 1 for a given SOURCEURL, weighing the results based on number of occurrences or other measures. In this specific scenario the fussiness of the graph is seen as a positive for reasons previously explained, and as such none of the mitigating features were implemented.

5.2.7 Checking for Similar URLs

Another strategy we employ to counterbalance the limitations of GDELT's feature extraction from news articles involves applying word embeddings to the article URLs themselves. In a significant number of instances within GDELT, the URL contains wordsoften the article's headlinewhich can serve as an additional resource for gauging the similarity between event 1 and event 2.

To carry out this analysis, we first deconstruct the URL. This process involves removing all characters that are neither letters nor numbers and treating the remaining elements as separate words. This approach generates many irrelevant words, such as "www" or "https", along with other non-word string artifacts. To address this, we first verify whether these words are in the word embedding model used. Following this, we implement the Hungarian algorithm, previously used with event codes, to assess the cosine similarity between the sets of words drawn from each URL.

However, this facet of the distance metric requires judicious application. A notable problem arises when we assign too much weight to this component of the distance metric: not all article URLs contain words that reflect the article's content. Furthermore, there's a less tangible issue where the headlines or URLs may not accurately represent the content of the article, potentially leading to false positives or negatives regarding event proximity or the lack thereof. Because of these drawbacks this distancing mechanism has only been assigned a small impact on the final result. The weight given to this distance metric is 5.

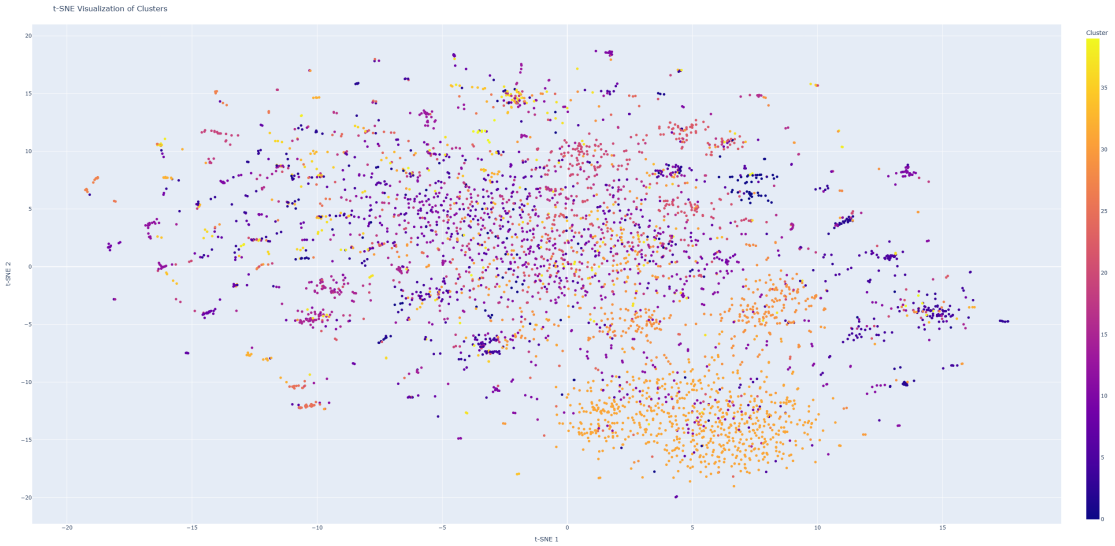

Figure 5.8: t-SNE Graph Using Word Embeddings on URLs

This metric feature distanced the clusters a slight distance from each other again, but also noticeably seems to have separated data points into clearer clusterings based on the more clear separation of color groups in the graph itself.

Chapter 6

Results

6.1 Cluster Analysis

Cluster analysis is a significant analytical tool for understanding the natural grouping in a data set. It involves various algorithms to partition data into sets or clusters where objects in the same cluster are more similar to each other than to those in other clusters. The goal is to segregate groups with similar traits and assign them into clusters.

In this thesis, cluster analysis will be used to identify and isolate related events from a large dataset. The primary challenge lies in the fact that these events need to be related within themselves but distinct from each other and unrelated events. It's a complex task that requires an efficient and robust clustering algorithm.

6.2 Quantitative Analysis

Quantitative measures like the silhouette score are vital to assess the quality of the clustering algorithm. The silhouette score is a measure of how close each point in one cluster is to the points in the neighboring clusters. It's an effective metric for evaluating the degree of intermingled clusters. Other measures like the homogeneity, the Fowlkes-Mallow Index as well as a purity measure will also be implemented to gather a full picture of the efficacy of the cluster analysis and distance metrics. These will more closely relate to the makeup of the clusters themselves.

6.2.1 Silhouette Score

For this thesis, the silhouette score will help gauge the effectiveness of the clustering algorithm in accurately grouping related events. It will provide an objective measure to optimize the machine learning model for the best performance.

Figure 5.1 visualizes the silhouette scores measured over 100 clusters. As seen by the data, the graph peaks at 56 and 76 clusters. To limit the degree at which more clusters do no more than just erase outliers, a cutoff at 56 clusters was chosen for the rest of the analysis as a high point of silhouette scores.

Silhouette score for the whole clustering: 0.0738166623182993

6.2.2 Homogeneity

The homogeneity of the clustering is 0.312122426287609. Considering a homogeneity has a maximum of 1 and a minimum of 0, this is not a stellar score. The reason for this is likely the same as the silhouette score, that is, a poor differentiation between clusters. The homogeneity score is not seen as critical as this thesis does not aim to optimize the clustering.

6.2.3 Purity Score

Average cluster purity, accounted for cluster size: 0.66577778 Average cluster purity, not accounted for cluster size: 0.68148706

Table 6.1: Cluster data and purity scores

Cluster Number	Number of Datapoints	Purity Score
0	620	0.7173913
1	46	0.7173913
2	160	0.6125
3	121	0.40714286
4	48	0.77083333
5	55	0.34545455
6	54	0.48148148
7	140	0.40714286
8	45	0.62222222
9	64	0.5625
10	29	0.93103448
11	42	0.5952381
12	58	0.94827586
13	92	0.77173913

Cluster Number	Number of Datapoints	Purity Score
14	49	0.46938776
15	162	0.7037037
16	35	0.48571429
17	151	0.94701987
18	34	0.94117647
19	170	0.80588235
20	54	0.62962963
21	774	0.60852713
22	29	0.5862069
23	60	0.46666667
24	29	0.55172414
25	91	0.49450549
26	35	0.82857143
27	73	0.52054795
28	25	0.44
29	93	0.41935484
30	94	0.43617021
31	47	0.38297872
32	41	0.7804878
33	49	0.81632653
34	27	0.55555556
35	84	0.38095238
36	15	0.8
37	40	0.925
38	29	0.55172414
39	34	0.44117647
40	17	1.0
41	95	0.94736842
42	44	0.61363636
43	16	0.875
44	96	0.61458333
45	27	0.66666667
46	17	0.88235294
47	51	0.92156863
48	32	0.6875
49	26	0.92307692
50	32	1.0
51	32	0.5625
52	68	0.94117647
53	10	1.0
54	26	0.73076923
55	13	0.76923077

The purity score differentiates itself from the silhouette and homogeneity score in that it can consider both the clusters on their own, but also the clustering as a whole, giving additional insight. The purity score overall is also not exemplary, but here we start to discern some values that are looking more promising. As the score is significantly larger than 0.333, we can extract that the clustering is doing better than simply stochastically assigning data points to clusters. Of specific clusters, there are a few that are worth taking notice of.

Cluster 0 and 21 hold significantly more data points than the other clusters, and cluster 0 in particular also has a high purity score of 0.781, meaning that appx. 78,1% of the data points in this cluster have the same classification. In this case the majority of both cluster 0 and 21 are data points regarding Hurricane Harvey.

Cluster 50 has 32 data points, all of which belong to the same classification, that being GDELT-events unrelated to either Hurricane Harvey nor Brexit.

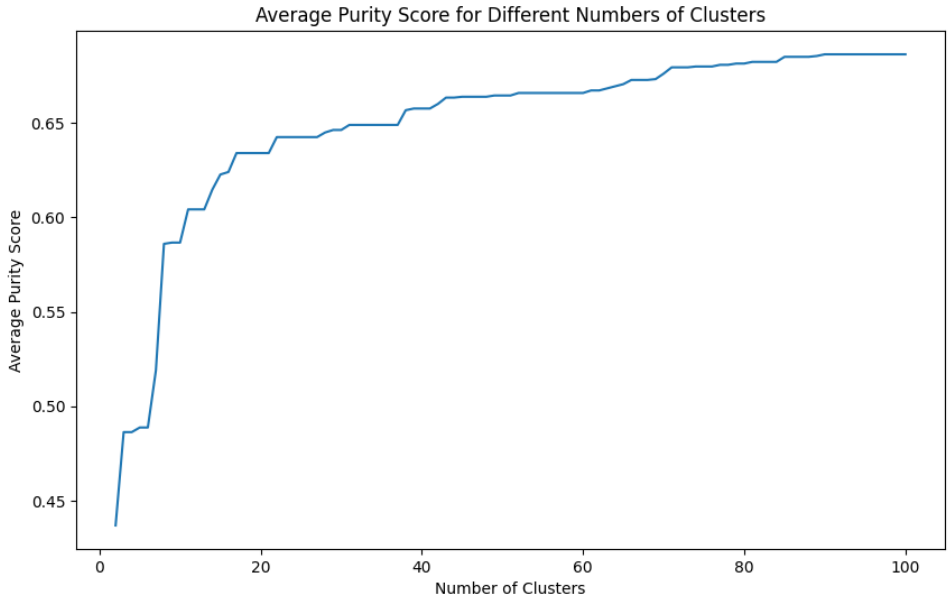

Figure 6.1: Graph Showing Average Purity Score as Cluster Sizes Increase

6.3 Qualitative Analysis and Visual Observations

After the computation of the various metrics, a clustering (depicted by a 2D graph via the application of t-SNE) similar to the one below might emerge. These are randomly sampled GDELT-events; thus, the graph will differ with each dataset, but there are still some broad patterns that can be discerned.

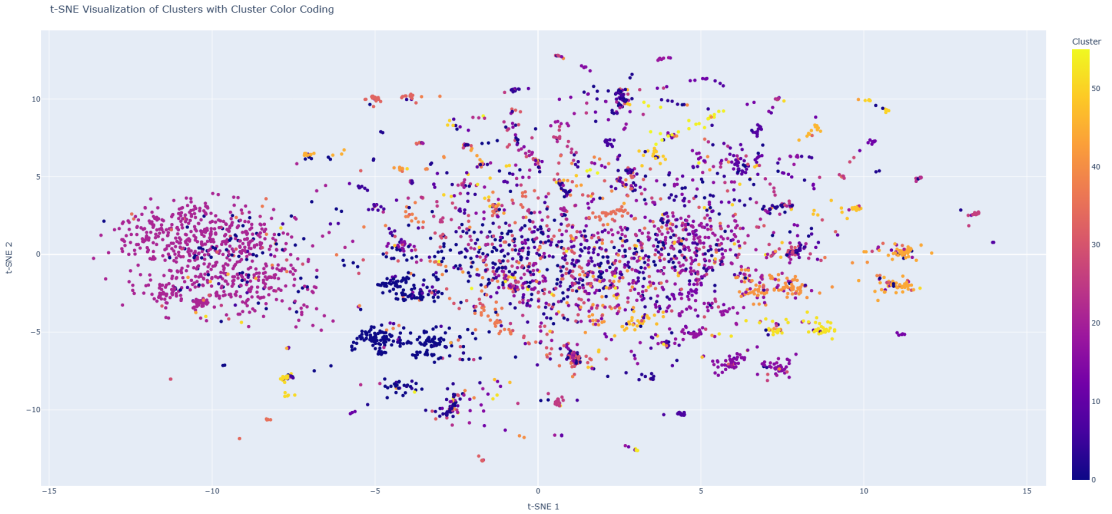

Figure 6.2: Visualization of a Clustering Using t-SNE, Automatically Colored

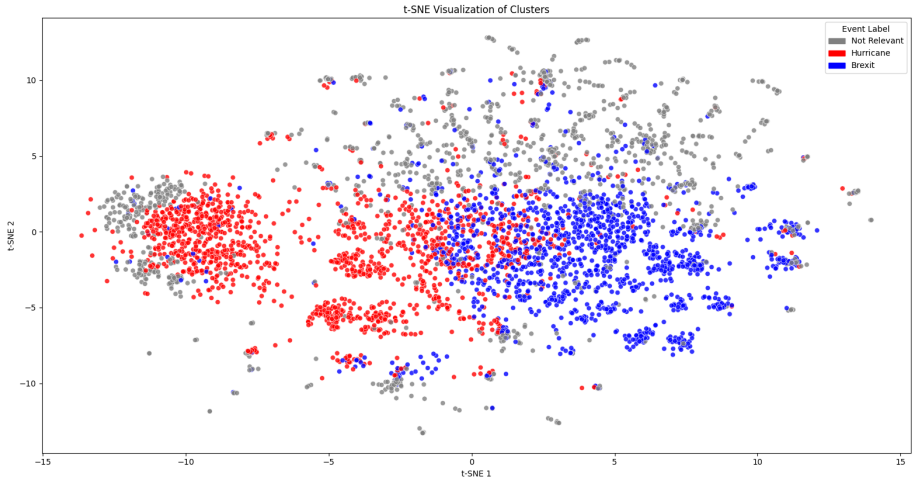

Figure 6.3: Visualization of a Clustering Using t-SNE, Manually Colored to Show Event Relevance

Figures 6.2 and 6.3 illustrate two different color encodings of the same clustering. Figure 6.2 has automatically color encoded clusters as determined by the linkage algorithm and the n_clusters parameter, which in this case is set to 56 clusters, ranging from 0 to 55. Some patterns become noticeable immediately.

However, figure 6.3 has a much more distinct structure, with only three colorations, which have been manually assigned: Hurricane Harvey is red, Brexit is blue, and unrelated events are colored gray. A certain separation of event types from each other is immediately apparent, but a substantial amount of gray mingles with both event groups. This might be because the algorithm does a poor job of separating outliers, but there is also the possibility that the function that labels GDELT-events has not labeled this data correctly. This would be difficult to know without inspecting each specific data point and inspecting the source article itself.

The majority of the clusterings seen in the t-SNE graph can be traced back to the significant distinction of event codes and actors. For instance, the sizable cluster to the left of the main, large cluster is entirely made up of instances where either Actor 1 or 2 is "United States". Moreover, the lower half of that larger cluster is separated from the top half by the fact that the lower cluster features "United States" as Actor1, while the top half has it as Actor2. This clustering around "United States" can be further demonstrated in figure 6.4.

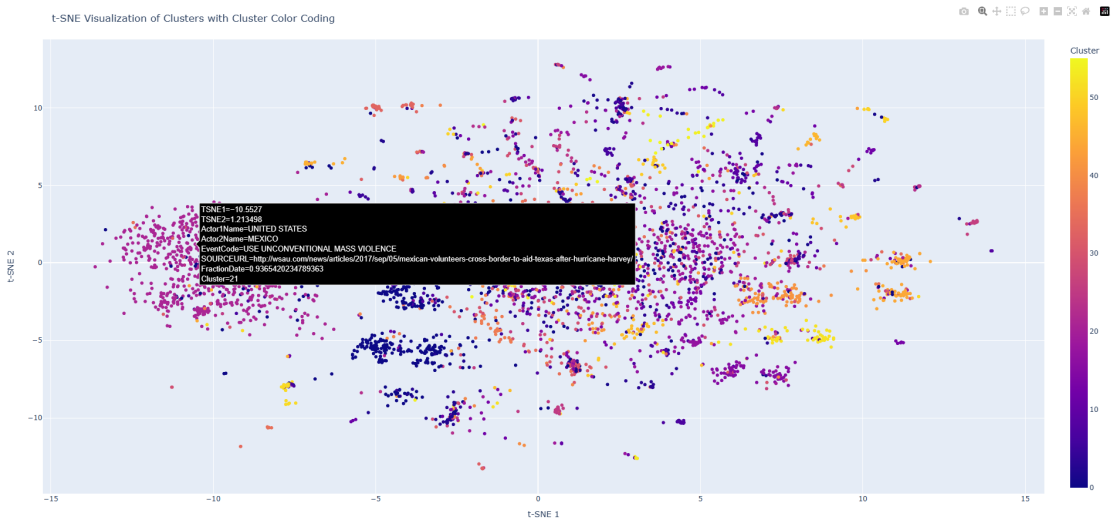

Figure 6.4: Event Data Being Shown, Actor1 is United States

Additionally, figure 6.4 showcases a closer view of the three smaller clusters found beneath the central cluster. The two clusters at the bottom are respectively "TEXAS" as Actor 1 and Actor 2, while the cluster above them is "HOUSTON" as Actor1 or 2. This could indicate that the word embedding is functioning properly, as Houston and Texas should exhibit a close semantic similarity. It is also noticeable that these have been merged to belong to the same cluster.

Figure 6.5: Event Data Being Shown, Actor1 is Texas

Furthermore, the cluster hierarchy has been translated into a 3D graph to better illustrate another form the dendrogram can assume when reduced to a lower-dimensionality graph. The three figures 6.6, 6.7 and 6.8 below present different perspectives. The different perspectives individually highlight either events related to Hurricane Harvey in red, Brexit in blue, and unrelated events in gray. It is immediately clear that there is a general separation of events from each other, but they still exist as parts of the same, large cluster, with unrelated events intermingling.

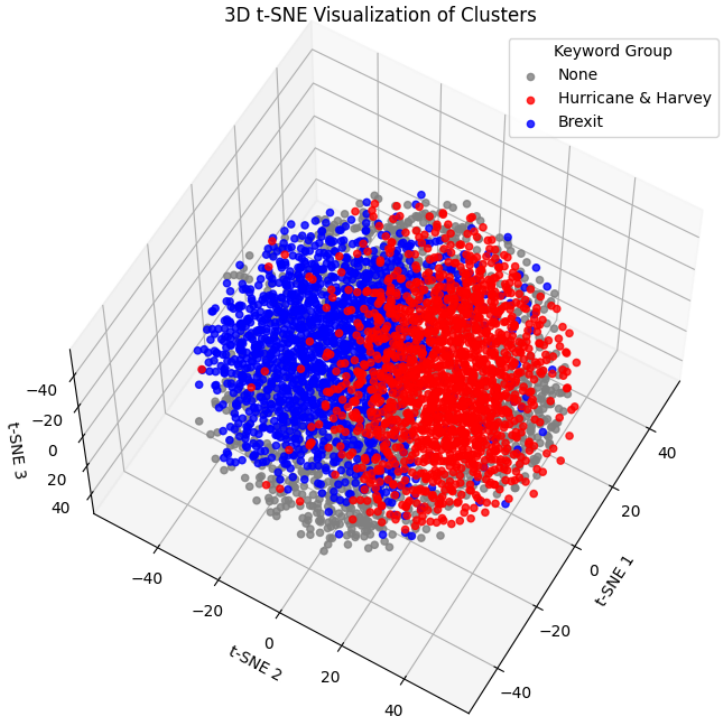

Figure 6.6: 3D t-SNE Graph With a Focus on Hurricane Harvey Related Events

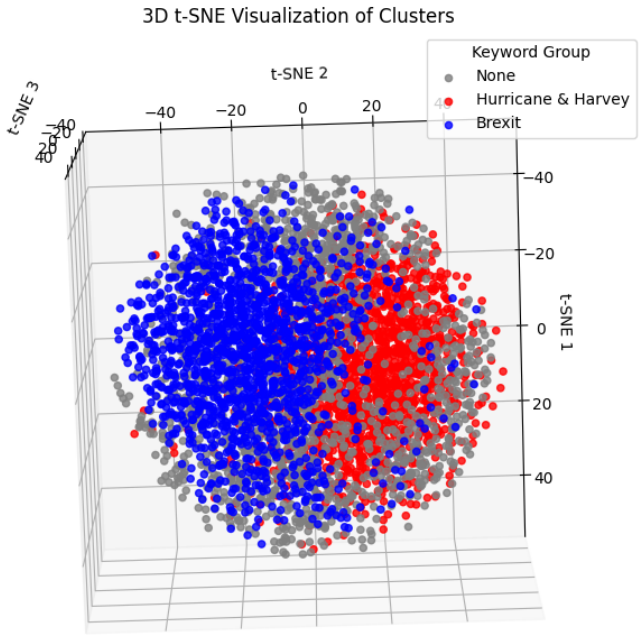

Figure 6.7: 3D t-SNE Graph With a Focus on Brexit Related Events

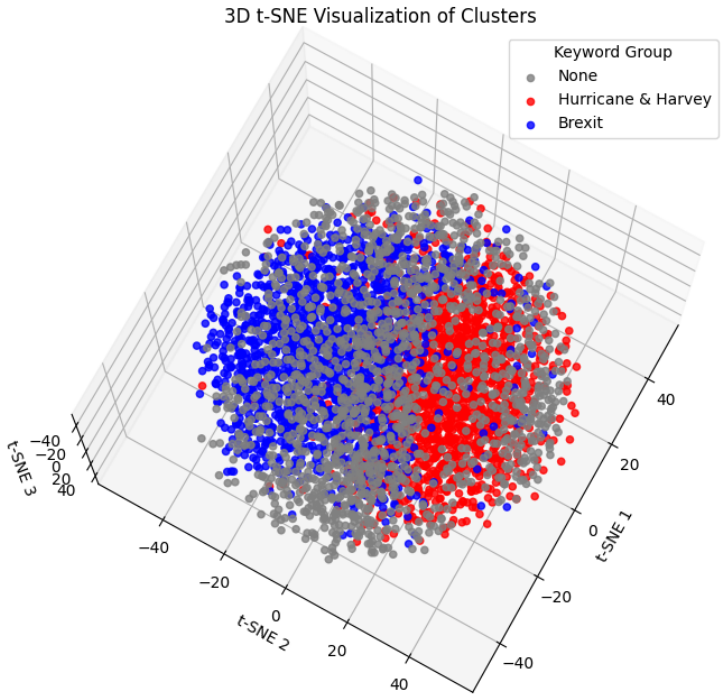

Figure 6.8: 3D t-SNE Graph With a Focus on Unrelated Events

Chapter 7

Discussion and Evaluation

7.1 Research Questions: Interpretation of Results

The results put forth by the quantitative and qualitative analysis paint an interesting picture. On one hand the clustering algorithm struggles with the large amount of noise in GDELT data, resulting in fuzzy and not clearly defined clusters, shown by the low silhouette and homogeneity scores, but the purity score illustrates that while the clusterings are not tight, they at least hold some value, averaging at 67

On the other hand, the graphs, both in 2D and 3D, that have been manually painted according to the labeling of the original GDELT-events paint a less nuanced picture. In these graphs the data points regarding events have formed two large groupings on either side of the central supercluster, while the unrelated events in large part are either intermingling, or have been pushed out to the outer edges.

7.1.1 RQ1: Is it possible to use GDELTs collected data to accurately identify news articles that report about the same real life event?

The first research question of this thesis is quite expansive and exploratory. The overarching aim of the project was to assess if GDELT's collected data could be utilized for accurate identification of news articles about the same real-world event. Regrettably, the results did not strictly meet the criteria of accuracy initially established. Nevertheless, this shortfall is partially mitigated when considering external metrics and visual evaluation of graphical representations and dendrograms, which seem to suggest a level of efficacy that is not fully captured by internal metrics.

While the internal metrics, specifically silhouette and homogeneity scores, did not reflect optimal clustering quality, the external metrics indicated some effectiveness. For instance, visual analysis of graphs and dendrograms revealed meaningful patterns and relationships. However, the silhouette scores suggested that the clusters' delineation was not well-defined, indicating overlapping or weakly separated clusters. Similarly, the purity scores suggested reasonable performance of the algorithm, yet there was still a degree of uncertainty. Specifically, while the purity scores were satisfactory, there was not a strong guarantee that most or all data points within a given cluster actually belonged to the same real-world event.

7.1.2 RQ2: Does GDELT feature data that is granular enough to be used directly in an unsupervised machine learning algorithm without extensive preprocessing?

Another primary aim of this thesis was to evaluate if the GDELT data could be effectively utilized in machine learning algorithms without requiring extensive preprocessing. This is a significant consideration given that numerous projects today heavily rely on extensive preprocessing, particularly in the context of supervised learning. Thus, this research sought to employ minimal and swift preprocessing as a strategy to reduce the computational load. For this project, the most involved preprocessing measures included dropping irrelevant features, and performing operations like min-max scaling on certain columns. These measures were adopted to ensure the data met the assumptions and requirements of the machine learning algorithm.

While the results of this approach didn't achieve exceptional performance, they were nevertheless promising. Specifically, the outcomes surpassed what would be expected from a random assignment of values. Thus, it appears that GDELT data, despite its potential limitations, can be leveraged directly in an unsupervised machine learning algorithm with relatively minimal preprocessing, and still produce better-than-random results. This finding underscores the potential utility of GDELT data in computational applications, even when resources for extensive preprocessing are limited.

7.1.3 RQ3: Is it possible to create a distance metric that can account for differences in region, actors, action codes as well as being robust against outliers?

The development of a versatile and adaptable distance metric was a critical component of this research project. This metric, as discussed in the Research Method chapter, can be modified to change the relative importance of various parameters. The distance metric in its current state, while capable of handling a range of features, relies predominantly on three main features: Actors, Event Codes, and to a lesser extent, the SOURCEURL of the GDELT events.

This heavy reliance on just a few of over 60 features of GDELT 2.0 presents certain challenges. The metric exhibits reasonable proficiency at distinguishing real-world events separated by both geographic distance and unique characteristics. However, its effectiveness dwindles when tasked with differentiating events occurring predominantly within the same region or pertaining to similar types of actions. This limitation is evident in the silhouette scores and the graphical representations of the clusters, which reveal that the clustering is not clearly delineated. Consequently, there is a justified concern that real-world events closely aligned in nature might be inseparable by the algorithm.

Therefore, while the current distance metric has demonstrated some level of effectiveness, its limitations suggest a need for further enhancements. These enhancements should aim to improve its ability to distinguish between closely related events, thereby increasing the robustness and utility of the metric in the application of unsupervised machine learning algorithms to GDELT data.

7.2 Limiations and Assumptions

7.2.1 Generalisation

This research employs a clustering algorithm pipeline that was developed and refined through the comparison and analysis of two real-world events: Hurricane Harvey and Brexit, as well as the corresponding data from the Global Database of Events, Language, and Tone (GDELT). When applying machine learning tools in such a context, it is imperative to question the potential biases that may emerge in the developed software and the data it processes.

During the course of this project, a considerable amount of work was devoted not merely to model development, but to the refinement of a distancing metric. This refined metric enhances our ability to segregate GDELT events based on their features. Throughout this process, meticulous care was taken to avoid "hard

coding" or manual input of any features that depend explicitly on event characteristics. Despite these precautions, it is possible that the semantic features of the two real-world events, and inherent GDELT features corresponding to them, may have influenced the outcomes.

The following sections will delve into these concerns, assessing the extent to which the inherent characteristics of Hurricane Harvey and Brexit, and their corresponding GDELT data, might have influenced the results. By examining these potential biases and their implications, we hope to shed light on the reliability and generalizability of the developed machine learning tool, and hence, ensure the robustness of our research findings.

7.2.2 Data Quality and Completeness

The core of this thesis hinges upon the data provided by the Global Database of Events, Language, and Tone (GDELT) Project. It is critical to pay attention to the quality and completeness of this data. The issues include a substantial amount of missing data, uncertainty about the natural language processor that GDELT uses to aggregate its events, and unclear access to the complete list of sources that GDELT employs or disregards in its data collation.

GDELT's database is characterized by considerable missing data, and this issue's impact varies across different categories. For instance, the absence of data about the ethnic characteristics or religious affiliations of Actor1 or Actor2 might not be overly detrimental for some research objectives. But it becomes problematic when the data on Actor1 or Actor2 themselves are absent, as seen in many GDELT events. This can lead to skewed clustering and data analysis, especially when these omissions are consistent or disproportionately affect certain regions.

We can expect some degree of missing data, given that it's not always clear who is involved in an event. Still, it's reasonable to expect a reputable data source like GDELT to provide data on both actors in most situations. Such gaps could indicate issues with GDELT's data collection or processing methods, suggesting a need for further investigation. Moreover, GDELT's opaque natural language processing technology raises questions about the accuracy of the data it produces. Also, the lack of a comprehensive list of sources used or overlooked by GDELT hinders our ability to evaluate the database's inclusiveness and robustness.

7.2.3 The "Black Box" of the GDELT Project

A black box is a term commonly used to refer to a process in which the observer only has control over the input and output, and does not know how the input gets converted into the output. This leads into another issue with utilizing GDELT's

collected data, which is the project's lack of transparency. Since GDELT is not open-source and offers limited insight into their coding, assuring the quality of the base facts upon which the dataset relies is challenging. GDELT's news article processing is understood at a basic level, but questions may arise about the frequency of missing data points, the possibility of enriching the dataset with more data points through code modifications, and potential biases in the articles that GDELT processes or selects. If GDELT's core does not operate on a non-discriminatory principle, all results derived from the project may be susceptible to perceptions of inaccuracies and biases.

7.2.4 Assumptions of Granularity

A key assumption in this thesis concerns the granularity of GDELT's data. It is presumed that GDELTs data is sufficiently detailed to enable the extraction of higher-level data. This is particularly relevant when considering GDELT events on a smaller scale, and the necessity for GDELT's data to maintain its granularity even when a large volume of GDELT data is used.

7.2.5 Dependance on Proper Preprocessing

An additional assumption pertains to the preprocessing phase. It is assumed that the preprocessing stage did not distort or damage the GDELT-events, or inadvertently remove event fragments that could be crucial. Any errors at this stage will lead to flawed results that may be difficult to identify as illegitimate. It is difficult to properly conduct inspections of the data to ensure that no data has been lost or changed when the GDELT dataset is a at times sparse dataset to begin with, and as such the most critical measure taken to ensure no flaws in the preprocessing stage has been a thorough error-handling workflow, as well as cursory inspections of produced datasets.

7.2.6 Assumptions about Word Embeddings

This thesis implementation of word embedding, particularly in relation to developing the distance metric, hinges on the assumption that word embeddings can meaningfully capture semantic similarities, and that the results of these word embeddings do not contain significant biases that might obstruct the clustering graph's functioning to the extent that it obscures results.

7.2.7 Scalability

As discussed in great length earlier, the computational tax imposed by HAC is a major concern when scaling the experiment to more than the lower thousands of data points

7.2.8 Interpretation of Results

Given the combination of both quantitative and qualitative methods of analysis in this thesis, there exists a possibility that results might not align between both methods, but instead contradict each other. This could stem from a variety of reasons, including faulty metrics for analysis, data biases favoring either method, the injection of personal biases by the analyst into the product that could skew results, among other issues. These considerations are important to bear in mind when analyzing data, to ensure that neither of them end up distorting the study's results. .

Chapter 8

Conclusions and Future Work

8.1 Summary

This project has encountered varying degrees of success in relation to internal and external metrics. This variance stems partially from the expansive array of possibilities within the workspace and partially from inherent challenges linked to using GDELT, unsupervised learning, and clustering algorithms. The research questions were formulated as they represented the most significant insights that could be derived from the project.

Research Question 1 investigates the potential of GDELT to accurately identify news articles reporting about the same event. This endeavor has met with partial success. The existing clustering algorithm is capable of approximately segregating events into smaller clusters that generally correspond with the same event type. This is confirmed when comparing these clusters against their objective labels post-clustering. Hence, while inspecting the data of the events within a cluster, it is generally possible to interpret the data as relating to the same event.

Research Question 2 is a more technical inquiry aimed at scrutinizing the preprocessing techniques employed in the development process, as well as the data provided by GDELT itself. GDELT has inherent issues, including missing data and a sometimes skewed processing of news material, leading to collected events that do not fully capture the reported event. Despite these challenges, the project has demonstrated that simple techniques such as dropping irrelevant data columns, min-max scaling, and preparing data for word embeddings can have a substantial impact on the clustering process, leading to discernible results.

Research Question 3 scrutinizes the technical efficacy of the distance metric used in the agglomerative clustering pipeline. The current distance metric performs adequately in producing recognizable clusters of related events when applied to events that differ in event type and geolocation. However, its effectiveness is questionable when applied to real-world events that are more similar in nature.

The simplistic structure of the distance metric may lead to closely related events being clustered together, pointing towards the need for further refinement of the metric.

8.2 Significance

This thesis has sought to pioneer new approaches to clustering GDELT data, recognizing the immense potential rewards that could be reaped from effectively addressing some of the inherent challenges of GDELT. While the literature has begun to acknowledge these issues, steps taken to fully utilize this potential and garnering the gains are relatively small.

The author harbors the hope that this thesis can serve as a catalyst for further research in this relatively uncharted domain in the ensuing years. Despite its noise and incompleteness, GDELT provides an immense, ever-expanding database that updates every 15 minutes, day in and day out. This extensive, rapidly updating data source, albeit imperfect, offers a uniquely rich foundation for machine learning applications. The hope is that, by leveraging an adequately large portion of this data, the issues related to data incompleteness could become inconsequential, ultimately enhancing our ability to accurately cluster events and better understand the complex dynamics captured within the GDELT database.

8.3 Recommendations for Future Work

Through the work on this thesis, several opportunities have made themselves available, but there has not been time enough to explore all routes, seeing as the possibility space of the field is large. On the other hand, there have been made observations about inefficacies and structural problems in the research that have been noted, but not deemed crucial enough to start the technical work over. Some immediately available options for future work has been noted, and these could play a part in future work attempting to fully utilize the level of depth and breadth that the GDELT datasets hold.

Finding better, or more expansive, news article collectors

The current process for scraping for news articles is extremely basic, and for a larger project that demands more accurate findings for relevant articles, it would prove itself productive to gather more news articles to gather GDELT-events from.

Some examples of these might be DBPedia, European Media Monitor, Google News and a host of other news aggregators.

Implementing features from more columns in the distance metric

The current distance metric process leverages a limited selection from the over 60 feature columns available in the GDELT dataset. The reasons behind this choice have been explained earlier in the thesis. However, a future project with a broader scope could consider incorporating a larger number of features, potentially enhancing the algorithm's effectiveness.

Several features inherently contain valuable information for clustering. For example, the Goldstein Scale scores and the Confidence values from the mentions.csv files could both offer additional depth. Another feature that could be utilized more is the Actor1Geo_Countrycode, which indicates the location of Actor1. This could help reduce the fuzziness of the clustering that currently arises when an actor is labeled as "Government" for both Actor1 and Actor2, even though the governments might be in different countries, such as China for Actor1 and Egypt for Actor2.

Another feature that could be easily incorporated is the AverageTone. This could ensure that news articles with similar emotional tones, whether "positive" or "negative", are placed closer together in the clusters. This could be particularly useful for objectively negative events like earthquakes or other natural disasters. However, the usefulness of this feature might be limited for politically charged events, where different news articles may present differing perspectives on the same issue.

Use a wider time space when regarding news events

The selection of real-world events for this thesis was constrained within a narrow timeframe to facilitate data handling during preprocessing. Expanding the scope and temporal range of events could significantly impact the clustering results, as the temporal features of GDELT were not fully exploited within this study's limited timeframe. This is a two-part problem, the first of which is the inherent closeness of the two events chosen for this study. The other is that GDELT offers such a large amount of data to use that the computational load will be very taxing unless computational cost-saving measures is put in place.

Use the GKG and/or mentions

This thesis did not employ two of the three main data sets provided by GDELT - the Mentions and Global Knowledge Graph (GKG) datasets. The Mentions dataset, which features metadata for most GDELT events, could have offered valuable insights had the project scope been broader. Similarly, the GKG could have been instrumental if the project's primary objective was to optimize the clustering process. Although the GKG is more complex to manage compared to the ex-

port.csv used for straightforward data usage, it offers a depth of information not found in the export files.

The GKG possesses a unique ability to grasp the context of an event, beyond just the raw data, and compare it to all other events. It also provides metadata that intrinsically links it to similar events, which could help address the relevance issue that is crucial in a clustering approach. Thus, future projects could significantly benefit from leveraging these untapped resources from GDELT.

More group to group metrics, instead of simply event to event

With one exception, this thesis has dealt with GDELT data on an event level, comparing one event to another in the distance metric. The one exception was built to handle strange or unique choices that the GDELT process had made in regards to Actor features. In dealing with not only GDELT-event to GDELT-event, there is the possibility of further enhancing the data gathered. Examples of these will be to use the groups of events with a similar SOURCEURL further, for example by adding a secondary distancing metric that works explicitly on this level. If the baseline clustering algorithm is also enhanced to a high enough degree, it would be possible to use the events that are clustered together to form the basis of new datasets to be used in further development, utilizing the connectedness that is first shown after the cluster process has made the connection that these events are relevant to each other.

This is a positive flow of information that can continue to enhance data to an almost infinite point, and as long as each step of the journey is well documented and outliers and noise is reduced, it should be possible.

8.3.1 Grasping Possibilites

It is the hope of the author that this thesis might serve as an inspiration to other people in the field of machine learning, cluster analysis, or for simple data analytic fanatics to scale the mountain of possibility that is GDELT. It contains the unrealized potential for serving as a powerful tool in obtaining an overview of the political, social and global landscapes we all operate in each day, and the future ahead only looks bright when seeing it through the lens of data analytics.

Bibliography

Bouguettaya, A., Q. Yu, X. Liu, X. Zhou, and A. Song (2015), Efficient agglomerative hierarchical clustering, *Expert Systems with Applications*, *42*(5), 2785–2797, doi:https://doi.org/10.1016/j.eswa.2014.09.054. 2.3.1

Galla, D., and J. Burke (2018), Predicting social unrest using gdelt, in *Machine Learning and Data Mining in Pattern Recognition*, edited by P. Perner, pp. 103–116, Springer International Publishing, Cham. 2.5.2

Jain, A., M. Murty, and P. Flynn (1999), Data clustering: a review, *ACM computing surveys*, *31*(3), 264–323. 2.3.1, 2.3.1

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2016), Bag of tricks for efficient text classification, *arXiv.org*. 2.4

Kwak, H., and J. An (2014), *A First Look at Global News Coverage of Disasters by Using the GDELT Dataset*, pp. 300–308, Springer International Publishing, Cham, doi:10.1007/978-3-319-13734-6_22. 2.5.4

Leetaru, K., and P. A. Schrodt (2013), Gdelt: Global data on events, location and tone, 1979-2012., accessed on: 2023-05-06. 2.1.1, 2.1.2

Leetaru, K., and P. A. Schrodt (2015), The gdelt event database data format codebook v2.0, accessed on: 2023-05-09. 4.1

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013a), Distributed representations of words and phrases and their compositionality. 2.4

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013b), Efficient estimation of word representations in vector space. 2.4

Qiao, F., P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang (2017), Predicting social unrest events with hidden markov models using gdelt, *Discrete Dynamics in Nature and Society*, *2017*. 2.5.1

Rowell, E. (2013), Know thy complexities!, accessed on: 2023-05-29. (document), 2.2

Schrodt, P. A. (2012), Cameo conflict and mediation event observations event and actor codebook, accessed on: 2023-05-28. 4.1

van der Maaten, L., and G. Hinton (2008), Visualizing data using t-sne, *Journal of machine learning research*, *9*, 2579–2605. 2.3.1

Winge, E. (2018), Word embedding models as graphs : conversion and evaluation. (document), 2.3

Yao, Y., Y. Zhang, J. Liu, Y. Li, and X. Li (2022), Analysis of spatiotemporal characteristics and influencing factors for the aid events of covid-19 based on gdelt, *Sustainability*, *14*(19), doi:10.3390/su141912522. 2.5.3

Zhang, Y. (2015), An introduction to python and computer programming. 2.2

Zhang, Y. (2016), Synthetic hierarchical clustering dendrogram conceptualization initialized with 10 data units. at the cutoff point shown, three clusters are identified., accessed on: 2023-05-26. (document), 2.1