applied sciences

MDPI

Article

Cluster-Based Knowledge Graph and Entity-Relation Representation on Tourism Economical Sentiments

Ram Krishn Mishra [1,*], Harshit Raj [2], Siddhaling Urolagin [1], J. Angel Arul Jothi [1] and Nishad Nawaz [3]

[1] Department of Computer Science, Birla Institute of Technology and Science Pilani, Dubai P.O. Box 345055, United Arab Emirates
[2] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
[3] Department of Business Management, Kingdom University, Riffa 40434, Bahrain
* Correspondence: rkmishra@dubai.bits-pilani.ac.in

Abstract: The tourism industry has experienced fast and sustainable growth over the years in the economic sector. The data available online on the ever-growing tourism sector must be given importance as it provides crucial economic insights, which can be helpful for consumers and governments. Natural language processing (NLP) techniques have traditionally been used to tackle the issues of structuring of unprocessed data, and the representation of the data in a knowledge-based system. NLP is able to capture the full richness of the text by extracting the entity and relationship from the processed data, which is gathered from various social media platforms, webpages, blogs, and other online sources, while successfully taking into consideration the semantics of the text. With the purpose of detecting connections between tourism and economy, the research aims to present a visual representation of the refined data using knowledge graphs. In this research, the data has been gathered from Twitter using keyword extraction techniques with an emphasis on tourism and economy. The research uses TextBlob to convert the tweets to numeric vector representations and further uses clustering techniques to group similar entities. A cluster-wise knowledge graph has been constructed, which comprises a large number of relationships among various factors, that visualize entities and their relationships connecting tourism and economy.

Keywords: tourism; economy; natural language processing; clustering techniques; knowledge graphs; entity relationship

Citation: Mishra, R.K.; Raj, H.; Urolagin, S.; Jothi, J.A.A.; Nawaz, N. Cluster-Based Knowledge Graph and Entity-Relation Representation on Tourism Economical Sentiments. *Appl. Sci.* **2022**, *12*, 8105. https://doi.org/10.3390/app12168105

Academic Editors: Duy-Tai Dinh and Uday Kiran RAGE

Received: 11 July 2022
Accepted: 10 August 2022
Published: 12 August 2022

1. Introduction

Tourism-based revenue is a significant measure of economic growth as it is a direct reflection of the overall economy's health. Economic growth is heavily influenced by the tourism sector as it brings crucial benefits in commercial activity that generate significant growth, development, and demand. As one of the world's fastest expanding industries, it helps economic development by creating employment, producing new income sources, and improving cross-border technology and information transmission [1].

In the early phase of the research into the link between tourism and the economy, theoretical and basic statistical studies such as the input–output framework and the notion of tourist multipliers were widely applied [2]. Social media, such as Twitter [3], is a great source of emotive data, such as those relating to sentiments [4], views, and shared perspectives on users' daily lives [5]. User-Generated Content (UGC), which includes any form of reviews, blogs, posts, and comments from social networking platforms, is one of the most crucial data sources for data analytics. UGC is composed of insightful input that is offered voluntarily by the population. This input information is readily accessible at little or no cost, and it may also be received in a simple and straightforward manner [6]. Twitter, on the other hand, generates spontaneous factual content that is considered incredibly beneficial for discovering new connections that would otherwise take a long time to be

documented on online resources. In the research, the keyword method was used to extract the unfiltered data, which assisted in discovering the relationship linkage between tourism and economy. The methodology used in this research takes a closer look into the use of knowledge graphs in order to find relevant and substantial relationships.

Knowledge graphs have evolved as an appealing concept for organizing the world's data and ideas, which are used in semantic analysis of data that is collected from diverse sources. Knowledge graphs help to understand the context of the subject, which gives more information about the topic under study, which in turn allows for a more thorough examination and understanding of the discovered interpretation and relationship. Knowledge graphs are becoming critical in expressing knowledge retrieved through natural language processing and machine learning techniques. Knowledge graphs that are inputted in advanced and evolved machine learning models help to make more accurate and precise predictions [7,8].

Clustering algorithms are also a vital part of establishing the key relationship linkage between tourism and economy. Clustering is fundamentally a type of unsupervised learning method that pulls references from datasets, which include input data with no labeled responses. It is often used as a procedure to discover relevant structure, explanatory underlying processes, generative features, and groups in a given dataset [9]. Clustering is the process of separating a population or set of data points into groups such that data points in the same group are more similar to other data points in the same group and different to data points in other groups. It is essentially a collection of items based on their similarity and dissimilarity.

The topics of natural language processing, clustering, entity relationship extraction, and knowledge graphs are introduced in Section 2—Related Work, where previously conducted studies are discussed. Section 3—Methodology discusses the process used to specifically find pertinent relationships between tourism and economy. The method is broken down into various subsections, data collection, data preprocessing, vectorization, sentiment labeling, clustering, selecting the best cluster, cluster wise entity relationship extraction, and knowledge representation. The following, Section 4—Experimental Result, shows the outcome of the study, and finally Section 5—Conclusion and Future Scope talks about the impact of the research and thoughts on the future of this dynamically evolving area of study.

2. Related Work

The economic impact of tourism has been examined from several angles in a huge number of research projects. According to the author in [10], long-term economic prosperity may best be achieved through tourism; hence, finding the linkage between tourism and economy can give light to many other factors that connect tourism to economic development. It is evident in todays' society that the tourism industry is evolving quickly. As more data is collected at an ever-increasing pace, both structured and unstructured, this data may be turned into information, which can then be used to build relations and explore factors that connect tourism to economy [11].

The exponential growth of digital technology has resulted in the access of information to most people in the globe. Every element in life, including the way one consumes, has been altered by social media. Businesses have been profoundly impacted by these advances, mostly because they have made new marketing methods possible. Undoubtedly a component of all of these is tourism, one of its most dynamic support systems for the economic sectors around the world. The authors in [12] introduce a vital and niche branch of tourism, digital tourism, which is an essential aspect in the economic sector. The integration of tourism and technology has resulted in significant economic advances and optimized various processes. Furthermore, current advances in data science technologies as part of the tourism sector have provided better experiences for visitors and analytical solutions for tourism service providers. The vast development of new algorithms and the growing simplicity in implementing data driven research approaches enlighten researchers to use

the updated approaches. Additionally, the tourism sector has adopted a data-based strategy due to the significant rise in information delivery, which has boosted the tourism-based economy through online resources and competition.

Text clustering is a vital stage in the processing of textual data and is intensively investigated by the industry of text miners. To deal with the high-dimensional and sparseness issues, many present text classification methods rely on the bag-of-words model, which does not consider text structure or sequence information. Authors in [13] used pre-trained word activations to build a deep feature-based text clustering. With the use of machine learning techniques, it is possible to define a classifier that learns to differentiate between positive and negative sentiments and then determines the polarity of new texts; because it is a rule-based approach, it is possible to derive the emotion from the specific terms used. To improve the accuracy and speed of identifying tweets based on their polarities, the authors in [14] implemented an ensemble classifier which made use of Twitter sentiment techniques. A ranking method [15] and Skip-gram meter, Word2Vec [16], were combined with a resource-based method using linguistic knowledge in their design. Text clustering techniques can be used to establish multi-level ontological linkages and identify semantic linkages between topic ideas in the framework of a knowledge graph [17].

Natural language processing (NLP) is an essential aspect in the analysis of text written by users on social media platforms. NLP has provided the capability to use relevant data for public policy decision-making by predicting economic conditions and illuminating the sentiments of the population, especially through the use of social media. NLP has been utilized in social media previously to analyze content submitted and written by the users. However, it has been shown that sentiment analysis by itself may not be sufficient to describe a group reaction in the absence of context. Authors in [18] have analyzed tweets using a model that incorporates both language and non-language elements. The authors use both texts and emoticons to analyze the sentiment of the audience, which can be a valuable indicator in finding pertinent relationships. The idea of sentiment analysis can aid in the prediction of future tourism based economic direction in terms of positive, neutral, and negative [19,20].

Within NLP, the entity relationship extraction is used to find unique relationships between two entities from the unstructured data collected. Relationship extraction is the process of identifying connections between elements in a given dataset. Crucial elements of the process include identifying relationships between entities and deciding which of them are important for the specific application problem. The authors of [21] suggested a method to identify the entities and the relationships between the entities in order to automate the ER diagram design process. The relationship extraction process is then used to create knowledge graphs, which are visual representations of the entity relationships of data, which assists in making logical inferences in order to retrieve implicit knowledge [22]. Authors in [23] focused on building approaches on the extraction, representation, fusion, and reasoning of the knowledge in the graphs, which included attaching entities and relations to knowledge graphs after data processing [24].

3. Methodology

In this section of the research, a discussion of a detailed stepwise approach is presented as shown in Figure 1. The process starts with collecting the data from Twitter followed by a few crucial components such as cleaning and preprocessing of the data. Then the collected the tweets are changed to numeric vector representations, where an optimized K-means clustering algorithm is used to group similar entities. Cluster-wise knowledge graphs are finally created to visually represent significant relationships in relevance to tourism and economical sentiments.

3.1. Dataset Collection

Using tweepy [25], an open-source Python [26] module, a total of 100,000 tweets were gathered and mined over the course of four months. Tweepy facilitates access to the

Python-based Twitter APIs [27], which employ consumer keys and private access tokens for authentication. The datetime library was used to include a custom script that was created to precisely extract 150 tweets every day and kept in a python list. When the search term "tourist" was used, all of the tweets containing related terms like "economy" and "tourism" were grouped together. A CSV file containing the tweets from 9061 different individuals was collected. An example of the dataset, which includes the characteristics Index, Datetime, and Tweets, is displayed in Table 1.

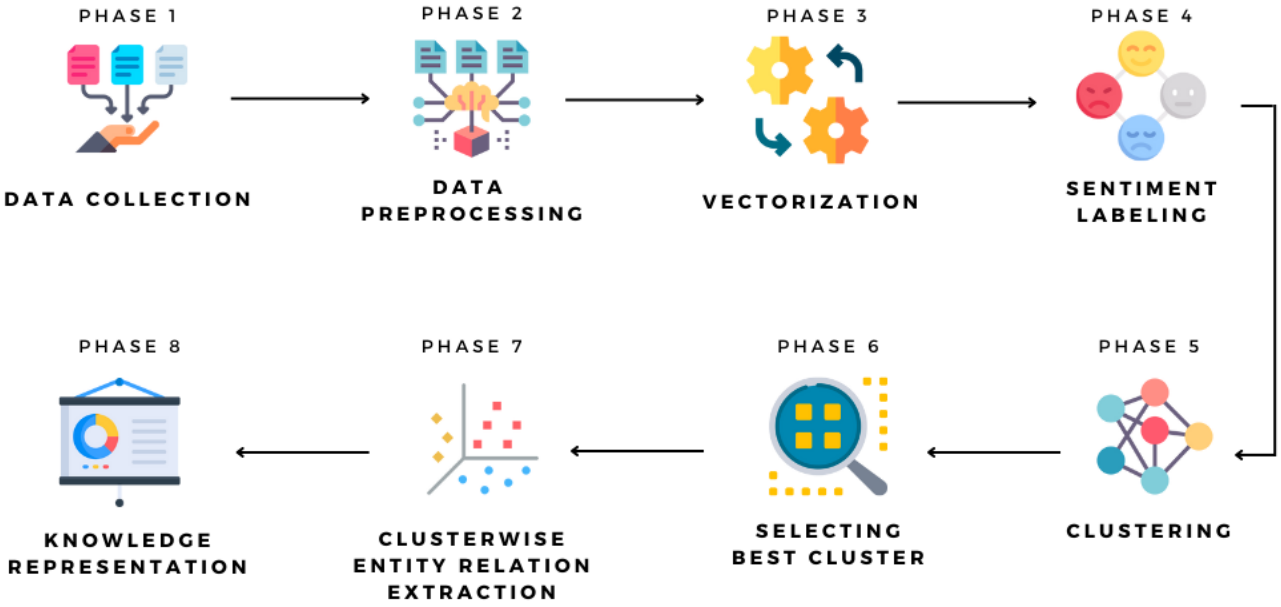

Figure 1. Methodology Overview.

Table 1. Dataset Overview.

Date/Time	Date Accessed	Tweet
31 December 2019	11 May 2022	India's tourism industry was the only promising aspect of economy, now that too has got a severe hit with Kashmir shut down in August and the CAA effect on Eastern India and other parts of India. What's happening to our Country's economy!
31 December 2019	11 May 2022	https://t.co/GDK6faOrW3 (Tourism becomes Taiwan's economy-Does the U.S. know that strengthening Taiwan tourism and immigration is a priority)
31 December 2019	11 May 2022	@paolo_lim @SenhorRaposa @OryxMaps @ElectionMapsCo @politicsluo @LoganZT1 If New England is weird, then VT is super-weird. It too has a lot of WWC voters and not too many educated suburbanites. But its tourism-based economy (and especially its skiing, with the associated concerns about climate change) will keep it solidly in the blue column. 3/?
31 December 2019	11 May 2022	@SportingTruth_ @DanMacPherson You know the bulk of the estimated $100 million has already entered the economy through tourism (and flow on), right?
31 December 2019	11 May 2022	Expanded air services to #SanJuan expected to add $124m to PR economy in 2020 #Tourism https://t.co/5X5FsLfTxV https://t.co/NYXtuMOcE7
31 December 2019	11 May 2022	@mumbaidilse @AUThackeray It's a very good idea to boost economy in 4× speed. Create more jobs, reduce crimes & also tourism
31 December 2019	11 May 2022	@travelfish @Tim_Hannigan I also found it to be an annoying/useful look at global tourism-for the exact same reasons as you. Becker is good at pinpointing problems, but her fixation with solving things by embracing high-end tourism reveals a lack of familiarity with the nuances of the travel economy.

3.2. Dataset Preprocessing

Data preprocessing is a mining activity that transforms the raw data into an understandable format [19,28]. To overcome the issue of redundant unstructured information through tweets preprocessing is important and it has been done with a tweet-preprocessor library, which is held responsible for identification of emoticons, URLs, and reserved words. Due to the unstructured information obtained from the tweets, the information must be preprocessed into quantifiable evaluation [29]. For each tweet, the orientation of emotions are positive sentiment, negative sentiment, and neutral sentiment. Table 2 describes the query with keywords along with their sentimental score.

Table 2. Calculated Data.

Number of Clusters	Silhouette Score
2	0.188
3	0.150
4	0.130
5	0.132
6	0.119
7	0.117
8	0.100
9	0.091
10	0.087
11	0.081

Stop Words are terms like "for" or "by" that have no real meaning. Eliminating such terms is usually believed to be an essential step, which can enhance the quality of the framework. Figure 2 shows all the necessary steps taken to preprocess the tweets.

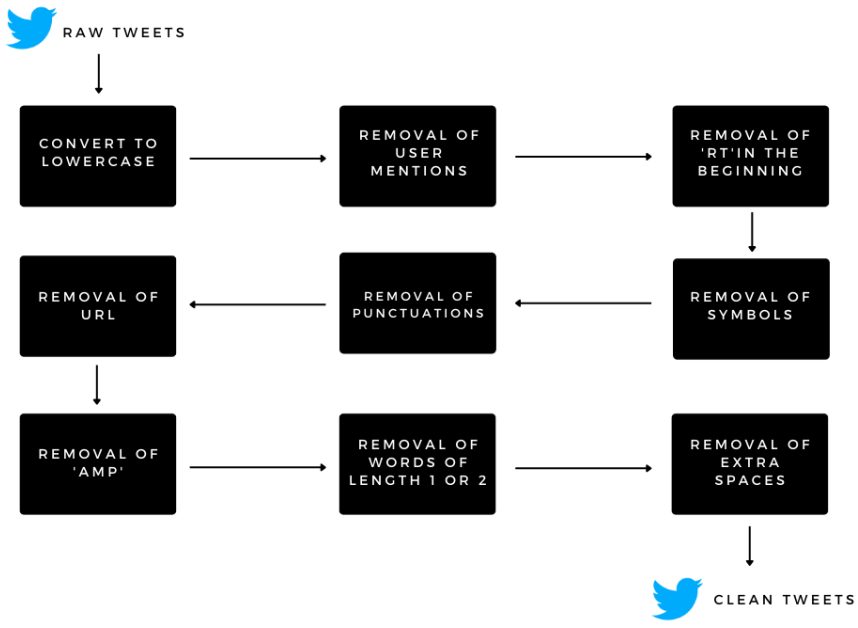

Figure 2. Data Preprocessing.

3.3. Tweet Vectorization

Machine learning methods work with a numeric feature space and need input in the form of a two-dimensional array with rows representing occurrences and columns representing features. To perform machine learning on text, first the documents must be converted into vector representations so that computation operational for machine learning may be applied [30]. This procedure is known as feature extraction or, more simply, vectorization, and it is a necessary initial step in language-aware analysis.

Punkt sentence tokenizer splits a given paragraph into a collection of sentences by building a model for abbreviated terms, compound words, and terms that initiate lines using an unsupervised method. Before utilizing the tokenizer, it should be trained on a substantial amount of input in the chosen language. A pre-conditioned Punkt tokenizer for English is included in the NLTK (Natural Language Toolkit) data collection [31–33].

Word2vec [34], shown in Figure 3, is used in order to convert the tweets into the vector format. Statistical computations on such word vectors allow us to discover connections between the terms. Word2Vec is a hybrid method that combines two methods: CBOW [35] (Continuous Bag of Words) [36] and the Skip-gram model [37]; both are neural networks that map one or more words to a target variable that is also a word. Weights that serve as term vector forms are learned in both approaches. CBOW predicts the chances of a term based on the context in which it is used, which can be a sole neighboring word or a series of words. The Skip-gram model operates in the opposite direction, attempting to predict the background for a given term [38].

Figure 3. Vectorization.

In order to employ textual content in machine learning methods, it must first be transformed it into a vector. The bag-of-words strategy is one of the techniques. Bag of words approach neglects syntax and word structure. First, a keyword bank is constructed from the whole sample of dataset that was gathered. As a result, each text or item is considered as a series of input vectors depending on the corpus's terminology [35].

3.4. Sentiment Labelling

Every year, the number of digital exchanges increases significantly, and content analysis gives a mechanism to comprehend the views, ideas, and reactions that lay beneath the online language. It is especially valuable in the era of social media, that may give you a sense of how people feel about certain topics. Data collected from sentiment studies on social networking sites has a wide range of applications, ranging from minor applications like improving brand building activities to bigger public concerns like guiding government ideology and forecasting economic growth [5,39]. As people are expressing their ideas and emotions more freely nowadays, sentiment analysis is quickly coming up as an indispensable tool for monitoring and understanding opinions in all forms of data, including social media.

Sentiment labeling can be considered as a subdivision of natural language processing that involves determining the underlying emotion in a textual content [40,41]. This involves measuring the sentiment polarity of an entity. Sentiment polarity refers to the inclination of the text, whether the emotion is more of negative or positive. Polarity score is a float value of the range $[-1.0, 1.0]$.

After converting the textual data into vector representations, the polarity of each text is measured based on lexical approach. NLTK is a python library that provides the user with simple access to large sources of lexical knowledge. Textblob is an NLP module present in NLTK that helps users to facilitates the investigation and manipulation of large amounts of textual data. Textblob returns sentiment polarity of a sentence based on positive and negative words used [42].

The polarity and subjectivity of a statement are returned by TextBlob: -1 denotes a negative emotion, whereas 1 indicates a pleasant emotion. This polarity is reversed when using negative terms. TextBlob is used as a labeling system to perform fine-grained analysis. For instance, emoticons, exclamation points, emojis, and like (button) are all examples. Between [0, 1] is where subjectivity is found. Quantifying the quantity of subjective information in a piece of writing is called subjectiveness. The more subjective the material, the more likely it is to be filled with personal opinions instead of facts. Intensity

is an additional TextBlob property. TextBlob uses 'intensity' to determine how subjective a piece of content is. The severity of a term affects whether it alters the following word or phrase. Adverbs are often employed in English as modifiers (e.g., "excess amount") [43,44].

3.5. Clustering

Clustering plays a huge role in data analysis and can be regarded as one of the most challenging aspects under unsupervised learning. The main aim of clustering to make groups of similar unlabeled data points called clusters. For the clustering process, the *K*-means clustering approach is used. *K*-means clustering is a well-known and strong unsupervised classification approach that may be used to a wide range of problems where each data point is added to a cluster. It is used in the solution of a wide range of complicated unsupervised machine learning issues [45]. The Figures 4–9 demonstrate how *K*-means clustering works.

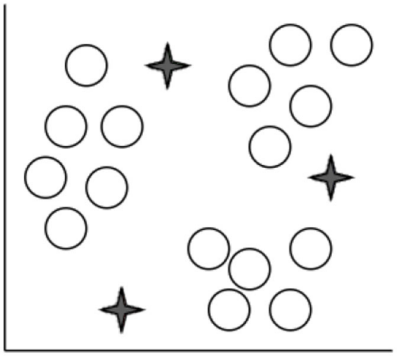

Figure 4. Step 1—*K*-means: $K = 3$ Step 1: Choose how many clusters (K value) to be formed to group the data points.

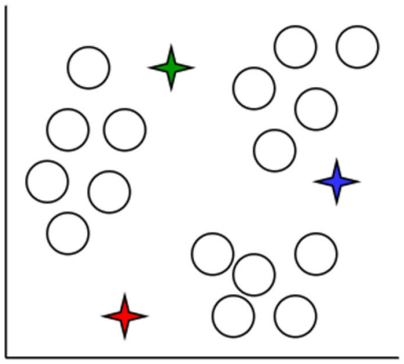

Figure 5. Step 2—*K*-means: $K = 3$ Step 2: Initialize 3 centeroids.

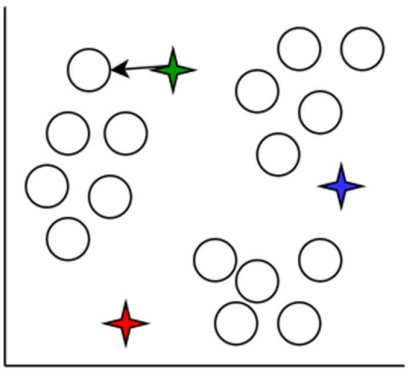

Figure 6. Step 3—*K*-means: $K = 3$ Step 3: Group data point to their nearest centroid by measuring the distance from each centroid using Euclidean Distance formula: $d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$ Euclidean Distance between data point X and Centroid C can be calculated using the formula.

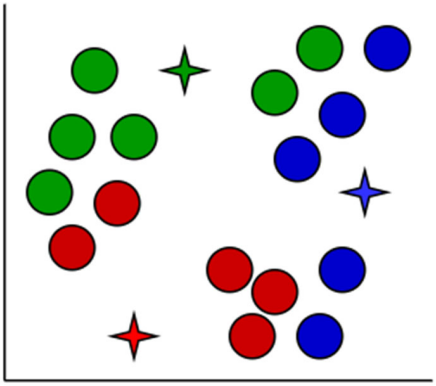

Figure 7. Step 4—*K*-means: $K = 3$ Step 4: Data point added to its nearest cluster.

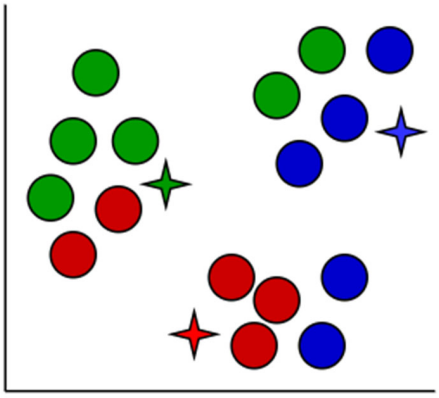

Figure 8. Step 5—*K*-means: $K = 3$ Step 5: Measure the average of data entities in each cluster and relocate the centroid.

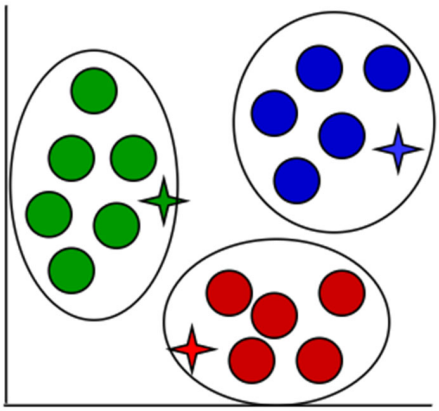

Figure 9. Step 6—*K*-means: $K = 3$ Step 6: Repeat previous steps until it has reached the point where the data points do not switch to another cluster. Here, it can be seen that there are three distinct clusters. While the figure shown has taken form in a short period of time, this approach normally requires several trials.

An iterative method of allocating each piece of data item to the categories, and over time, data points begin to clump together based on similar characteristics. The goal is to reduce the total of gaps among the data items and the group center to the smallest possible value in order to determine which group every data item should fall to.

K-means clustering function can be stated as

$$\sum_{k=1}^{K} \sum_{x_i \in \pi_k} ||x_i - \mu_k||^2$$

K—Total number of clusters
k—Current cluster
π—All data point
x_i—Selected data point
μ_k—Centroid of cluster k
π_k—Data points in cluster k.

Silhouette score [46] is used to evaluate the quality of clusters created using the clustering algorithms. In this study, the K-means clustering method is used where it uses optimized centroids to calculate silhouette scores. The silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are. The intra-cluster distance between the sample and other data points in the same cluster is also taken into account by the silhouette score. The score is within the range $[-1, 1]$, where negative values imply that data belonging to clusters are possibly incorrect. A silhouette score of 0 suggests that the clusters are overlapping and a score 1 means that the cluster is dense and neatly separated. After the calculation of the silhouette score, the scores can be plotted which aids the selection of the most optimal value of K, the number of clusters in the K-means clustering method. In this study, there are 11 iterations of the clustering process where the optional cluster size for the given dataset is 5 ($K = 5$) [47].

3.6. Selecting the Best Cluster

The K-means clustering technique is one of the most adopted cluster analyses for its simple technique and quick resolution. To ensure that the grouping is successful, the K-value must be provided initially. This has a serious influence on the simulation outcome.

As mentioned in Step 1, K-value must be decided before the beginning of the procedure. After classifying based on the sentiment polarities, the optimal number of clusters for the data to be grouped is found by measuring the silhouette score of a range of clusters in an iterative method [38]. The silhouette score, seen in Table 2 and Figure 10, for a cluster can be measured using the formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$

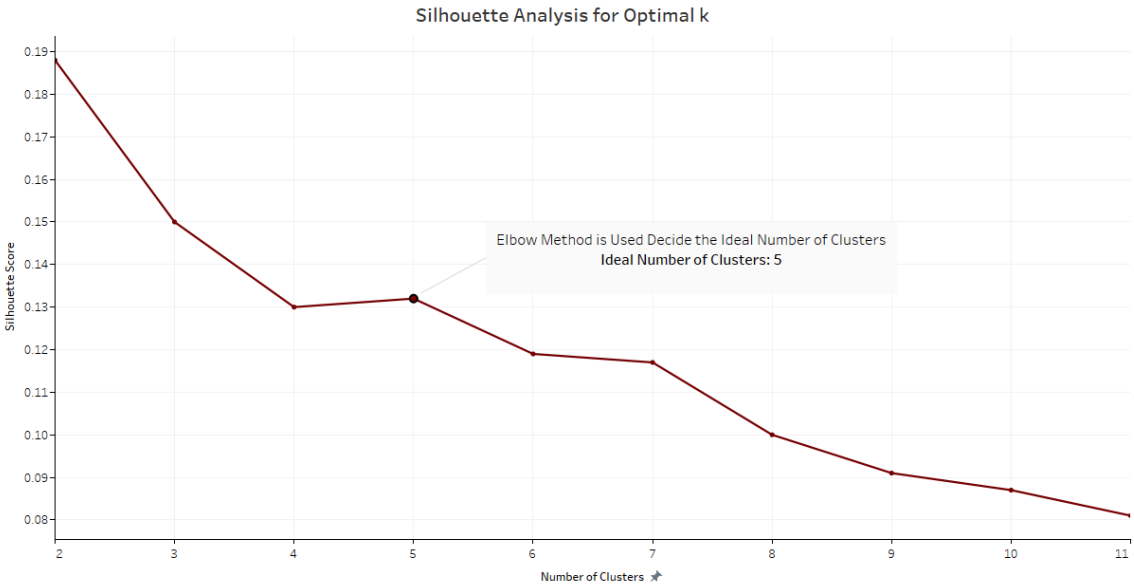

Figure 10. Silhouette analysis to find the optimal k value.

3.7. Cluster Wise Entity Relation Extraction

With regard to the applications of natural language processing, there is a particular emphasis on data analysis for digital media or online data mining, with relationship extraction being one of the most essential aspects of data analysis. The process of identifying the relationships between two different elements in a text is known as relationship extraction. A sentence is broken down into two entities and then the relation analysis is performed between the subject and object that have been recognized in the text [25].

In order to create a knowledge representation graph, the entity pairs need to be extracted from the sentences. Functions using SpaCy matcher are built to extract the entities.

Step 1:

The main objective is to pull out an entity pair, when the sentence is being analyzed. This chunk contains several declarations of empty variables. The variables 'prv_tok_dep' and 'prv_tok_text' will store the dependency label of the preceding word in a phrase as well as the prior word itself, respectively, in the context of a sentence. The text that relates to the entities of the text will be contained within the prefix and modifier.

Step 2:

Check for punctuation marks in the text. If a punctuation is present disregard that token and move to the next. Check if that word is a compound word. Compound words are a set of words that come in together with different meanings. When a subject or an object is discovered while analyzing the text, the compound word will be added as a prefix to it.

Step 3:

Extracting the subject as the first entity.

Step 4:

Extracting the subject as the second entity.

After all the steps are completed, an entity pair extraction is created, seen in Figure 11. SpaCy is a Python library for sophisticated NLP techniques that is available as a free, open-source download. SpaCy is specifically built for usage in production environments, and it assists developers in creating systems that analyze and understand massive volumes of text. It may be used to develop information extraction and natural language comprehension systems, as well as to pre-process text in preparation for deep learning applications.

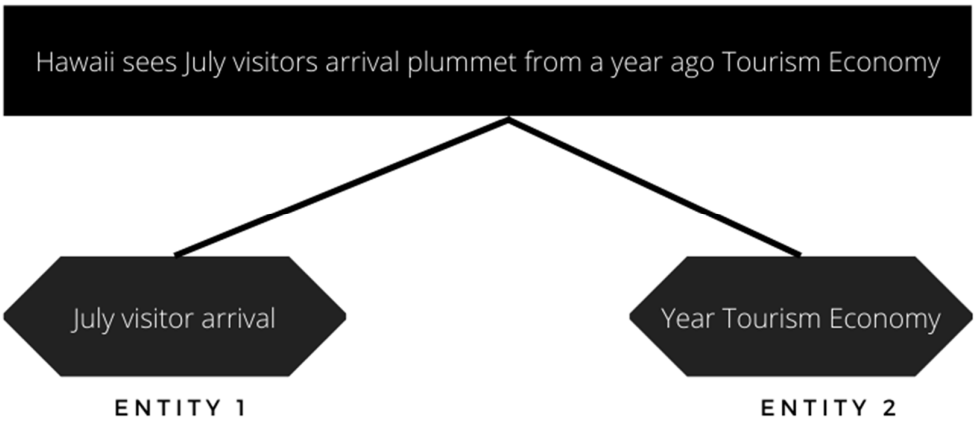

Figure 11. Entity Pair Extraction.

The first part of the work is completed with entity extraction. The nodes must be linked in order to construct a knowledge graph (entities). These edges represent relationships between nodes that are next to one another. The architecture of the statement, including items like the subject, object, modifiers, and components in the statement, must be understood in order to extract the needed details. The matcher lets the user locate terms or expressions by applying rules that describe the token properties of the words and phrases. In addition to lexical characteristics, rules may relate to token annotations (such as text or part-of-speech tags) and other types of markings.

Now a pattern is added to the matcher object based on what needs to be extracted. This pattern is made using a specific format mentioned in Spacy. Adding a pattern to the Matcher, seen in Figure 12, involves adding a collection of dictionary definitions. In each dictionary, one token and its properties are described.

```
pattern = [{'DEP':'ROOT'},
           {'DEP':'prep','OP':"?"},
           {'DEP':'agent','OP':"?"},
           {'POS':'ADJ','OP':"?"}]
```

Figure 12. Pattern Added to the Matcher.

The syntactic organization of a statement may be represented by obtaining its dependency parse. Headwords and their dependents are linked by this rule. The root of a statement refers to the core of the statement, which is independent of the rest of the statement. The main part of the phrase is generally the verb, and the root serves as a hub for all other terms.

Sentences are parsed to determine the underlying root word using the function shown in Figure 13 and Table 3. After it has been identified, it is checked to see whether a preposition or an agent word comes after it. If so, the suffix will be appended to the root term.

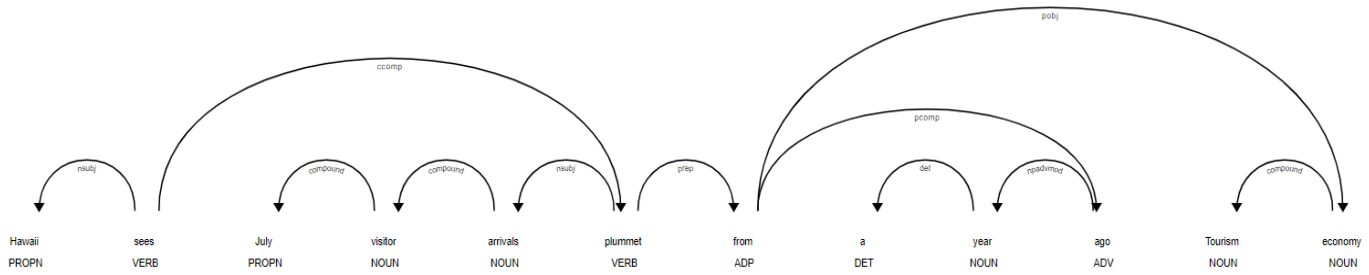

Figure 13. Sentence Parsing.

Table 3. Sentence Parsing.

Word	Parsing
Hawaii	nsubj
sees	ROOT
July	compound
visitor	Compound
arrivals	nsubj
plummet	clomp
from	prep
a	det
Year	npadvmod
Ago	pcomp
Tourism	compound
economy	pobj

"Hawaii sees July visitor arrivals plummet from a year ago Tourism economy".

From the entities and relations extracted shown in Table 4, a network of directed graphs is built using the Networkx library in Python. Directed graph implies that relation

is only from one entity to the other. The nodes are regarded as the entities and the link between them is considered to be the relation between the entities.

Table 4. Frequency of Relations Extracted.

Relation	Frequency
is	5305
are	1025
s	929
Have	841
Be	769
think	612
need	575
said	474
has	448
let	415
was	375
says	368
know	350
see	322
help	313
want	285
going	243
make	232
read	226
hope	212

3.8. Knowledge Representation

When it comes to building knowledge graphs, one can often find data from a wide variety of platforms, each with different data structures. To make sense of a wide range of data, formats, names, and semantics should all be taken into consideration collectively. As the knowledge graph's building blocks, schemas establish nodes' identity, and the setting establishes their location in relation to each other; terms that have multiple meanings benefit from the inclusion of these features. For example, Google's search engine system can distinguish among the color orange and the fruit orange [48].

The text data is aggregated on a graph-based data structure in the case of semantic processing. Entities are represented as nodes in the knowledge graph, and certain pairs of entities are connected to each other in some way. Edges are used to show these connections. The integration of data around knowledge graphs may aid in the development of new knowledge by revealing relationships between data points that may not have been apparent previously [49].

Knowledge graphs have evolved as an appealing concept for structuring the world's largest user acknowledges, as well as a method for integrating data retrieved from a variety of platforms, among other things. Knowledge graphs have begun to carry a significant role for describing the material retrieved via the use of NLP and artificial intelligence techniques, among other methods. Domain information represented in knowledge graphs is being fed into neural network models in order to make more accurate predictions [50].

The suggested solution relies on the identification of entity pairs and their relationships, and the use of spaCy's rule-based matching makes this possible.

Using knowledge graphs in conjunction with machine learning approaches, this research demonstrates that it is possible to identify sentiment in brief portions of text. Knowledge graphs are capable of capturing the structural information included in a tweet as well as a fraction of its underlying meaning. The most significant focus of this research is linked to knowledge graphs, which have the benefit of not being impacted by the quantity of the content or the usage of accents and being able to be graphically examined. There are nodes and connections in the knowledge graph, and each node's context serves as a vector representation of each item and relation. Consistent models of both local and global relations are maintained during the synchronization process, while precise feature learning information is also included in the resultant knowledge graph from the constructor.

4. Result and Discussion

4.1. Experimental Setup

The experimental study was done on a high computing server-Intel® Xeon® Silver 4210R CPU @ 2.40 GHz 2.39 GHz with installed RAM size of 32 GB. As part of the Python code editor for development, Anaconda and Visual studio were utilized as these tools provide an integrated development environment to deal with the multiple dependencies and python packages.

To process the high volume of tweets, few of the popular python libraries, such as pandas and NLTK, played a major role. Pandas was used in the data cleaning and processing aspects of the methodology and the graphical visualization of the relationships, the most important phase of study was done using visualization libraries, such as matplotlib, Spacy [51], and Networkx [52]. Other libraries were used, such as re for regular expression and tqdm for time duration, to deal with data more efficiently.

4.2. Experimental Results

Experimtal flow of this study have been followed as per Figure 14. This section starts from clustering of tweets, then counts the sentiments as per the clusters and then extracts the entities & relationship which has leads to visualization and knowledge analysis.

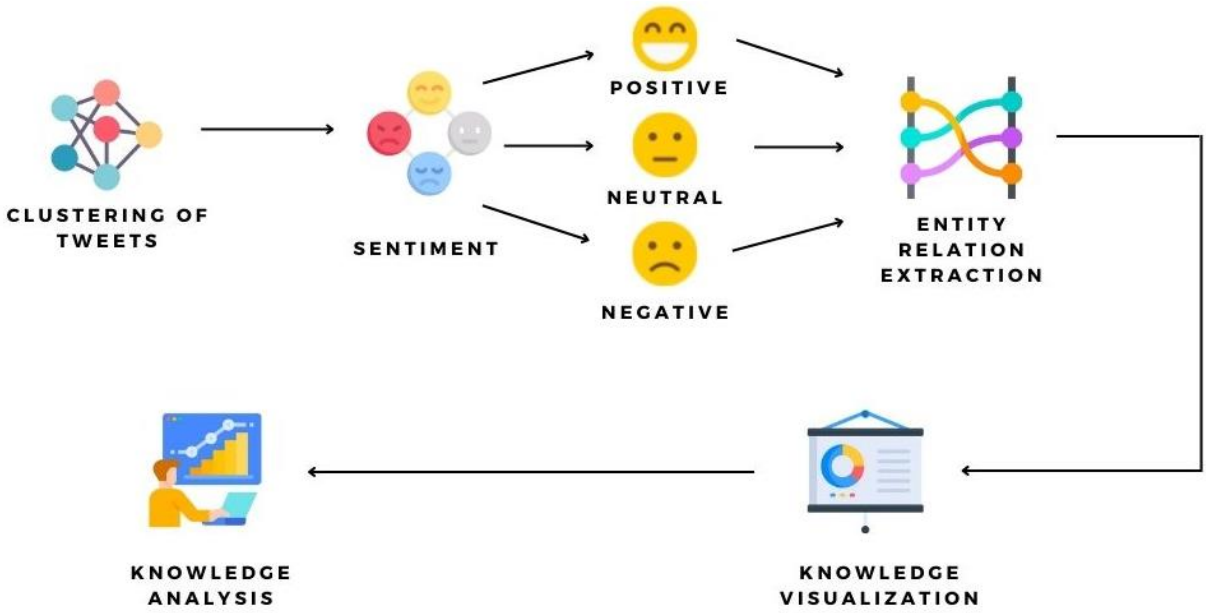

Figure 14. Experimental Flow for Result.

4.2.1. Clustering of Tweets and Sentiments

In order to select the ideal number of clusters, the elbow method was used. It is evident that the ideal number of clusters for this dataset is 5. Table 5 shows the number of tweets that have been assigned to each cluster.

Table 5. Number of Tweets in Each Cluster.

Cluster	Tweets	Sentiment
1	19,328	Positive–0; Neutral–13; Negative–19,315
2	24,548	Positive–24,536; Neutral–12; Negative–0
3	16,339	Positive–0; Neutral–13; Negative–16,326
4	27,385	Positive–27,375; Neutral–10; Negative–0
5	6982	Positive–1776; Neutral–6; Negative–5192

4.2.2. Knowledge Visualization and Analysis

The interactive representation of entities has been shown in Figures 15–19 for Clusters 1, 2, 3, 4, and 5, respectively, where the identification of name entities, which determines which words belong to what entity types, have been experimented using Python3 inbuilt libraries, Spacy, pretrained NLP, and machine learning models. Spacy provides a useful library that categorizes and highlights specific words in terms of name entities. In Figure 15, it is evident that keywords in the text are classified using name entities, such as "saudi arabia" as "Geopolitical Entity (GPE)" and "more than 100 million" as "Money". The classification of name entities is an essential step in the extracting relationship and highlighting these relationships in visual forms using knowledge graphs.

Figure 15. Name Entity Representation for Cluster 1.

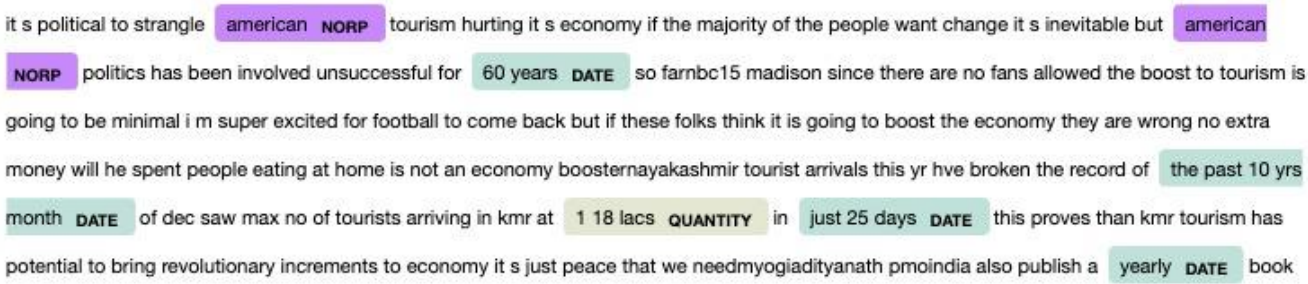

Figure 16. Name Entity Representation for Cluster 2.

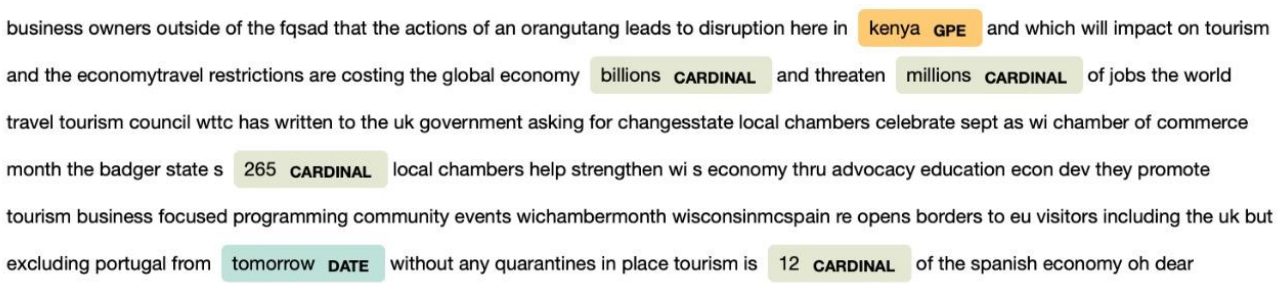

Figure 17. Name Entity Representation for Cluster 3.

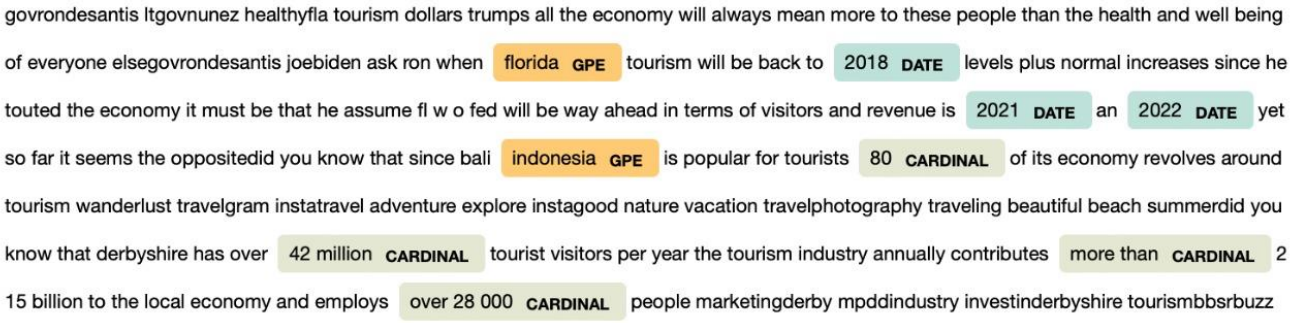

Figure 18. Name Entity Representation for Cluster 4.

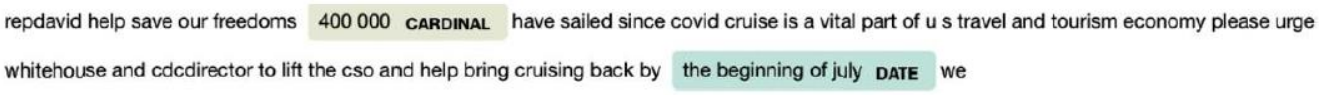

Figure 19. Name Entity Representation for Cluster 5.

Tables 6–10 and Figures 20–24 show the visual representation of extracted entities and relationships from Clusters 1, 2, 3, 4, and 5, respectively. Knowledge graph representation has been performed on extracted source nodes and target nodes, which are connected through edges. Since most of the tweets were captured during the COVID-19 era, the impact of the pandemic is a crucial aspect when determining relationships between tourism and economy, hence for all clusters the relationship used is COVID-19.

Table 6. Cluster wise Entity Relation Extraction–1.

Source	Target	Edge
indefinitely everyone	it	suffer
don it	good grief	want
quietly they	policy error	admit
Cuban they	3000 year	is
local economy	1800s	is
tourism	worse economy implications	is
huge Florida economy they	huge Florida economy state	going
that	dealers	is
big crash economy	tourism	be

Table 7. Cluster wise Entity Relation Extraction–2.

Source	Target	Edge
calm Catalina Island coronavirus	francineorr tourism economy latimesphotos	on
marine bea economy	more 235 tradeonlytoday	news
than 100 he	than 100 million	added
breaking	high uhuru kenyatta areas	comes following high
logistical this	real trust austria uk	is
rich folks	retirement	are
blue blanket	aviation fuel economy cop26	speaks
now Houston livestock these	major tourism perspective	cancelled

Table 8. Cluster wise Entity Relation Extraction–3.

Source	Target	Edge
tourism	Spanish economy	is
heavily pakhtunkhwa pm this	youth team patriots	is on
them	how much	s hard
now outlook	traveling after covid travel industry	is
even shouldn t	healthcare	is key
low nitaqat fm	long tourism nipah	is
time someone	such mandir economy	wrote
heavily pakhtunkhwa pm this	youth	is on
tourism	economy	tolerate such

Table 9. Cluster wise Entity Relation Extraction–4.

Source	Target	Edge
economy	well everyone	trumps
far it	ahead visitors	seems
80 travel Instagram adventure	nature vacation travel photography	know
28,000 28 people	Mpdd industry invest inder by shire tourism	know
tribal place	once you	cherish
too you	local visitor visit bathbiz	thanks
so iso pas	safely tourists isostandards	know
country work	American women	agree
why though you	it	know

Table 10. Cluster wise Entity Relation Extraction–5.

Source	Target	Edge
how dorsetcounciluk s digital	visitors	learn
hospitality who	on coronavirus	is key
tourism	1 economy	is
tourism industry	65 date	is
royal family	1 8 flows	clivebull
how dorsetcounciluk s digital	visitors	learn
1 air safety this	air tourism	informed
small that	local economy	is
tourism	economy	tolerate such
how dorsetcounciluk s digital	visitors	learn

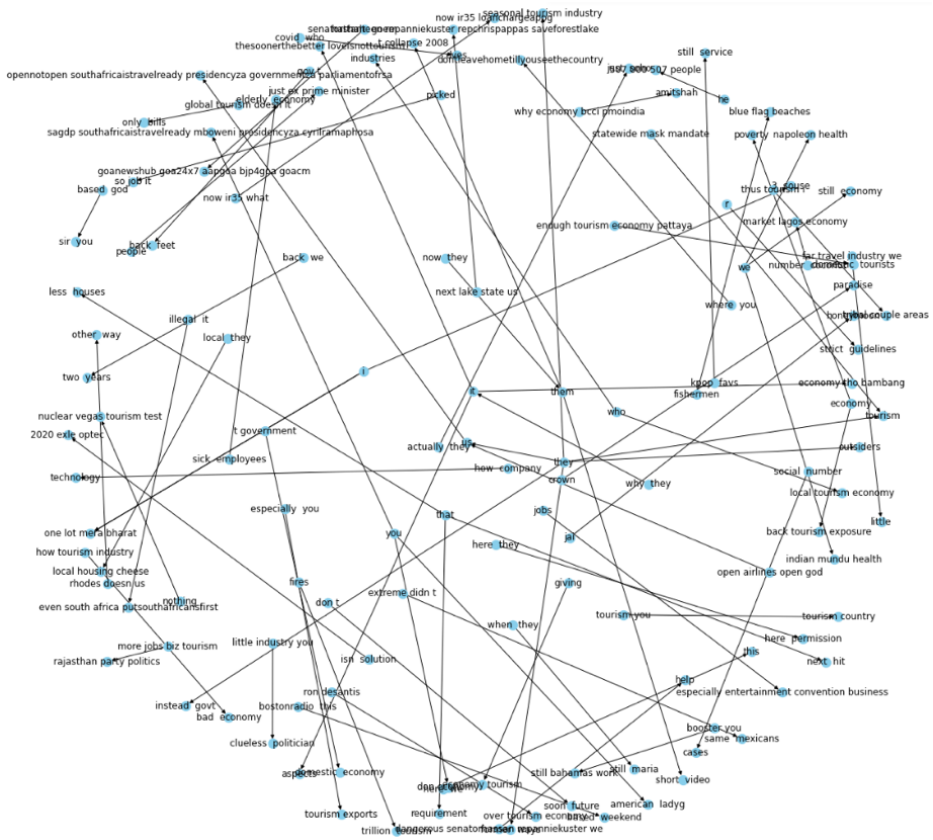

Figure 20. Knowledge Graph for Cluster 1.

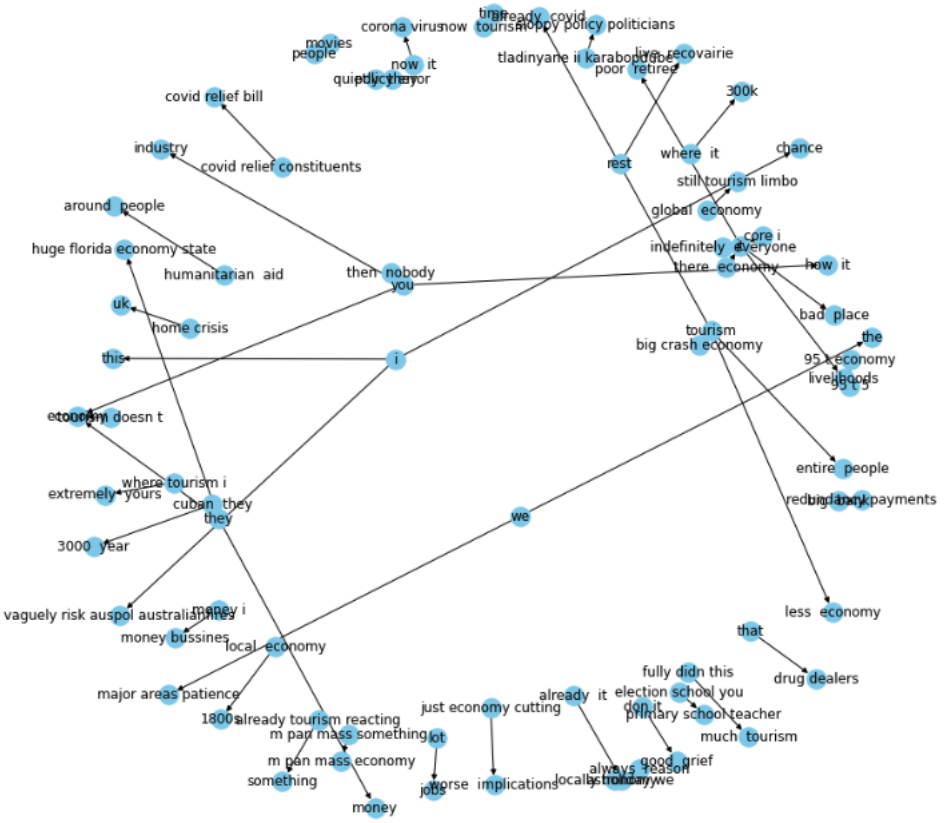

Figure 21. Knowledge Graph for Cluster 2.

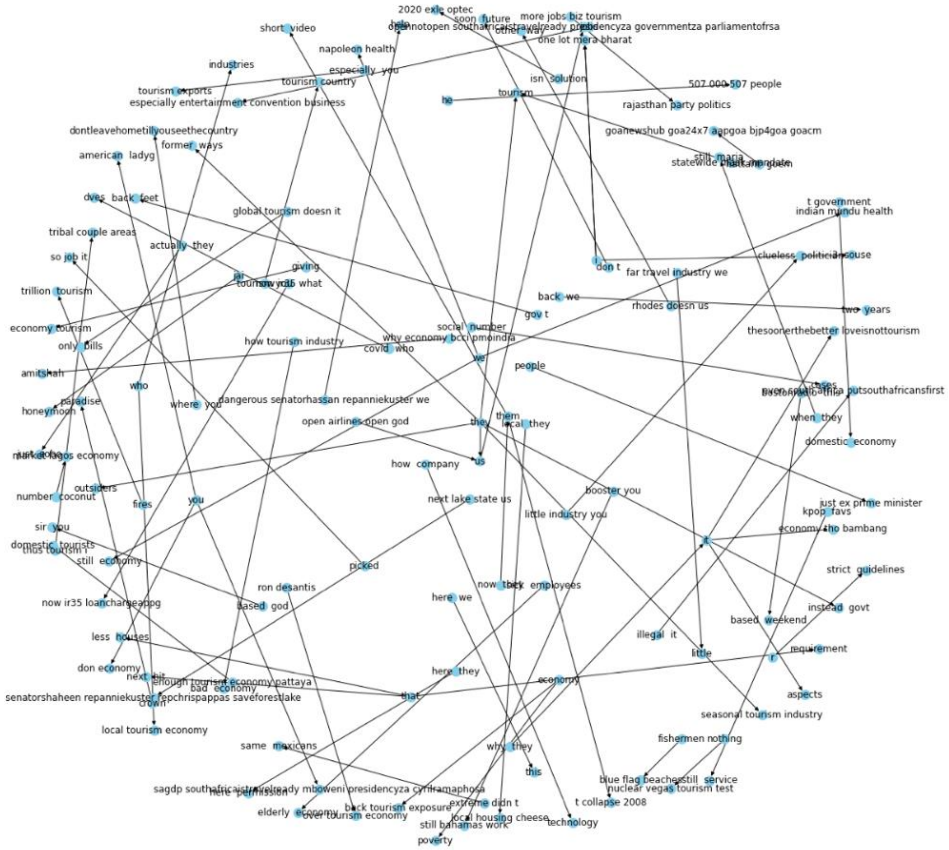

Figure 22. Knowledge Graph for Cluster 3.

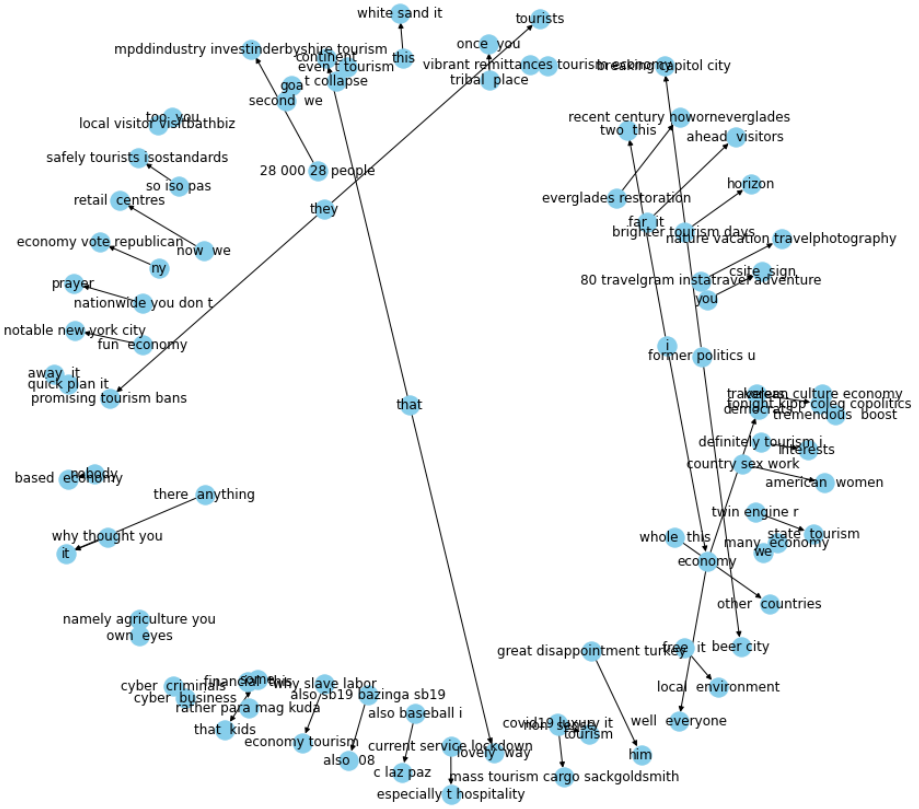

Figure 23. Knowledge Graph for Cluster 4.

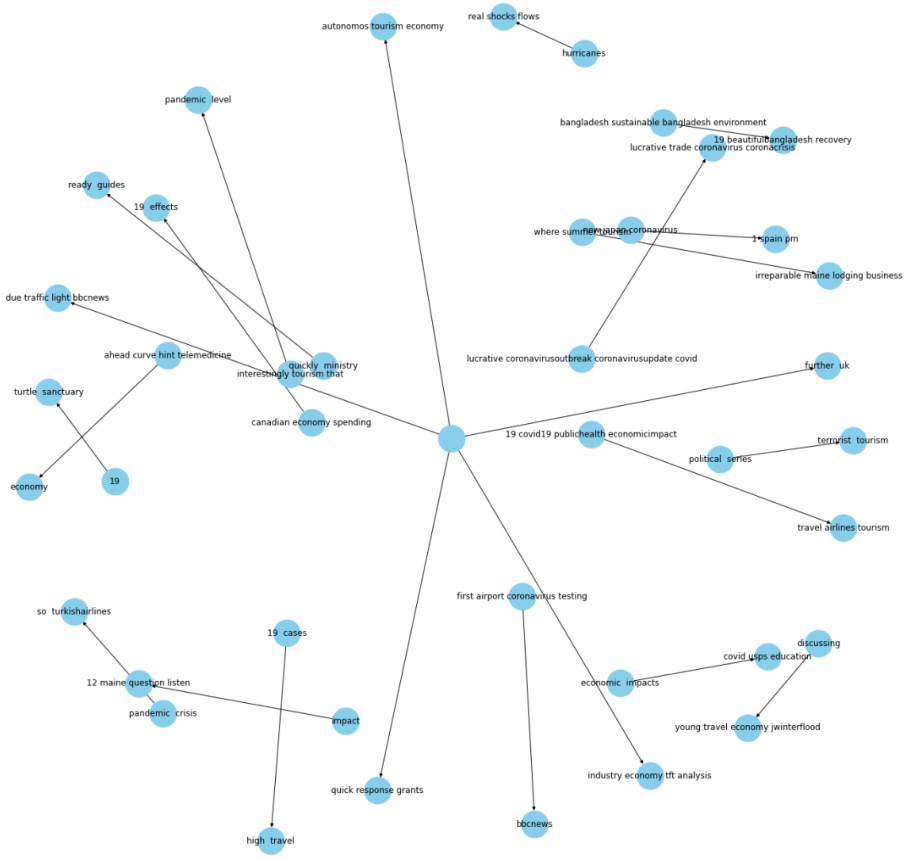

Figure 24. Knowledge Graph for Cluster 5.

In Cluster 1 a key relationship identified using the knowledge graph between "big crash economy" (source node) and tourism (target node) where relationship is identified through "be" (edge). This suggests that due to COVID-19 there was economic crisis that effected the tourism industry.

For Cluster 2, an important relationship recognized is between "rich folks" and "retirement" with the edge as "are". The relationship suggests people who have acquired significant amount of wealth seek early retirement and generally enjoy travelling.

Within Cluster 3, the relationship between "new outlook" and "travelling after covid travel industry" with the edge as "is" implies that people must look at travelling differently after the pandemic, where more precautions are necessary for everyone's safety.

As part of Cluster 4, an essential relationship found is between "service" and "hospitality" that indicates that an important part of the tourism industry is service and hospitality, where service and hospitality results in more tourists, hence improving the local economy.

In Cluster 5, a vital connection distinguished is between "air safety" and "air tourism" with the edge as "informed". This correlation insinuates that air travel after the pandemic must take into accounts safety precautions, such as wearing masks and regular sanitization.

4.3. Discussion

Reviewing previous works and studies conducted on knowledge graphs seems to be the first that attempts to finds correlations between entities using such graphs, which in this case is tourism and economy. This study successfully identifies key relationships between tourism and economy, which can be used for further studies in similar fields. Furthermore, the study provides a comprehensive overview of new methodologies, which are utilized in this article, making the article relevant and insightful for readers who are interested in the domain of finding relationships between entities through knowledge graphs or knowledge graphs in general.

The limitation of this study is that the number of entities extracted is limited to two keywords, therefore, in the future multiple knowledge graphs could be integrated to create a powerful system that would give out faster and more accurate results. Additionally, future work can also include the integration of more complex forms of machine learning algorithms and advanced clustering methods in order to get more precise results. Furthermore, this research only evaluates English language tweets, tweets in other languages and emoticons can also be analyzed, where the dataset collected will be multidimension.

5. Conclusions and Future Scope

During the past few years, Twitter has risen to be the most popular social media platform for users to voice their views and thoughts on a variety of businesses and services. As a result, it draws a large number of scholars who utilize it as a data source for sentiment extraction and data mining research investigations.

Knowledge graphs and *K*-means clustering are used in this study to create a novel sentiment analysis approach that brings out the linkage of the tourism sector and economy. The structure of a tweet review and some of its meaning may be captured using knowledge graphs. The key focus would be in the area of knowledge graphs, which have the benefit of not being impacted by the quantity of the content or the usage of languages, and of being able to be graphically reviewed throughout their formation.

Graphs have the added benefit that they will be traversed swiftly. If supplied with the necessary degree of schema, graphs could be implemented to integrate information taken from multiple sources in an automatic and clean way as a foundation for analysis. So, graphs are a useful approach to express contextual information for usage across the industry and are well equipped for providing data source in a solution independent and future proof manner. This research paper is a gateway to the newly introduced topic of knowledge graphs and entity relations extraction.

Author Contributions: Conceptualization, R.K.M.; methodology, R.K.M.; software, R.K.M.; validation, R.K.M., H.R., S.U. and J.A.A.J.; formal analysis, R.K.M.; investigation, R.K.M. and H.R.; resources, R.K.M. and H.R.; data curation, R.K.M.; writing—original draft preparation, R.K.M. and H.R.; writing, R.K.M., H.R., S.U., J.A.A.J. and N.N.; visualization, R.K.M. and H.R.; supervision, S.U. and J.A.A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Aratuo, D.N. Three Essays on Tourism Demand and Economic Development in the United States. 2018. Available online: https://researchrepository.wvu.edu/etd/3687/ (accessed on 8 August 2022).
2. Comerio, N.; Strozzi, F. Tourism and its economic impact: A literature review using bibliometric tools. *Tour. Econ.* **2018**, *25*, 109–131. [CrossRef]
3. Dabade, E.A.P.M.S. Sentiment Analysis of Twitter Data by Using Deep Learning and Machine Learning. *Turk. J. Comput. Math. Educ. TURCOMAT* **2021**, *12*, 962–970. [CrossRef]
4. Zainuddin, N.; Selamat, A.; Ibrahim, R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Appl. Intell.* **2018**, *48*, 1218–1232. [CrossRef]
5. Adwan, O.Y.; Al-Tawil, M.; Huneiti, A.; Shahin, R.; Abu Zayed, A.A.; Al-Dibsi, R.H. Twitter Sentiment Analysis Approaches: A Survey. *Int. J. Emerg. Technol. Learn.* **2020**, *15*, 79–93. [CrossRef]
6. Guo, Y.; Barnes, S.J.; Jia, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tour. Manag.* **2017**, *59*, 467–483. [CrossRef]
7. Lovera, F.A.; Cardinale, Y.C.; Homsi, M.N. Sentiment Analysis in Twitter Based on Knowledge Graph and Deep Learning Classification. *Electronics* **2021**, *10*, 2739. [CrossRef]
8. Wang, X.; He, X.; Cao, Y.; Liu, M.; Chua, T.-S. KGAT: Knowledge Graph Attention Network for Recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019. [CrossRef]
9. Huang, J. Design of Tourism Data Clustering Analysis Model Based on K-Means Clustering Algorithm. *Lect. Notes Data Eng. Commun. Technol.* **2022**, *136*, 373–380. [CrossRef]
10. Ayyub, K.; Iqbal, S.; Nisar, M.W.; Munir, E.U.; Alarfaj, F.K.; Almusallam, N. A Feature-Based Approach for Sentiment Quantification Using Machine Learning. *Electronics* **2022**, *11*, 846. [CrossRef]
11. Martín, C.A.; Torres, J.M.; Aguilar, R.M.; Diaz, S. Using Deep Learning to Predict Sentiments: Case Study in Tourism. *Complexity* **2018**, *2018*, 7408431. [CrossRef]
12. Akhtar, N.; Khan, N.; Mahroof Khan, M.; Ashraf, S.; Hashmi, M.S.; Khan, M.M.; Hishan, S.S. Post-COVID 19 tourism: Will digital tourism replace mass tourism? *Sustainability* **2021**, *13*, 5352. [CrossRef]
13. Guan, R.; Zhang, H.; Liang, Y.; Giunchiglia, F.; Huang, L.; Feng, X. Deep Feature-Based Text Clustering and Its Explanation. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 3669–3680. [CrossRef]
14. Bibi, M.; Abbasi, W.A.; Aziz, W.; Khalil, S.; Uddin, M.; Iwendi, C.; Gadekallu, T.R. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognit. Lett.* **2022**, *158*, 80–86. [CrossRef]
15. Coenen, L.; Verbeke, W.; Guns, T. Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods. *J. Oper. Res. Soc.* **2021**, *73*, 191–206. [CrossRef]
16. Tang, Y. Research on Word Vector Training Method Based on Improved Skip-Gram Algorithm. *Adv. Multimed.* **2022**, *2022*, 4414207. [CrossRef]
17. Fu, Y.; Hao, J.-X.; Li, X.; Hsu, C.H. Predictive Accuracy of Sentiment Analytics for Tourism: A Metalearning Perspective on Chinese Travel News. *J. Travel Res.* **2018**, *58*, 666–679. [CrossRef]
18. Akilandeswari, J.; Jothi, G. Sentiment Classification of Tweets with Non-Language Features. *Proc. Comput. Sci.* **2018**, *143*, 426–433. [CrossRef]
19. Neogi, A.S.; Garg, K.A.; Mishra, R.K.; Dwivedi, Y.K. Sentiment analysis and classification of Indian farmers' protest using twitter data. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100019. [CrossRef]
20. Stirparo, D.; Penna, B.; Kazemi, M.; Shashaj, A. Mining Tourism Experience on Twitter: A case study. *arXiv* **2022**, arXiv:2207.00816.

21. Kashmira, P.G.T.H.; Sumathipala, S. Generating Entity Relationship Diagram from Requirement Specification based on NLP. In Proceedings of the 2018 3rd International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 5–7 December 2018. [CrossRef]
22. Kejriwal, M. Knowledge Graphs. In *Applied Data Science in Tourism*; Springer: Cham, Switzerland, 2022; pp. 423–449. [CrossRef]
23. Zou, X. A Survey on Application of Knowledge Graph. *J. Physics Conf. Ser.* **2020**, *1487*, 012016. [CrossRef]
24. Bharadi, V.A. Sentiment Analysis of Twitter Data Using Named Entity Recognition. In *Computing and Communications Engineering in Real-Time Application Development*; Taylor & Francis: Oxfordshire, UK, 2022; pp. 101–122. [CrossRef]
25. Tweepy. Available online: https://www.tweepy.org/ (accessed on 8 August 2022).
26. Top 10 Open-Source Python Libraries for Machine Learning. Available online: https://blog.hackajob.co/top-10-open-source-python-libraries-and-frameworks-for-machine-learning-in-2022/ (accessed on 8 August 2022).
27. Twitter API Documentation | Docs | Twitter Developer Platform. Available online: https://developer.twitter.com/en/docs/twitter-api (accessed on 8 August 2022).
28. Pradha, S.; Halgamuge, M.N.; Vinh, N.T.Q. Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. In Proceedings of the 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam, 24–26 October 2019. [CrossRef]
29. Duong, H.-T.; Nguyen-Thi, T.-A. A review: Preprocessing techniques and data augmentation for sentiment analysis. *Comput. Soc. Netw.* **2021**, *8*, 1. [CrossRef]
30. Murillo, E.C.; De León, A.L.; Raventós, G.M. Evaluation of potential features present in short texts in spanish in order to classify them by polarity. *Appl. Sci.* **2017**, *40*, 21–32. [CrossRef]
31. Miah, S.U.; Sulaiman, J.; Bin Sarwar, T.; Naseer, A.; Ashraf, F.; Zamli, K.Z.; Jose, R. Sentence Boundary Extraction from Scientific Literature of Electric Double Layer Capacitor Domain: Tools and Techniques. *Appl. Sci.* **2022**, *12*, 1352. [CrossRef]
32. El Rahman, S.A.; AlOtaibi, F.A.; AlShehri, W.A. Sentiment Analysis of Twitter Data. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS 2019), Sakaka, Saudi Arabia, 3–4 April 2019.
33. Mishra, R.K.; Urolagin, S.; Jothi, A.A.J. A Sentiment analysis-based hotel recommendation using TF-IDF Approach. In Proceedings of the 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE 2019), Dubai, United Arab Emirates, 11–12 December 2019; pp. 811–815. [CrossRef]
34. Paliwal, S.; Mishra, A.K.; Mishra, R.K.; Nawaz, N.; Senthilkumar, M. XGBRS Framework Integrated with Word2Vec Sentiment Analysis for Augmented Drug Recommendation. *Comput. Mater. Contin.* **2022**, *72*, 5345–5362. [CrossRef]
35. Jang, B.; Kim, I.; Kim, J.W. Word2vec convolutional neural networks for classification of news articles and tweets. *PLoS ONE* **2019**, *14*, e0220976. [CrossRef]
36. Menon, T. *Empirical Analysis of CBOW and Skip Gram NLP Empirical Analysis of CBOW and Skip Gram NLP Models*; PDXScholar: Portland, OR, USA, 2020. [CrossRef]
37. Yang, X.; Yang, K.; Cui, T.; Chen, M.; He, L. A Study of Text Vectorization Method Combining Topic Model and Transfer Learning. *Processes* **2022**, *10*, 350. [CrossRef]
38. Lei, S. Research on the Improved Word2Vec Optimization Strategy Based on Statistical Language Model. In Proceedings of the 2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), Xi'an, China, 14–16 August 2020; pp. 356–359. [CrossRef]
39. Mehta, V.; Mishra, R.K. Machine Learning Based Fake News Detection on COVID-19 Tweets Data. In Proceedings of the International Conference on Computational Intelligence and Data Engineering, Vijayawada, India, 12–13 August 2022; pp. 89–96. [CrossRef]
40. Sharma, A.; Ghose, U. Sentimental Analysis of Twitter Data with respect to General Elections in India. *Proc. Comput. Sci.* **2020**, *173*, 325–334. [CrossRef]
41. Mishra, R.K.; Urolagin, S.; Jothi, J.A.A.; Neogi, A.S.; Nawaz, N. Deep Learning-based Sentiment Analysis and Topic Modeling on Tourism During COVID-19 Pandemic. *Front. Comput. Sci.* **2021**, *3*, 100. [CrossRef]
42. Mujahid, M.; Lee, E.; Rustam, F.; Washington, P.B.; Ullah, S.; Reshi, A.A.; Ashraf, I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Appl. Sci.* **2021**, *11*, 8438. [CrossRef]
43. Twitter Sentiment Analysis on Coronavirus Using Textblob. Available online: https://www.researchgate.net/publication/339998775_Twitter_Sentiment_Analysis_on_Coronavirus_using_Textblob (accessed on 3 January 2022).
44. Rakshitha, K.; Ramalingam, H.M.; Pavithra, M.; Advi, H.D.; Hegde, M. Sentimental analysis of Indian regional languages on social media. *Glob. Transit. Proc.* **2021**, *2*, 414–420. [CrossRef]
45. Moldagulova, A.; Sulaiman, R.B. Using KNN algorithm for classification of textual documents. In Proceedings of the 2017 International Conference on Information Technology (ICIT), Amman, Jordan, 17–18 May 2017; pp. 665–671.
46. Shahapure, K.R.; Nicholas, C. Cluster Quality Analysis Using Silhouette Score. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 6–9 October 2020; pp. 747–748. [CrossRef]
47. Ogbuabor, G.; Ugwoke, F.N. Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *Int. J. Comput. Sci. Inf. Technol. IJCSIT* **2018**, *10*, 27–37. [CrossRef]
48. Turki, H.; Taieb, M.A.H.; Ben Aouicha, M.; Fraumann, G.; Hauschke, C.; Heller, L. Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata. *Front. Res. Metrics Anal.* **2021**, *6*, 36. [CrossRef] [PubMed]
49. Kejriwal, M. Knowledge Graphs: A Practical Review of the Research Landscape. *Information* **2022**, *13*, 161. [CrossRef]

50. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space | Papers with Code. Available online: https://paperswithcode.com/paper/rotate-knowledge-graph-embedding-by (accessed on 3 January 2022).
51. SpaCy—Industrial-Strength Natural Language Processing in Python. Available online: https://spacy.io/ (accessed on 8 August 2022).
52. NetworkX—NetworkX Documentation. Available online: https://networkx.org/ (accessed on 8 August 2022).