Vol. 04, No. 01, pp. 34 –53 (2023)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Document Clustering in the Age of Big Data: Incorporating Semantic Information for Improved Results

Saad Hikmat Haji[1,*], Adel Al-zebari[2], Abdulkadir Sengur[3], Shakir Fattah Kak[2], Nasiba Mahdi Abdulkareem[2]
[1] *Department of Computer Sciences, Cihan University - Duhok, Duhok Kurdistan Region, Iraq, saad.hikmat91@gmail.com*
[2]*IT Dept., Technical College of Informatics Akre, Duhok Polytechnic University, Duhok, Iraq, (adel.ali@dpu.edu.krd, shakir.fattah@dpu.edu.krd, nasiba.mahdi@dpu.edu.krd)*
[3]*Electrical – Electronics Engineering Dept., Firat University, Elazig, Turkey, ksengur@firat.edu.tr*

Abstract

There has been a meteoric rise in the total amount of digital texts as a direct result of the proliferation of internet access. As a direct result of this, document clustering has evolved into a crucial method that must be used in order to successfully extract relevant information from big document collections. When employing the document clustering approach, documents are automatically sorted into groups whose members have a high degree of similarity to one another. These groups are created by applying the document clustering technique. Because they do not take into account the semantic linkages that exist between the texts, traditional clustering approaches are unable to provide an acceptable description of a collection of texts. This is because traditional clustering techniques. Document clusters, in which texts are ordered according to their meaning rather than their use of keywords, have been extensively utilized as a means of overcoming these challenges as a result of the incorporation of semantic information. This has been possible as a result of the fact that document clusters can group together related texts. In this investigation, we looked at a total of 27 distinct papers that were published over the previous five years and categorized the documents based on the semantic similarities that existed between the various pieces. A detailed literature evaluation is included to each and every one of the publications that were selected for further consideration. Comparative research is carried out on a wide variety of evaluation strategies, including as algorithms, similarity metrics, instruments, and processes. Following that, there is a drawn-out discussion that analyzes the similarities and differences between the activities.

Keywords: *Document Clustering, Semantic Document Clustering, Semantic Similarity, WordNet, Cloud computing.*
Received: November 11, 2022 / Accepted: February 15, 2023 / Online: February 19,2023

I. INTRODUCTION

In the context of today's world, the amount of data sets continues to expand at an exponential rate. Every single day, tens of thousands upon tens of thousands of news websites, both large and little, coming from all four corners of the globe upload their most recent news pieces to the internet [1-4]. These news websites are known as "news aggregators." Because it is physically impossible for a single person to see everything that takes place during an event or series of events, it is impossible for a single person to know all there is to know about the event or series of events. The amount of material, such as news articles and other types of content that can be found on the internet is expanding at the same pace as the number of websites that give information on the internet. This growth is occurring at the same time [5-7]. Text clustering is an approach that we make use of in order to find a solution to these issues, and we hope that it will be successful. This approach enables the grouping of individual articles into collections on the basis of the topics that are common to all of the individual articles in the collection. Clustering techniques are often used as one of the various ways that are used in the quest to make sense of the mountain of information that has been accumulated [8-10]. Clustering the items refers to the process of splitting a collection of things into smaller groups that all share similar features. This procedure is carried out by breaking a set of things into smaller groups. This process is carried out on a group of things that have been gathered together [11-13]. Clustering is a technique that can be used to organize data structures into a number of groups that are incompatible with one another and are referred to collectively as clusters [14-16]. Clustering is a strategy that may be employed. Items that belong to clusters that are diametrically opposing to one another are quite different from one another, in contrast to those that belong to the same cluster as one another and are thus comparable to one another. Both the technique of arranging connected documents into meaningful groups and the approach known as document clustering may be used to accomplish the same end goal [17-19].

Document clustering is a method that is somewhat comparable to the process of arranging linked documents into meaningful groups. These compiled sets of data may be used to develop explanations for a broad variety of topics and topics in general [20-22].

Conventional techniques of clustering employ the words in texts, which are also referred to as terms, as vectors; however,

doi:10.38094/jastt401143

they do not take into account any semantic correlations that may exist between terms. It is unable to generate cohesive clusters, and as a result, it suffers from a range of issues, some of which include synonymy, polysemy, and ambiguity, to name just a few [23-25]. This is because of the fact that it is unable to produce cohesive clusters. As a direct result of this, clustering algorithms are necessary in order to improve the process of paper clustering by injecting meaning into the clustering technique. This is a direct consequence of the fact that clustering algorithms are required [26-29].

When organizing information into coherent clusters, it is of the utmost importance to take full use of the semantic connections that exist in the texts between different words and the concepts that they represent. These connections are mentioned in the texts at various points. To be successful in accomplishing this goal, there must be semantic overlap between the concepts involved [30-32]. The semantic similarity that exists between two different pieces of literature or ideas is a metric that can be accurately measured, in contrast to the fuzzy similarity that can be quantified using reflect grammar. This is because semantic similarity refers to the degree to which two things share the same meaning (for example, string format). It is possible to offer a numerical representation of the semantic links that exist between language units and the concepts or occurrences that they refer to by making use of statistical techniques [33-35]. This is something that has been shown to be possible. As a direct result of this, a numerical definition is formed. This definition is founded on a comparison of the facts that support the relevance or depiction of life of the linguistic unit that is the subject of the current debate [36-38].

In the process of identifying the degree to which two distinct concepts are semantically connected to one another, WordNet is only one of many standard ways that are utilized. Other methods that are often used include: The WordNet database only includes English-language dictionaries since that is the language in which it was built [39-41]. The dictionaries in question are not offered in any other language editions. It organizes the words into groups of terms that have meanings that are comparable to one another, develops a distinct semantic link between the various classes of synonyms, and organizes the information in a manner that is consistent across the board. It also provides information that is standardized and concise [42-44]. WordNet performs quite well when it comes to similarity processes. This is because of the way that it controls nouns and verbs inside the "is a" relational hierarchy. The fact that this is the case gives WordNet an advantage over its rivals. WordNet organizes words into groups of interchangeable meanings (also known as a "synset" or "concept") based on the underlying ideas that are expressed by the words [45-47]. These groups of interchangeable meanings are then classified according to their respective categories. There are a few broad categories that may be used to classify the WordNet-based techniques in connection to the one-of-a-kind technical components and computational components that each method has. These categories can be used to organize the approaches in a more organized manner [48, 49]. These categories consist of things like: These categorizations are determined by the length of the route, the nature of the information, and the characteristics of the features themselves. In the following paragraphs, further information on each of these categories will be provided. The usage of WordNet is by far the most frequent way for detecting meanings and phrases that are semantically equivalent to one another [50-52]. This is true despite the fact that the literature study addressed a variety of different approaches to solving this problem. This approach is still the one that receives the greatest amount of support. In this article, we take a look at a few recent studies that use semantic similarity as a clustering approach. These publications have been published quite recently [53-55]. These articles were just recently made available online. In light of the findings presented here, we evaluate the merits of using semantic similarity in order to achieve the objective of this attempt. We discuss with each and every one of them some facts that we see as being of the utmost importance [56-58].

This paper provides an overview of numerous semantic document-clustering strategies. The second section covers the fundamentals of document clustering, such as document clustering, the standard approach, and the semantic approach. The third segment goes over the related work. In this section, we review previous research in semantic text clustering and the results of each study. Section four provides an outline and discussion of the articles that have been studied. Section five also serves as the paper's conclusion.

II. THEORITICAL BACKGROUND

This section addresses the classical and textual clustering of documents. The advantages and disadvantages of clustering documents are also discussed.

A. Document Clustering based on Traditional way

For representation documents, the standard text cluster model uses bags of words (BOW) [59, 60]. Sadly, the crucial drawback of this model is that it lacks semantic relationships between the wordings. It is observed that few algorithms significantly improve the effectiveness of document clustering. Traditional clustering algorithms draw from the fact that two different clusters cannot be identified. Clustering techniques or algorithms of traditional document clustering use features such as terms, phrases, and series. The vector space model is used in traditional document clustering approaches [61]. The text is presented as a vector in this model using a weighting scheme dependent on term frequency. However, a frequency-based system for weighting should only monitor the number of instances of a text; thus, semantical similarities of the document's content cannot be used in full for that model [62-64].

1. Semantic and Semantics Analysis

The academic discipline known as semiotics focuses on the study of signs and symbols that have a role in many forms of human communication. When taken to its most comprehensive level, semantics may be seen as a subdivision of semiotics. Because of this, not only do we have a semantics of utterances or actions in natural languages, but we also have a semantics of paraverbal or nonverbal behavior [65, 66]. This is a result of the fact that we have a semantics of utterances or actions in artificial languages. This encompasses a variety of mediums such as gestures, images and videos, logical systems or computer languages, sign languages used by the deaf, and potentially social contact in general. The phrase "the concept of

interpretation" is perhaps the most all-encompassing phrase that can be used to refer to the specific topic that is addressed by a semantic theory [67]. This is because the word "the idea of interpretation" comes from the Latin phrase "the notion of interpreting." This is due to the fact that the idea of interpretation includes both the literal and the metaphorical meanings of words. There is the possibility for an exceedingly wide variety of different interpretations, and these interpretations might vary greatly depending on the area of study or the theoretical framework that is being used [68].

On the other hand, semantic discourse analysis contains an element that may be interpreted in either an extensional or referential manner, depending on the context. To put it another way, we are interested in the potential referents of a large variety of different sentence chains that may be encountered in a wide variety of different discourses. These chains can be found in a vast variety of different places. Many referents of sentences have been assigned truth values, such as true and false, respectively, as a result of investigations that have been carried out throughout the course of philosophical and logical inquiry all throughout the course of history. Following that, an investigation into the semantics of the connectives that were used in the formulation of the compound propositions was carried out in order to ascertain whether or not the compound propositions were legitimate (e.g., logical and, or, if . . . then) [69]. If we follow this train of thought, we might need discourse semantics in order to determine the criteria by which the truth value of a discourse can be determined in a manner that is independent of the truth values of the sentences that make up the discourse. This would be possible only if the truth values of the sentences that make up the discourse were taken into account. This is due to the fact that without discourse semantics, defining the truth value of a conversation in such a way would be difficult. In spite of the fact that this may be an ideal objective, there are a number of compelling reasons why a strategy like this should not be used in the context in which it would be employed. For instance, the words and propositions that comprise a discourse are related to one another in a variety of additional ways in addition to the logical ones [70]. These connections exist in addition to the logical ones. Not only that, but a truth functional approach is excessively limited because it would only apply to discourses used in positive circumstances (as speech acts of assertion), and it would not apply to discourses such as inquiries, commands, promises, compliments, or accusations. In addition to that, a truth functional approach is excessively limited because it would only apply to discourses used in positive circumstances (as speech acts of assertion) [71]. Under addition to this, a truth functional method has an extremely narrow scope since it can only be used to discourses that are utilized in favorable conditions (as speech acts of assertion). In addition to this, the scope that a truth functional method may cover is too restricted.

Semantics is meaning study. It focuses on the association between meanings like words, sentences, signs, and symbols [72]. Semantic interpretation is analogous to vocabulary or logic. It tries to explain the essence of language as part of its construction and how it is expressed. Semantics examines the meaning of language in isolation as well as in the language itself. To understand the relationships between words, semantics examines the various ways in which their meanings can be related to one another. Sentences may be semantically related to one another in various ways [73].

The semantic analysis involves comparing syntactic structures of words, clauses, phrases, and paragraphs to their independent linguistic meanings. And also, It consists of removing the essential characteristics of such linguistic and cultural contexts [74].

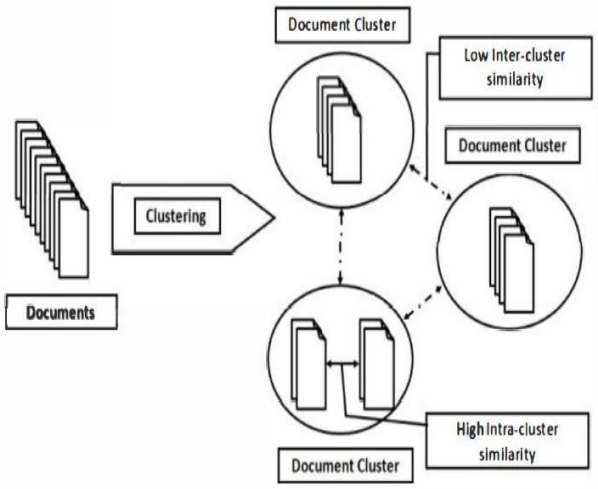

Fig.1 Document Clustering [62].

2. *Semantic Clustering*

Semantic Clustering is a strategy for generating keywords that concentrate on linked, associative keywords and phrases. Semantic Clustering focuses on grouping points within a data collection in separate groups (clusters) to ensure that two points in the same category are semantic equivalent [62, 75].

B. *Information Resources*

The availability of information resources is a critical factor in determining the semantic similarity of words. Many researchers have used WordNet as an information resource since their work on semantic similarity calculation between words. More recently, some have used a web search engine [41].

1. *WordNet*

WordNet is a computer lexicon that keeps track of the relationships that are shared by individual words. WordNet establishes a broad range of semantic links between the words in its database. Synonyms, hyponyms, and meronyms are some examples of the relationships that may be made between words. Synonyms are organized into synsets, and each synset has a brief explanation as well as various examples of how the term could be used. WordNet may be seen as an encyclopedia due to the fact that it incorporates and expands upon the concepts that are presented in dictionaries and thesauruses. People are able to access it by using a web browser; however, its principal use is in automated text analysis and artificial intelligence systems. People are able to access it by using a web browser. The first iteration of WordNet was developed in English; both its database and its software tools are presently offered for free

download on the WordNet website. They are licensed in a way that is analogous to that of BSD. There are word dictionaries available for more than two hundred distinct languages at this time [76].

There are many people who continue to argue that WordNet is an ontology, despite the fact that the people who created it have always denied that this is what it is. It is possible to see the interactions that take place between the hypernyms and hyponyms of noun synsets as distinct ties between a varieties of different types of cognition. As a consequence of this, the discipline of computer science has the potential to comprehend and make use of WordNet in the capacity of a lexical ontology. However, in order for such an ontology to be useful, it will first need to be updated so that the hundreds of fundamental semantic errors that it now has may be fixed. Among these flaws are things like generalizing common specializations to exclusive categories and duplicating specializations in the hierarchy. Both of these are examples of overgeneralization [40]. It is possible to convert WordNet into a lexical ontology that is suitable for the representation of knowledge by, among other things, decomposing the specialization relations into subtypes of instances of relations and assigning each of these subtypes IDs that are intuitively distinct from one another. In this way, WordNet can be transformed into a lexical ontology that is adequate for the representation of knowledge. Even though these sorts of corrections and transformations have been performed and documented as part of the integration of WordNet 1.7 into the cooperatively updatable knowledge base of WebKB-2, the majority of projects that claim to re-use WordNet for knowledge-based applications (typically, knowledge-oriented information retrieval) simply re-use it directly. This is despite the fact that such corrections and transformations have been performed as part of the integration of WordNet 1.7 into WebKB-2. In spite of the fact that WordNet 1.7 has been subjected to these sorts of repairs and alterations, this remains the case [77].

WordNet has been converted into a formal specification by utilizing both a bottom-up and a top-down approach to automatically extract association connections from WordNet and interpret these associations in terms of a set of conceptual relations, which are explicitly described in the DOLCE fundamental ontology. This process has allowed WordNet to be transformed into a formal specification. These connections have been understood in terms of a set of conceptual relations, which have been clearly specified in the DOLCE core ontology.

The vast majority of publications that make the claim that they have integrated WordNet into ontologies have not merely updated the content of WordNet when it was deemed necessary; rather, WordNet has been significantly re-interpreted and updated whenever it is relevant [78]. This is true for the vast majority of publications that make this claim. This occurred happen, for instance, either when the WordNet top-level ontology was reformed using the OntoClean-based technique or when WordNet was employed as a substantial source for generating the lower classes of the SENSUS ontology. Both of these instances are examples of when this took place. In each of these procedures, the OntoClean tool played an essential part [79].

A WordNet is a lexical dictionary of English. WordNet was created and managed by Professor George A. Miller of psychology at Princeton University Cognite Science Laboratory. Unlike other traditional lexicons, its groups terms into sets of synonyms known as synsets, offers short and general meanings, and documents the numerous semantic relationships between these synonym sets. Since it categorizes nouns and verbs into hierarchies of IS-A relationships, WordNet is exceptionally well suited for semantic similarity measurement. Figure 1 depicts a section of the WordNet2.1 IS-A hierarchy [80, 81].

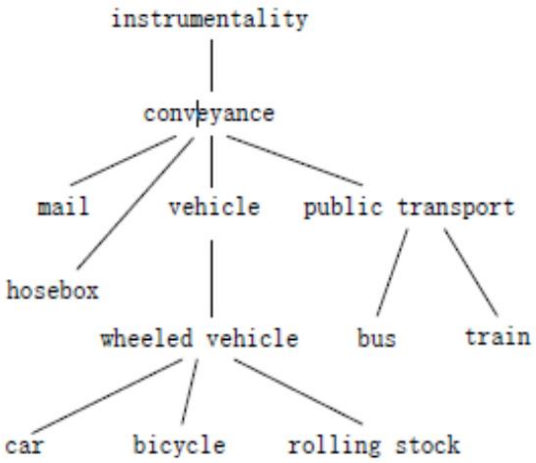

Fig.2 A fragment of WordNet 2.1 [82] .

2. Web Search Engine

Web search engines are specialized computer servers that are designed to search through the material of the World Wide Web in an effort to locate particular information. The bulk of the time, the search results that are returned to the user are presented to them in the form of a list (sometimes called hits) [81]. There is a potential that the hits comprise not just photographs and web pages, but also a wide variety of other file kinds. Some search engines will also go through directories and databases that are accessible to the general public, and they will only provide results that are relevant from these kinds of sources. It is vital to bear in mind that online directories and search engines are not the same thing. Search engines come in all shapes and sizes. Human editors are responsible for compiling online directories, while search engines rely on either algorithms or a mix of human and machine curation to choose which results to return [83].

Because of internet search engines, it is now considerably less difficult to discover particular material on the World Wide Web. Investigating the possibility of a link existing between two concepts has been the focus of a significant number of studies, many of which have made use of the results given by web-based search engines. The majority of the information that can be discovered in the search engine results that can be accessed on the internet is comprised of page counts and snippets [84]. The page count that is supplied by a search gives an approximation of the total number of pages that include the words and phrases that were searched for in that particular search. Snippets are bite-

sized chunks of text that are taken from a website by a search engine surrounding a query phrase and give context for the query [43]. Snippets may be found on the results page of a search. The "snippets" part of the results page of a search engine is the place where one may find snippets. The finding of semantically related measures across snippets has been put to use for a number of purposes, including the expansion of query scope, the disambiguation of names, and the mining of communities. These are just a few examples. When compared to uploading complete web pages, which may take a substantial amount of time depending on the size of the pages processing snippets is often a more efficient use of time [85].

In order to carry out semantic comparisons, a number of academics working in the academic sector have started to rely on snippets as a source of data. However, there have been studies that have just relied on the total number of pages, while other research have combined the total number of pages with a variety of other factors [82].

Because of the ever-changing nature of the Internet, assessing the World Wide Web is a challenging endeavor. This is something that everyone of us is well aware of and hence should come as no surprise. In 1999, there were over 40 million computers in more than 200 countries that were linked to the internet. More than three million of these computers are necessary for the operation of the server architecture that supports the World Wide Web. There are two primary causes that are primarily responsible for the meteoric rise in the number of Web servers [86]. Because of the availability of virtual hosts, a single server may be used to host several separate websites; but, due to the presence of virtual hosts, only a subset of those websites may actually be visible from the outside world. This presents the first problem. The second reason is that in 1998 there were a significantly smaller number of websites compared to what there are today since the URLs of certain websites do not begin with the prefix "www [87]." In 1998, industry experts predicted that there will be around 350 million websites on the Internet by the year 2000. The number of websites saw a growth rate of 20 million per month between 1997 and 1998, which resulted in the overall number of websites experiencing a growth rate that was more than double. On the internet, documents are most often found in the HTML format, followed by GIF and JPG (both of which are picture formats), ASCII files, Postscript, and ASP. By a significant margin, the most used format is HTML. The software GNU zip, Zip, and Compress are the ones that are used for compression the majority of the time. The vast majority of HTML pages are regarded as being non-standard since they do not properly comply to the specifications that are required by HTML [88]. When you initially view an HTML file, the description of the document type may not be the very first item that you see on the screen. They are typically rather little, with a mean size of around 5 kilobytes and a median size of just 2 kilobytes. The standard deviation for their sizes is approximately 1 kilobyte. There is a need for at least one picture to be included on an HTML page, and the maximum number of links that may be placed on that page ranges from five to fifteen. This suggests that the overwhelming majority of these links will take you to further websites that are housed on the same web server as this one [89].

C. Architecture and Organization of Cloud and Information Storage

The phrase "cloud storage" refers to a specific paradigm for the storage of digital information in computers, in which the information is held in distinct data centers that are located in distant locations. The information is kept on a large number of servers, each of which might be housed in a separate data center. Cloud service providers are responsible for ensuring that the data is always available and can be accessed, as well as for the facility's safety, maintenance, and proper operation. In order to maintain the safety and integrity of user, corporate, or application data, businesses and individuals often rent or buy storage space from third-party suppliers [90].

Cloud computing is commonly regarded as having been pioneered by Joseph Carl Robnett Licklider, who is credited with developing the concept via his work on ARPANET in the 1960s. 1983 was the year that CompuServe first started offering its consumer users free disk space for the purpose of data uploading and storing. AT&T established PersonaLink Services in 1994 as a web-based platform designed to facilitate interpersonal and commercial connections as well as corporate endeavors. They were the first companies to market their services using the term "the cloud," which referred to the electronic data storage facility that resembled a cloud and was used by the company. AWS S3, the cloud storage service offered by Amazon Web Applications, was first introduced in 2006. Since that time, it has been more popular as a storage provider for several well-known applications, such as SmugMug, Dropbox, and Pinterest. In 2005, the business cloud content management and file sharing service known as Box was first made available [91]. Since cloud storage is built on a similarly virtualized underlying architecture, it is identical to other aspects of cloud computing, including interfaces, almost immediate elasticity and scalability, multi-tenancy, and metered resources. This is because cloud computing as a whole is based on virtualization. Accessing and storing data in the cloud may be accomplished using either off-premises services (such as Amazon S3) or on-premises installations. Both are effective approaches (ViON Capacity Services). Cloud storage may take many forms, including object storage services hosted elsewhere, file storage, and block storage, just to name a few. The use of each of these unique types of cloud storage comes with its own set of advantages [92].

Block storage systems, such as Amazon Elastic Block Store (EBS), are essential to the operation of a significant number of commercial applications, such as databases, which need dependable access to massive volumes of data while experiencing as little lag as possible. This situation is comparable to that of storage area networks and direct attached storage (DAS), in certain respects (SAN) [93]. Organizations are only required to pay for the amount of storage space that they actually use, which is often the average over a certain period of time such as a month, quarter, or year. The fact that cloud storage incurs operating expenses rather than upfront capital expenditures does not mean that it is less expensive. By putting their data on the cloud, businesses have the potential to cut their overall use of energy by as much as 70 percent, making them friendlier to the environment [94].

Companies have the option of selecting off-site cloud storage, on-site cloud storage, or a hybrid combination of the two for their data storage needs. This choice can be made according to decision criteria such as continuity of operations (COOP), disaster recovery (DR), security (PII, HIPAA, SARBOX, IA/CND), and records retention laws, regulations, and policies. Initial direct cost savings potential is also a factor in this decision. Because they are already a part of the architecture of object storage, the additional effort, money, and technology that are necessary to ensure availability and security may potentially be avoided. The storage management obligations, including the purchase of additional storage space, will now be handled by the service provider. • This shift in responsibility will take effect immediately. Cloud storage provides users with a web service interface that enables them to have rapid access to a variety of resources and applications that are stored in the infrastructure of a different company [95].Using cloud storage, virtual machine images may be transferred between cloud and on-premises environments, or they can be imported from an on-premises environment into the cloud image library [96]. Alternatively, virtual machine pictures can be exported from the cloud to an on-premises environment. The usage of cloud storage allows for the transfer of virtual machine images across the accounts of different users as well as between different data centers. Cloud storage offers protection against data loss in the case of a natural disaster since it often utilizes a large number of backup servers located in various parts of the globe. Users are able to access their cloud storage in the same manner as if it were a local drive thanks to WebDAV. It is possible for companies that have more than one location to utilize it as a centralized file server so that they may exchange files with one another. Concerns that may arise: An Exhaustive Analysis of Information Security Protection in the Cloud. When you keep your data away from the physical location of your business, you leave yourself up to the possibility of being attacked [97].

Because it is kept in a greater number of locations, distributed data has a greater risk of being physically corrupted. Because data is constantly being replicated and moved, the use of a cloud-based architecture, for example, significantly increases the possibility that illegal data recovery may occur [98]. There are several instances of this, including the recycling of hard drives, the reuse of obsolete computers, and the reallocation of data storage space. The manner in which a consumer's data is replicated will be determined by both the service provider they use and the service level they choose. The use of encryption may secure critical data while still maintaining users' privacy. One of the available options for discarding sensitive information is known as crypto-shredding (on a disk) [99].

The likelihood of data being compromised (via improper methods, such as bribery or coercion) increases at an exponential rate whenever more people gain access to the information. The cloud storage provider will have a much larger team of technical staff with physical and electronic access to nearly all of the data at the entire facility or possibly the entire company. This is in contrast to the small team of administrators, network engineers, and technicians that might be responsible for the data of an individual company. If you compare this to the team that might be responsible for the data of an individual company, you'll notice that it's much smaller [100].

When the user, rather than the service provider, is in possession of the decryption keys, access to the data by staff members of the service provider is limited. If users are going to be able to share numerous data sets in the cloud with a variety of users, then a large number of decryption keys will need to be sent to the users over encrypted channels before they can be used [101]. The users will also need to be able to safely store and control these keys on their own devices. It is necessary to keep these keys in a secure place, which may result in additional costs. It is possible to get around this issue by using a cryptographic mechanism known as key aggregation. It expands the number of different networks that are available for information to go across. When it comes to connecting the data that is stored in the cloud, a wide area network (WAN) is required rather than just a local area network (LAN) or a storage area network (SAN) [102].

When data storage and networks are shared with a significant number of other users, it is possible for other customers to have access to the data that is stored there. It is possible that it is the result of hostile behavior; however, it is also possible that it is the consequence of human mistake, a malfunction, or a fault. This risk applies to any storage media, including cloud storage, and can't be completely eliminated [103]. When data are encrypted, the risk of information being stolen while it is in transit is significantly reduced. The data that is being sent from your device to the cloud service may be encrypted while it is in transit to ensure its safety [104]. Information that is retained with a service provider may be protected against unauthorized access by using encryption while it is "at rest." When data is encrypted in a system that is on-premises and connected to a cloud service, users have access to both varieties of encryption protection [105]. There are a number of various options available to you if you want to stay away from potential dangers. One option is to make use of a private cloud rather than a public cloud. This is one possible course of action. In addition, information may be imported in a secure manner, with the decryption key being kept locally. As a result, on-premise cloud storage gateways that include encryption options are often used in order to get access [106].

III. RELATED WORK

In this section, you will be able to see the outcomes of any and all research that has employed document clustering to generate a measure of semantic similarity. These findings will be presented to you in chronological order. The subjects of these investigations might be anything. Academics are placing a significant lot of emphasis on the semantic features of the articles that they are analyzing as a method of gaining more accuracy in the clustering efforts that they are engaged in. As a consequence of this, the authors draw from a variety of other lexical resources in addition to their own work. WordNet, Wikipedia, and Lucene are a few examples of the kind of resources that fall under this category. Within their work, the authors of the study made use of a broad range of clustering methods, techniques, similarity measures, datasets, and assessment metrics. In addition, the authors of the research made

use of a diverse selection of indicators for evaluating performance.

The bulk of typical keyword-based algorithms, as stated by Mahapatra et al. in 2020 [107], don't pay much attention to semantic meanings, which makes it tougher to reconstruct the real text. Since of this, there is a dilemma because it is necessary to grasp the semantic meanings of words in order to comprehend the context of a phrase. The authors of this paper presented a model for the retrieval of semantic information that is fuzzy and cluster-based. This approach examines the user query to evaluate the true relevance of the information that is being sought by the user, and it then offers the right resources depending on the conclusions of the inquiry. The traditional Boolean paradigm and the usage of a fuzzy cluster in the approach that is being discussed here are compared and contrasted in this section of the article. According to the data, it would seem that the strategy that consisted of maintaining the status quo is less successful than the paradigm, which consists of changing nothing. Having said that, the functionality of the system may be improved by taking into account the semantic linkages that exist between the various concepts. If you apply one of the numerous efficient ranking algorithms that can be found on this website, you will have a greater chance of achieving the results you are looking for. This is an additional approach you might use to get better outcomes. Throughout the whole of the procedure for determining the model's overall degree of performance, the two metrics of precision and recall serve as the primary foci of attention. The model that has received a lot of praise has a memory retention rate of 88% of the time, in addition to having an accuracy rating that is consistent across the board of 89%. After the input corpus has been lemmatized and processed by two stemmers, the embeddings that are generated by Doc2Vec are guaranteed to be independent of the technique that was used to prepare them. This accuracy of the model was evaluated in comparison to the TF-IDF model by Radu et al., 2020 [108], who integrated the Doc2Vec model with four different clustering algorithms in order to conduct their research (DBSCAN, LDA, K-Means, and Sp According to the results of the trials, one strategy for enhancing the continuity of a cluster is to maintain the linkages between the texts and the internal structures of the cluster while simultaneously cutting down on the cluster's total size.

In the year 2020, Fatimi and colleagues [109] shown that the technologies of the semantic web may be effectively deployed to successfully classify and evaluate vast volumes of text. It included persistent attempts to change the RDF code and served as the motivation for the building of a pipeline of semantic operations for RDF-based semantic text clustering. In addition to that, it brought attention to the current efforts that are being made to modify the RDF code. The most important advantage is that it offers a structure for semantic text clustering that is based on RDF data modeling. This is a big advantage. This is a highly practical aspect of the product. This is the most significant advantage that one could possibly have. This technique mixes a broad variety of research methodologies into a framework that is not only visually attractive but also extremely useful, and it is able to analyze textual data. In addition, this method is able to do data analysis on a variety of different types of information. The methods of machine learning and the concepts of the semantic web are used in order to achieve this goal. In addition to a wide range of other characteristics, such as RDF representation, topic modeling, clustering, and the capability to summarize document clusters, the system allows knowledge retrieval via the use of various reasoning methods. In addition to that, the system supports querying in RDF. We plan to achieve our goal of improving both the procedure of finding material as well as semantics by making use of the semantic web. This will allow us to fulfill both of our objectives simultaneously.

In addition, Zheng et al. (2019) [110] propose a model for an evolving neural network that would categorize the recorded picture examination pairings and the pathological tests by identifying semantic similarities involving overlapping body regions. This would be done using the results of the pathological tests and the recorded picture examination pairings. In order to do this, the outcomes of the pathological tests and the recorded pairings of the picture examinations would be used. Employing the recorded pairings of the photo inspections would be necessary in order to achieve this goal. Our model's performance and NER-based structure showed out to be highly outstanding when contrasted with more conventional models like keyword mapping, LSA, LDA, Doc2Vec, and Siamese LSTM. They have been able to make progress with their project by using the integrated diagram approaches to the medical ontology knowledge base that they have obtained from other sites. This has allowed them to make progress and has allowed them to advance their project. In addition to this, we made use of the LIME approach so that we could visually inspect the behavior of our model. According to the findings, the proposed model was able to properly infer semantic information from texts in an automated and objective manner, which allowed it to derive the required conclusions. This was shown by the fact that the model was able to correctly draw the appropriate conclusions. [Cause and effect] It is likely that when this information is used in combination with other medical data, it will make it much easier to evaluate the diagnostic accuracy of a picture. The conclusions reported before are given further weight by research that was carried out by Sarasvananda and colleagues [111]. The purpose of the research is to divide patients into various subsets by analyzing the particular medical issues that they have in common with one another as the criterion for grouping them together. ICD codes are often used as a point of reference by medical practitioners when making diagnoses for their patients. In order to establish a measure of the degree to which two ICD codes are connected to one another, K-means conducts an examination of the degree to which the two sets of codes share semantic properties. Within the scope of this study, analyses of semantic similarity have been carried out making use of the methodologies of Chodorow and Girardi, Leacock and Jaccard, and Rada, in that order. An assessment of the degree to which the clusters may be depended upon is carried out with the use of a method known as the silhouette coefficient. If the semantic similarity data are not defined, then the analysis of those data may provide clustering findings that are of a higher quality than those that would be acquired in the case that those data were not determined. This is based on the presumption that the results of the experiment will be the same as those that were predicted. The highest degree of accuracy that can be achieved without the use of semantic similarity is 91.78 percent, while the greatest

degree of accuracy that can be achieved without the utilization of semantic similarity is 84.93 percent.

Lwin, 2019 [112] The results of the experiments that used the approach were provided by utilizing the methodology of Swarm Optimization (PSO), which makes use of semantic links in the process of function extraction, and the domain Ontology, which makes use of semantic linkages in the process of function extraction. Both of these methodologies make use of semantic linkages in the process of function extraction (PSO). According to the findings of this research project, the method of using intracluster similarity is extremely successful. The fact that there are a great deal of similarities between the clusters lends further to the credibility of the result as a statement that can be relied upon. When applied to the dataset that News Group supplied, the proposed approach generated an F-measure score that was rather high (100 documents). In addition to this, the findings of Park et al., 2019 [113] need to be taken into account. Document clustering is a strong method that allows the collecting of data for natural language processing more easy to do (NLP). The gathering of readily accessible word representations to be included in the dataset is one of the many steps that are involved in the process of document clustering. The TF-IDF approach has become the industry standard for encoding words in text and has thus become the de facto norm. TF-IDF estimates the frequency of each word in the dataset and displays the results in a truncated one-dimensional vector in order to provide the bare minimum of functionality while avoiding the evaluation of the linkages and dependencies that exist between the words in the dataset. This allows it to provide the bare minimum of functionality while avoiding the evaluation of the linkages and dependencies that exist between the words in the dataset. Because of this, the data may be represented with the highest possible degree of precision. In the above example, there are several words that have meanings and grammatical properties that aren't absolutely necessary but also aren't immediately obvious to the reader. In recent years, there has been a proliferation of research approaches devoted to getting a knowledge of how the contextualization of data sets affects the semantic representation of word vectors. These methodologies have been developed in an effort to learn more about these impacts. The word vector concatenation study is an example of this kind of research technique. The embedded terms generators Word2vec and GloVe are the ones that are used the most often while dealing with a frequency matrix that only has one dimension. A vector representation of each of the words that are included in the dataset is provided in each line and row of the dataset's data representation. This representation might be located wherever inside the dataset itself. Wordplay has the ability to be carried out without any hitches when dealing with such a vast blank slate as the one being discussed. We used k-medoid to create cluster balance, which allowed us to use it once again for creating interword similarities, word embeddings via word2vec, and GloVe representations of the word vectors included inside the datasets. In addition, k-medoid was used in order to achieve the objective of cluster balancing, as Wrzalik and Krechel mention [114]. This was done in order to ensure success.

Wang and Koopman, 2017 [115] employed a few different categorization algorithms in order to uncover clusters that were included inside the Astro dataset. The author of this study built a semantic article representation and searched for clusters of similar topic articles as a part of this analysis. These two actions were carried out in their entirety. Both of our clustering solutions, OCLC-31 (K-Means) and OCLC-Louvain (an algorithm for finding Louvain groups), are founded on two techniques that have garnered a lot of attention in recent years: K-Means and Louvain group detection. K-Means was developed in the 1970s, and Louvain group detection was developed in the 1980s. An method known as OCLC-Louvain has been developed for the purpose of determining Louvain groups. OCLC-Louvain is a technique that may be used to determine whether or not a group fits the definition of Louvain (standing for Louvain).

Kolhe and Sawarkar, 2017 [116] based on the idea of text data, the authors in this paper presented a novel approach dubbed 'Semantic Lingo.' Wordnet is used to carry out this concept-driven approach. The established method defines the dominant notion and generates clusters based on these notions automatically. The paper validates the proposed definition using the widely accessible datasets ODP and AMBIENT. The proposed concept resulted in data clusters with high precision and recall. The efficiency of clusters was evaluated using three performance metrics. Precision, recall, and purity. Furthermore, K. and Chidambaram, 2016 [117] Presented A hybrid approach for the measurement of semantic similarity among documents. In information retrieval and text processing, Semantic similarity plays an important role. This work gave an overview of semantic similitude and its approaches. The test of semantic similarity measures the similarity of words, phrases, and documents. There are two divisions in the methods proposed: In the first instance, the proposed method utilizes an ontological model of similarity and a counting model of similarity to measure document similitudes. The scheme proposed uses ontology and corpus to assess the paper's similarity in a second fold. The proposed method achieves a high level of precision in document similarity estimates because of the hybrid approach.

In [118] Explained that a semantic similarity metric based on the documents displayed on the subject maps quickly becomes an industry standard for representing information, emphasizing previous research and removal. The documents are converted to map-based information, and the similitude of the two documents is understood as a connection between the typical patterns (sub-trees). Text mining dataset experiments have shown that this new similarity measure is more effective than before to classify text similarity measures. This test concludes that the proposed map-based similarity measurement in clustering document selection is highly promising because it provides a more consistent clustering than human structures.

This information was obtained from Al-Azzawy and Al-Rufaye, 2017 [119], which is the reference that you should use for the source. The words in a piece of written work may be organized into clusters using a technique that was developed by taking into account a number of criteria, including their morphological, syntactic, and semantic similarities. After that, separate word clusters were categorized according to the manner in which the aforementioned attributes interact with one another. Clustering is a kind of unsupervised machine learning that picks words on the basis of the common properties they all possess in order to calculate the distances that exist between each of the

words. The Similarity Function will yield clusters, and these clusters will be utilized to make a judgment of which attributes are comparable to one another. The authors classified each word using k-means clustering, which enabled them to determine how distant each phrase was from the others and how far apart each word was from one another. Additionally, they were able to determine how far apart each phrase was from one another. After the authors had completed their analysis of the usefulness of the system, they moved on to the next step of the evaluation process, which was the calculation of traditional evaluation metrics such as accuracy, reminder, and F-measurement.

A. Zandieh and S. Shakibapoor provided a suggestion for an automated system that would categorize texts into various different categories in their paper [120] that was published in 2017 and can be seen here. The results that they acquired by using more traditional methods are inferior to the results that they achieved by combining various parts of the issue. However, the results that they achieved by combining different aspects of the problem are superior. Clustering becomes more reliable and accurate when both the names of the phrases and the concepts that lay behind them are taken into consideration, as is the case with the semantic TF-IDF matrix. In other words, it provides a higher level of accuracy. Traditional clustering approaches address issues such as node naming, accuracy, efficiency, output impressibility, and cluster irrationality in hierarchies because of the intent- and context-based nature of these algorithms.

Reference: Nanayakkara and Ranathunga, 2018 [121]. The first stage in the method that is discussed here for clustering Sinhala news items is selecting which articles to utilize and then grouping those pieces together. This is done in order to cluster the news items. The execution of this method may be broken down into two distinct parts. The incorporation of object categorization into the system became possible as a result of the use of a similarity analysis approach that was based on a corpus. Because it accurately categorized 77% of the news pieces that originated from 9 distinct sources, it is quite an accomplishment for such a basic system to have accomplished this task. A statistical test known as the F-score was used so that we could evaluate how precise the categorization was. It is possible for the F-scores to fall anywhere between 0.1 and 0.1, with larger values suggesting more efficient clusters. The F-score of our algorithm was compared to the scores obtained by a variety of different approaches, and this comparison served as one of the primary foci. This was used in order to analyze the development of the course and make any modifications that were deemed necessary.

Afreen and Srinivasu (2017) [122] developed a method for clustering that was based on meanings that were not open to interpretation and lexical chains. This method was used to group together similar concepts. Lexical chains are utilized in order to extract essential semantic components that articulate the subject of texts, the number of clusters generated, and the assignment of suitable explanations for the clusters that are produced. Lexical chains are also utilized in order to assign suitable explanations for the clusters that are produced. In addition, a novel term-based semantic similarity test that is supplied with the intention of separating the many meanings of a word is presented. They show, in particular, that the coherence between major and minor characteristics in content clustering may be greatly enhanced by making use of lexical chain features. This improvement can have a considerable impact on the clustering results. This enhancement is seen in a number of scenarios, one of which is content clustering. They are able to provide evidence of this particular point (core semantics). Despite the fact that lexical chains have been used to a large extent in a variety of settings, this particular piece of research is one of only a select handful that has investigated the possibility that they have an effect on text categorization. WordNet will be the focus of our attention since Stanchev (2018) [80] gave detailed instructions on how to construct a probabilistic network by making use of the data that is already available in WordNet. However, our focus here will be on WordNet. By attaching an expression tag to each word in addition to the meaning of the phrase, they were able to make improvements to algorithms that had been employed in the past. Following that, they provided a demonstration of how to include information from DBPedia into the development of the graph. In addition to a methodology that relied only on WordNet, a cosine similarity metric approach was also used in order to assess the efficiency of the strategy that was suggested. The use of data obtained from DBPedia, as shown by the findings, leads in an improvement to the consistency of the Reuters-21578 benchmark. As a consequence of this, it is possible for the benchmark to approach the characteristics of the algorithm.

Ali and Melton (2018) [123] offered an alternative method for classifying articles according to the extent to which they are semantically connected to one another as compared to other articles. The mental science and theory that was laid forth in the graph served as the basis for the line of reasoning that they followed. In order to explore semantic memory in humans, they used a method known as ICAN, which is a mathematical model of the interaction between the mind and meaning. At the level of the text, ICAN semantic graphs were generated by using the methodology that was supplied by ICAN. In order to get rid of the semantic meaning that was connected to the ICAN graphs, an approach was used that combined the theory of graphs with the concept of feelings. After the ICAN graphs have been clustered using the Louvain community-detection technique, a corpus-level graph development procedure will be described in order to create the clustered documents required for the generation of corpus-graphs. This procedure will be carried out in order to create the graphs that will be used in the generation of corpus-graphs. These papers will serve as the foundation for the graphs that are created based on the corpus after it is completed. This step is performed after the ICAN graphs have been clustered in the previous stage. Their method was proved to be more accurate in clustering the data when compared to the LSA-based technique. This was shown via the use of the purity and entropy variables throughout the evaluation phase. The usage of the method of assessment allowed for the accomplishment of this goal. The authors highlighted the application of graph theory in emotional research for semantically based text categorization as an extra result of their study, which may be considered as a good development. This was done as an additional consequence of their work. This was carried out as an additional result of the study that they did. It is important that you be aware of this fact since the methodology that is based on ICAN cannot be regarded totally data-driven because it is constructed on top of WordNet's lexical ontology. In their paper from 2016, [124], Desai and Laxminarayana describe a document clustering approach that can be used to

gather and arrange documents that are highly similar under a single roof. This technique can be used to collect and organize documents that are extremely similar to one another. Since this is the initial stage of the method that is recommended, it is recommended that you begin by defining the corerefs that are included inside each unique document that is a part of the series. This will serve as the first step in the procedure. By combining WordNet and Semantic Similarity to determine the exact meaning of a word depending on the context in which it is being used, one may avoid the complications that are caused by polysemy and synonymy. When applied to the c1assic4 dataset, the clustering strategy that was recommended produces much better results. In addition to this, Bai and Jin provided a semantic network structure for the purpose of text representation in their 2016 paper [125]. This structure adds to the overall optimization of the network architecture by using a semantic similarity matrix. This is one of the ways that this structure helps. This structure was developed in order to allow for the representation of text inside it. The greatest common sub-graph approach, which is taken from graph theory, is then applied to the structure of the semantic network in order to determine the degree to which it is similar. This is done in order to figure out how similar the structures are to one another. This is done in order to ascertain the degree to which the two structures are connected to one another. The K-means methodology offers the possibility of expanding and improving text clustering in Chinese when used in the appropriate context. The fact that this approach makes use of K-means makes this option a real possibility. It would appear, on the basis of the findings of the trials, that the method that was proposed, which is based on the structure of a semantic graph, is more successful in describing the semantics of the data. This conclusion was reached in light of the fact that the proposed technique is based on the structure of a semantic graph. This verdict was arrived at after taking into consideration the fact that the method that was suggested is based on the composition of a semantic network. The employment of this tool will, as a direct result of its use, result in an increase in both the accuracy of the test for text similarity and the efficiency of the method for text clustering. The process of grouping text will be improved by making finer-grained adjustments to the appropriate parameters.

As was indicated in Bafna et al., 2016 [126], in order for the researchers to achieve their objectives with regard to this study, they used semantic document clustering on a broad variety of actual and simulated datasets. These datasets were compiled from a variety of sources, some of which being NEWS 20, Reuters, emails, and research journals. Additional approaches, such as fuzzy K-means, a hierarchical algorithm, and the strategy of Term Frequency-Inverse Document Frequency, were applied. In addition, a huge number of other methods were used. In the course of the testing phase of the project, cluster analysis is first performed on a limited dataset to see whether or not the project can really be completed. We use the strategy that has been shown to achieve the greatest level of overall success. The silhouette coefficient, the entropy, and the F-measure pattern have all been shown using various clusters of articles that have something in common with one another. This was done so that the functioning of the algorithm could be better understood in respect to each data set. Kumar and Bhatia (2020) [127] have presented an innovative method for identifying new material that can be included into web crawlers that are currently in use. This tactic may be found in the work that they have done. First, the text is summarized with the help of the ontology. Next, wordnet 3.0 is used to evaluate the level of semantic overlap that exists between the summary and the original text. This process is repeated until the text has been completely analyzed. After that stage is finished, the value of the hash is computed with the help of the winnowing approach. Before being included into the computation of the similarity index, the hash value of the document is first subjected to a multiplication by the Dice coefficient. This adds further evidence to support the hypothesis that the index is accurate. The text is categorized as a book or not based on whether or not it fulfills the requisite degree of similarity for novels. If it does not, then it is not classed as a book. The technique that was recommended was dreamt up by employing Visual Studio 2012 as the front-end programming tool and SQL as the back-end development tool. Both of these tools were employed for the building of the system. According to the results, the approach that was recommended reduces down on the amount of memory that is needed as well as the number of records that are collected, which in turn cuts down on the amount of time that the user spends looking for information. It has been suggested that the strategy that may be used to minimize the quantity of results that are irrelevant may also be used with other search engines such as Google, Yahoo, Bing, and Alta Vista.

Assigning a score or weight to each individual user query and taking into account the semantic connection that exists between the different queries that are entered by users is the method that Banik et al. 2018, [128] developed in order to improve the quality of search results. This method can be found on page 37. This was done in order to make the results more relevant to the subject that was being asked at the time. Articles that are collected by utilizing the Wikipedia API will be subjected to moderation with the help of this system. Utilizing keyword data and doing a computational comparison between the two publications may help establish the degree of resemblance that exists between two pieces of literature. This can be done by comparing the two publications side-by-side. This is something that may be done in order to evaluate the degree of resemblance that exists between the two different pieces of literature that have been compared. Since the authors admit looking into other potential solutions, it is reasonable to expect that the system the authors suggest will be an improvement on the work that has gone before it and will display cutting-edge skill in the relevant area. In addition, it is reasonable to expect that the system will display cutting-edge skill in the relevant area. When compared to earlier approaches to finding similarities, the fact that the proposed technique has a high Pearson correlation coefficient positions it in a position of superiority above those earlier approaches. In any case, Blokh and Alexandrov's 2017 [129] technique for classifying news data on genuine Facebook mass media news based on a similarity estimate created from an ontology methodology is useful. This method was developed to classify news based on actual Facebook posts. In order to evaluate the authenticity of news shared on Facebook, this approach was devised. The presentation of this proposal had the goal of compiling all of the authentic news stories about Facebook that have been published in major media. Due to the fact that we used all of these different

strategies in the course of our job, we were in a position to deliver rapid reactions to a wide variety of news clusters. For the purposes of this study, the researchers gathered a total of 415 000 messages from official news media accounts that were posted on Facebook. The information used in this investigation was taken from those accounts during the months of January 2014 and May 2017. The results of an inquiry showed that communications may be organized into groups of topics that are connected to one another, with each topic group concentrating on a different facet of the subject matter that is being investigated. When sorting these issues into their respective groupings, the total number of news bulletins was one of the factors that was taken into account. Since the topic has been gathered together, we are free to explore how the mass media of our choice covered it on several occasions within the time period that is being researched. Since the topic has been grouped together, this gives us more freedom.

Users may be able to overcome the challenges that are inherent in traditional keyword searches by using the ontology and document clustering strategy, which was published by Zafar et al., 2018 [130]. This method was developed by Zafar and his colleagues. The aforementioned study work devoted considerable attention to analyzing this tactic in great depth. The key goals of this tool are to cut down on the amount of time spent looking for information and to provide assistance in locating the particular pieces of information that you want. This study presents a method for semantically-based document clustering by basing it on the K-means clustering algorithm and a definition weight matrix that was constructed using a modified version of the TF-IDF method. The method is based on the K-means clustering algorithm and the definition weight matrix. Throughout the course of this investigation, these two aspects were refined and improved. The ontology that WordNet uses is constructed up of relations and traits, and each of these components has its own distinct weight in the system. The silhouette coefficient is a handy tool that can be used to measure the amount of cluster purity that is presently present. This can be done by comparing the original cluster with a purified version of the cluster.

In this study (Sravanthi and Srinivasu, 2017 [131], we examined and contrasted three different semantic strategies: path-based, feature-based, and cosine similarity. Route-based techniques are those that are based on a path, while feature-based techniques are those that are based on features. These strategies were analyzed and contrasted with one another in order to have a deeper comprehension of the similarities and differences between them. The suggested approaches begin with the preprocessing of pairs of sentences in order to identify the bag of words, and then continue to make use of similarity measures such as cosine similarity or path-based similarity measures. In the end, the bag of words is determined based on the provided methods. The ultimate purpose of these approaches is to generate a single sentence that is capable of recognizing new sentences as they are written. Comparisons between various similarity measures that have been utilized in the past and a feature-based assessment that is based on Wordnet are the key contributions that have been made by the method that has been proposed. These comparisons have been made possible by the method that has been proposed. They tag and lemmatize the nouns and verbs in order to compute the value of the similarity between the two sets of words. The approach that they use is one that is predicated on characteristics, and they use this technique in order to calculate the value. The author determines whether or not their project was successful by contrasting and comparing three distinct methods, as well as doing an analysis of the existing indicators based on a variety of factors. On the other hand, the findings of the study showed that the characteristic evaluation led to a higher semantic score.

One of the contributions that Duan and colleagues (2020) [132] made was the presentation of a technique for measuring semantic similarity, which was included in their paper. This method employs the usage of a learning representation matrix in conjunction with a semantic matrix in order to map various types of connections to the same semantic space. combining these two matrices into a single one. The results of the experiments show that the method that utilizes a wide variety of links of varying types achieves a higher level of accuracy in semantic computation when compared to more conventional approaches to semantic computing. This is demonstrated by the fact that the method uses more than one type of link. This is the inference that may be made in light of the results of the experiment that was carried out.

An examination of the effectiveness of the linguistic pre-processing and similarity functions that were used to organize Arabic tweets was carried out by researchers working with Abuaiadah et al. For the purpose of this particular study, the researchers used a regular optimization form of the K-Means approach in order to classify tweets as either positive or negative. This conclusion was arrived at by analyzing the content of the tweets. In every single experimental setting, it was shown that root-based stemming was far more effective than light-based stemming. The results produced by the Kullback-average Leibler similarity function are superior than those obtained by the Cosine, Pearson, Jaccard, and Euclidean similarity functions. This is the case when all of these functions are compared to one another.

Gao et al., 2015 [133] proposed a new technique for evaluating semantic, and while it is based not only on the notion of information content but also on the idea of border count. This method was presented in the context of assessing semantic. The suggested metric alters the least weighted distance between the ideas that were investigated in order to arrive at the conclusions concerning the semantic similarity; the measure also thoroughly discusses the link between parameters and the correlation value. The findings of the experiment make it abundantly clear that the proposed method achieves a high correlation value in terms of human grades, and it does so with distributions that are more even than the vast majority of comparable works found in the body of research pertaining to the correlation coefficient. Additionally, the results of the experiment make it abundantly clear that the proposed method achieves a high correlation value in terms of human grades with respect to the correlation coefficient. In addition, the data make it clearly evident that the suggested technique accomplishes a high correlation value in

terms of human grades by using a method that has been proved to accomplish a high correlation value in terms of

Kherwa and Bansal (2017) [134] used the concept of single value decomposition as a method to explain latent semantic analysis in their explanation of the concept. This was done in order to provide an explanation for the notion. The major emphasis of the approach known as latent semantic analysis is on making use of the larger-scale structure of a piece of written work in order to derive meaning from it. The goal of a latent semantic analysis is to get an understanding of the connection that exists between the words and the documents that are part of a given collection. This was achieved by throwing light on previously hidden linkages within the texts themselves, which was done in order to achieve this goal. In other words, this goal was reached by shedding light on previously concealed links. This research made use of Latent Semantic Analysis in order to determine the relationships that exist between the various words that were included in a large corpus of academic papers that were produced by a variety of natural language processing technologies. These papers were compiled using a variety of these technologies. The material included in these documents served as a resource for the investigation. Because it combines several words that have the same semantic, LSA is able to classify terms that are capable of having a variety of meanings and describe documents in a lower dimensional mental space than would be possible using any other method. This makes it possible to categorize terms that are capable of having more than one meaning. Because of this, it is now feasible to organize concepts that may take on a number of different connotations. As a result of LSA joining together words that have the same meaning, the process works as described above.

Nejad et al., 2017 [135] highlighted how they used text mining algorithms to extract important sections from these agreements and offer them to help in licensing assessment. The license review process was made easier as a result of this action. The full version of the term EULA is End Users Licensing Agreements. This is what the acronym stands for. In their work, the writers of this work provide a technique for extracting and classifying comparable sections of documents like licenses, taxes, and limits. This approach is provided by the authors of this work. This approach is described in detail within the bounds of the scope of this inquiry. We employ a large word embeddings database in conjunction with a distributional semantics method in order to carry out the clustering that is based on the semantic similarity of the words. This allows us to do the clustering in a manner that is accurate and efficient. Because of this, we are able to do the grouping based on the semantic similarities shared by the terms. An evaluation's findings show that the method significantly increased human comprehension, and improved feature-based clustering further decreased the amount of time required to digest the EULA.

In general, there are a significant number of writers (Zeng et al., 2017) [136] As was mentioned before, the TF-IDF methodology is used in order to get the document vector. After that comes the cosine similarity approach, which is used to determine how similar the texts are to one another. On the other hand, this mathematical method does not take into account the possible semantics of the words or articles that are being evaluated. This method focuses on the fundamental idea that is communicated by the phrase itself, which serves as the focal point of the methodology. Latent Semantic Analysis, on the other hand, makes use of the computation of TF-IDF in order to apply the semantic space. This is done to check that the information is correct. By using a strategy known as singular value decomposition, it is possible to assign each individual word as well as a portion of the text its very own location in the semantic space (SVD). If we go in this direction, we will be able to do semantic analysis, document clustering, and interaction between semantic class and document class all at the same time if we choose this course of action. Just before we go on to the latent semantic analysis, we are going to undertake a rapid review of the methods that are used to establish the way in which two texts are comparable to one another. This review will take place just before we move on to the latent semantic analysis. When compared to TF-IDF by itself, the findings suggest that LSA is capable of producing more accurate predictions of the laws than its counterpart.

In their work from 2018, Nie et al. [137] give a brand new semantic similarity that is produced not only from HowNet but also from previously established ones. This similarity is derived from both HowNet and the ones that were already in existence. When discussing memes, it's common to classify each instance as falling into one of these four broad groups. The writers of this work devise a method that generates a system that produces the system in order to bring about incremental changes in the result. This method takes into consideration the relative significance of a number of distinct groups and makes a system that produces the system. This way of thinking. This lends even more credence to the idea that the two ideas are connected to one another in a way that is meaningful in some way. In addition to that, the study discusses a method referred to as unit matching, which is an approach that may be used to identify texts that are similar to one another in some way. The findings of the study indicate that the approach that has been updated results in text clustering that is more accurate than the one that was previously used.

To further improve the accuracy of text clustering, Wang and Yu, 2019 [138] proposed using an approach to text clustering that is based on the semantic web. This would be a technique to further raise the accuracy of text clustering. The purpose of this was to further increase the accuracy of the text clustering, therefore this was done. In order to construct the text similarity matrix, we make use of spectral clustering, which is based on a semantic matrix that takes into account the meanings that are shared across the different texts. The parallels between the passages are what inspired the creation of this matrix (SS-SC). This technique increases the accuracy of clustering, generates a new system for sorting texts, and offers specific recommendations while taking into consideration the inter-word relationships. The results of this investigation go in that direction. This research illustrates that there is a link between using a weight measurement approach and enhanced clustering efficiency by demonstrating that there is a positive relationship between the two. Text clustering methods such as TCUSS (text

clustering based on somaticized similarity) and SS-SC were tested on the Google text corpus to see how well they perform their respective tasks. Both of these algorithms classify texts into groups according to the semantic connections that are shared by the texts under consideration. According to the findings, the accuracy of the conventional technique to clustering might need some improvement and could be improved by various alterations.

In line with the results that were obtained from the research that Yang et al. (2019) [139] carried out. While you are attempting to come up with a recommendation for an algorithm, it is helpful to take advantage of geographical commonalities. This strategy determines the degree to which a large number of words are related to one another by using spatial propagation techniques that are analogous to one another in their operation. In order to do this, it bases its method on the co-occurrence matrix, which serves as its basis. A cluster analysis such as this one makes it feasible to extract topical hotspot data from large amounts of text output in a manner that is not just realistic but also effective in a way that is made possible by the fact that it is now possible to do so.

IV. Discussion and Analysis

It should be noted that extensive feedback has been provided for the articles that are now being evaluated, as can be shown in Table 1. All of the information on the 27 papers, including similarity measures, datasets used, clustering methods, and assessment criteria, is presented in the form of a table below. In addition to that, there was a presentation that included a more in-depth explanation of the most significant findings from the study.

Both the TF-IDF matrix and the cosine are the similarity measures that are used the most in authors' work, with the cosine being the more well-known metric of the two. Our research provided evidence to support what was previously thought to be true, namely that the TF-IDF matrix and the cosine are the similarity measures that are used the most in authors' work. This was shown by the findings of our research, which offered proof (see Fig. 3). When combined with the cosine approach, the TF-IDF method improves the precision of the grouping process while also boosting the speed at which it may be completed. The TF-IDF and the WordNet ontology have been combined in an innovative way in order to enable researchers to discover phrases that are semantically related to one another and other phrases. The researchers came up with this approach on their own. It is envisioned to serve as a motor that propels the movement toward oneness. Through the use of the TF-IDF methodology, the phrases that are most important to the corpus are extracted from the database, and the analysis does not take into account the words that appear in the corpus the most often. In addition, the value of a tag controller in one data set service is compared to the value of that tag controller in other services so that the level of importance it has in that particular service may be determined. This is the simplest way that the idea may be put into practice; using the fundamental expression frequency of a sheet of paper in this way is the most easy application of the notion.

The term frequency–inverse document frequency (TF–IDF) model is now the approach that is employed the majority of the time when displaying documents in the vector space. This model can be found in the acronym TF–IDF. The traditional methods of clustering, on the other hand, are afflicted by a problem that is sometimes referred to as the "dimensional curse." This issue makes it very difficult to replicate the findings that are produced by employing these methods. Word embedding is a relatively new method of text analysis that promises to maintain the semantic ties that already exist between words while simultaneously lowering the dimensions of the document [108]. Word embedding is a relatively new method of text analysis that promises to maintain the semantic ties that already exist between words. Researchers at Microsoft Research are responsible for the development of this approach. In addition to this, we make use of the TF-IDF approach, which comprises removing the words that appear in the corpus the most frequently in decreasing order in order to hone down on the phrases that have the most relevance [108]. This allows us to zero in on the phrases that have the most impact.

The authors also used the cosine similarity test in order to evaluate the degree of correspondence that exists between the two distinct manuscripts. This evaluation was done in order to assess the degree of correspondence that exists between the two different manuscripts. It was shown that the performance of the Cosine similarity parameter was much greater than that of the other options that were presented. The cosine similarity is a measurement of proximity that is used in a wide variety of applications for the purposes of recovery and grouping [109]. It is one of the most often used methods for determining closeness.

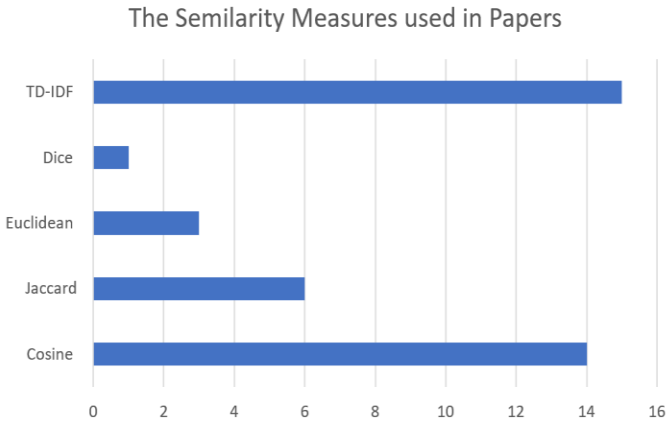

Fig. 4 Similarity Measures used in Papers

In spite of this fact, a number of writers use the Jaccard distance similarity scale in addition to the Euclidean distance similarity scale in the works that they have developed. What these writers have created may be seen here. The Jaccard similarity coefficient [123] is a metric that defines the degree to which a collection of texts and an ontology may be compared to one another based on the similarities that they have. This comparison may be made on the basis of the similarities that both of these entities have.

In addition, as can be seen in Figure 5, WordNet is the approach that is the most well-known and is used by academics on the most regular basis. This is because it allows for the most accurate results. The incorporation of the WordNet ontology provides the clusters with new meanings, which improves their quality and makes it feasible for users to browse the system in a number of various ways [130]. This component makes use of WordNet in order to discover synonyms and clear up any ambiguities that may have arisen about the intended meaning of the text. In order to do this, it first examines the surrounding information in order to determine which of the various alternative meanings of a word is the one that is the most similar to the meaning that is intended, and then it utilizes that meaning rather than any of the others [124]. When the WordNet ontology is included, more meanings are introduced into the clusters. As a consequence of this, users are provided with a wider variety of options when doing searches [130]. If you want your clustering to be more precise and effective, you might think about using Wordnet to construct the labels for your clusters [116].

In addition, many writers employ K-means and the hierarchical clustering algorithm in their work on document clusters because it is an easy technique of clustering that is also highly scalable [111]. This is because K-means is a simple method of clustering and the hierarchical clustering algorithm is. Both of these methods are simple and straightforward to implement.

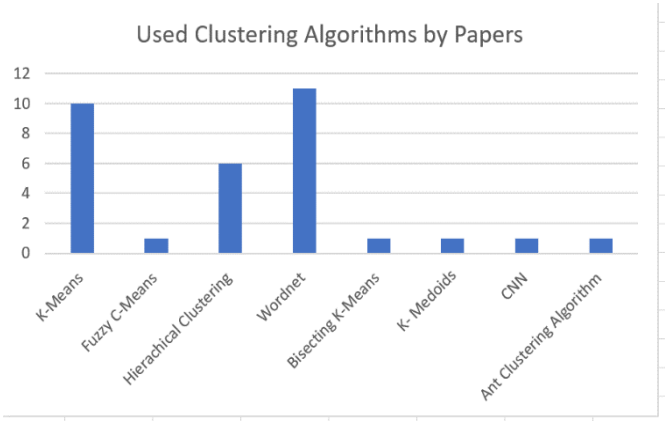

Fig.5. Used Clustering Algorithms by Papers.

writers the most often was K-Means. Fuzzy C-Means and Bisecting k-Means were utilized by a much lesser number of authors. The Bisecting K-Means Algorithm is a variation of the K-Means technique that may be used. This particular algorithm bears the name. In addition to being able to generate partitioned clustering, it can also construct hierarchical clustering. The technique has the capability of locating clusters that have a broad variety of sizes and organizational patterns. It's not completely out of the question to use this algorithm. When it comes to estimating the total amount of entropy, it performs much better than K-Means [124].

The authors of suggest a fuzzy-cluster-based semantic information recovery model that incorporates semantic user query awareness in order to infer the user's intent, determine the pages that are most relevant to the user's search, and place those pages into the appropriate categories as a means of increasing the level of search precision. This model can be found in [107]. The authors of suggest a fuzzy-cluster-based semantic information recovery model that incorporates semantic user query awareness in order to infer the user's intent

As can be seen in Figure 6, many authors make use of a wide variety of metrics while performing an analysis of the findings of their study in order to provide the most accurate picture possible. Some of the metrics that are covered here include recall, precision, F-score, F-measure, purity, the Rand Index, and entropy. Also included are these metrics. The accuracy and recall tests are two methods that are used often and to a large degree in the process of establishing whether or not the clusters are helpful and reliable. This is done in order to determine whether or not the clusters are useful and reliable.

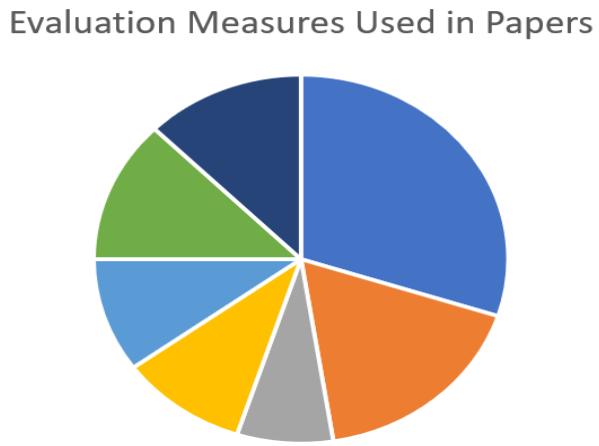

Fig. 5 Evaluation Measures Used in Papers.

TABLE I SUMMERY OF REVIEWED PAPERS USING SEMANTIC SIMILARITY BASED ON DOCUMENT CLUSTERING

Author	Year	Semantic Approach	Data Set	Clustering Algorithms/ Methods	Semantic Measures	Evaluation Measures
Fatimi et al [109]	2020	Semantic Text Clustering	DBPedia RDF	K-Means, Hierarchical Clustering	Cosine, TD-IDF	Precision
Radu et al. [108]	2020	Semantic Document Clustering	News Data from internet	K-Means	Cosine, TF-IDF,	Adjusted Rand Index
Kumar and Bhatia [127]	2020	Novelty Detection based on semantic Similarity for search engine	Random Document in different fields	Wordnet	Dice	Redundancy Removal, Memory overhead

Mahapatra et al. [107]	2020	Information retrieval based on semantic document clustering	----------	Fuzzy C-Means	Cosine	Precision and recall
Zheng et al. [110]	2019	Medical Text Semantic Clustering	16354 Image and Pathology Report from 1926 Patients	CNN	Jaccard and Cosine	Precision, Recall, F1-Score
Sarasvananda et al [111]	2019	Hospitalization cost estimation using semantic similarities	244 Patient Data	K-Means	silhouette coefficient, Jaccard	Precision and F1 Score
Lwin, [112]	2019	Document Clustering in web	News Dataset (100 Documents) from Amazon, Alibaba, BBC	Particle swarm optimization (PSO), Hierarchical Clustering	Jaccard	Precision, F1-Score, Recall
Park et al., [112]	2019	Advanced Document Clustering	FakeNewsAMT, SQuAD, Yahoo Answers, Reuters	Own Algorithm	Cosine	Entropy
Wrzalik and Krechel, [114]	2019	Semantic Document Representation	20 News, Reuters, Classic , Amazon and wiki	K- Medoids	TD-IDF Similarity	Fuzzy and sharp
Nanayakkara and Ranathunga [121]	2018	Sinhala News Clustering articles	Articles of Sinhala News	Wordnet	TD-IDF, Cosine	Adjust Rand Index
Stanchev [80]	2018	Document clustering based on semantic similarity	Reuters-21578 benchmark	K-Means	TD-IDF, Cosine	Precision, Recall, F-Measure
Ali and Melton, [123]	2018	Using Sematic Similarity for document text clustering	BBC, BBC Sport, Classic 300	K-Means	Jaccard, Euclidean,	Purity and Entropy
Banik et al [128]	2018	Wikipedia Documents Clustering based on semantic similarity measures	Wikipedia	Ant Clustering Algorithm	Cosine	F-Measure , Precision, Recall
Zafar et al [130]	2018	Semantic Document Analysis and clustering	Banks Annual Reports	Wordnet, K-Means	TD-IDF	F-Score
Wang and Koopman, [115]	2017	Article Clustering based on semantic Similarity	Astro dataset	Louvain Method and K-Means	Cosine	F1- Score and Adjusted F1 Score
Kolhe and Sawarkar [116]	2017	Web Search Clustering	Open Directory Project(ODP), AMBIENT	Wordnet	TD-IDF, Cosine	Precision , Recall and Purity
Al-Azzawy and Al-Rufaye, [119]	2017	10 news text and 10 short stories	Arabic News	K-Means	TD-IDF	Precision, F-Measure, Recall
Sravanthi and Srinivasu, [131]	2017	Similarity between sentences based on semantic	1000 pair of sentences from Microsoft research corpus	Wordnet	Cosine, path based, and feature based	----------
Zandieh and Shakibapoor [120]	2017	Text clustering based on semantic	100 Documents from 20 newsgroups	Hierarchical Clustering , Wordnet	TD-IDF, Cosine, Jaccard	Adjust Rand Index
Afreen and Srinivasu [122]	2017	Document clustering based on semantic measures	Group of 20 News	WordNet	TD-IDF	----------
Nejad et al., [135]	2017	EULA Summarization based on Sematic	1000 EULA from internet	Hierarchical Agglomerative Clustering	----------	F-Measure and Rand Index
Blokh and Alexandrov [129]	2017	News Semantic Clustering	Several official Facebook Mass Media News	Wordnet	----------	----------
Rafi et al., [118]	2016	Semantic Similarity measure for clustering of documents	NEWS20, OSHUMED, Classic, Webkb, Reuters	Hierarchical clustering technique	Euclidean, Cosine, Jaccard, KLD, Tm-sim	Purity, Entropy
K. and Chidambaram [117]	2016	Semantic Similarity measures between documents	50 documents from Australian Broadcasting corporation news mail service	Wordnet	Jaccard, cosine, dice, TD-IDF	F-Measure, Precision , Recall

Desai and Laxminaray ana [124]	2016	Document Clustering based on semantic	Classic4	Bisecting k-means, Wordnet	Cosine, TD-IDF,	Purity
Bai and Jin [125]	2016	Semantic Text Clustering	2000 articles including sports, politic, computer and environment	K- Means, hierarchical clustering	TD-IDF	Recall and Precision
Bafna et al [126]	2016	Clustering Documents using Semantic Approach	45 Documents from Research paper, News20, Reuters and Emails	Fuzzy K- Means, hierarchical agglomerative clustering	TD-IDF, Cosine	Entropy, F-Measure, Purity

V. CONCLUSION

The purpose of this essay is to offer a comprehensive analysis of the method of semantic text clustering, and to do so, we will be looking at it in great detail. In order to realize the objective of semantic text clustering, research will be carried out into both tried-and-true methods and pioneering new ways of doing things. Before being arranged into clusters, individual words are stopped, stemmed, and tokenized as part of the initial phase of both techniques. Despite this, the English WordNet dataset is one of the tools that is used most often for successful clustering. This is because it contains a large number of words. Because of this, various different approaches to clustering are used in order to bring the process to a successful conclusion. The K-mean approach is the one that is used the most because of how easy it is to put into action. This is due to the fact that it provides the most accurate results. As a direct result of this, it is the one that sees the most

action. It is likely that if you use additional methods, such as the bisecting K-mean, the fuzzy C-mean, and the hierarchical agglomerative clustering, you will be able to get even closer to the clusters than you were able to before. These methods include the fuzzy C-mean, the hierarchical agglomerative clustering, and the hierarchical bisecting K-mean. Word meaning disambiguation and wordnet are two of the tools that semantically based methods to clustering use in order to document it. [Clustering] These approaches provide results that are more accurate and clusters that are more consistent when compared to processes that are more traditionally used. This is because these methods are more consistent with one another. We examined the efficacy of the various approaches to clustering by contrasting and comparing them with one another. We did this so that we could get a more in-depth grasp of the benefits and drawbacks that are linked with each strategy.

REFERENCES

[1] M. M. Sadeeq, N. M. Abdulkareem, S. R. Zeebaree, D. M. Ahmed, A. S. Sami, and R. R. Zebari, "IoT and Cloud computing issues, challenges and opportunities: A review," Qubahan Academic Journal, vol. 1, pp. 1-7, 2021.

[2] O. Alzakholi, H. Shukur, R. Zebari, S. Abas, and M. Sadeeq, "Comparison among cloud technologies and cloud performance," Journal of Applied Science and Technology Trends, vol. 1, pp. 40-47, 2020.

[3] A. A. Salih, S. Zeebaree, A. S. Abdulraheem, R. R. Zebari, M. Sadeeq, and O. M. Ahmed, "Evolution of mobile wireless communication to 5G revolution," Technology Reports of Kansai University, vol. 62, pp. 2139-2151, 2020.

[4] H. Shukur, S. Zeebaree, R. Zebari, D. Zeebaree, O. Ahmed, and A. Salih, "Cloud computing virtualization of resources allocation for distributed systems," Journal of Applied Science and Technology Trends, vol. 1, pp. 98-105, 2020.

[5] A. A. Yazdeen, S. R. Zeebaree, M. M. Sadeeq, S. F. Kak, O. M. Ahmed, and R. R. Zebari, "FPGA implementations for data encryption and decryption via concurrent and parallel computation: A review," Qubahan Academic Journal, vol. 1, pp. 8-16, 2021.

[6] R. R. Zebari, S. R. Zeebaree, and K. Jacksi, "Impact analysis of HTTP and SYN flood DDoS attacks on apache 2 and IIS 10.0 Web servers," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 156-161.

[7] S. R. Zeebaree, K. Jacksi, and R. R. Zebari, "Impact analysis of SYN flood DDoS attack on HAProxy and NLB cluster-based web servers," Indones. J. Electr. Eng. Comput. Sci, vol. 19, pp. 510-517, 2020.

[8] L. M. Haji, S. Zeebaree, O. M. Ahmed, A. B. Sallow, K. Jacksi, and R. R. Zeabri, "Dynamic resource allocation for distributed systems and cloud computing," TEST Engineering & Management, vol. 83, pp. 22417-22426, 2020.

[9] B. T. Jijo, S. Zeebaree, R. R. Zebari, M. Sadeeq, A. B. Sallow, S. Mohsin, et al., "A comprehensive survey of 5G mm-wave technology design challenges," Asian Journal of Research in Computer Science, vol. 8, pp. 1-20, 2021.

[10] H. M. Yasin, S. Zeebaree, M. Sadeeq, S. Y. Ameen, I. M. Ibrahim, R. R. Zebari, et al., "IoT and ICT based smart water management, monitoring and controlling system: A review," Asian Journal of Research in Computer Science, vol. 8, pp. 42-56, 2021.

[11] L. M. Haji, O. M. Ahmad, S. Zeebaree, H. I. Dino, R. R. Zebari, and H. M. Shukur, "Impact of cloud computing and internet of things on the future internet," Technology Reports of Kansai University, vol. 62, pp. 2179-2190, 2020.

[12] R. R. Zebari, S. Zeebaree, K. Jacksi, and H. M. Shukur, "E-business requirements for flexibility and implementation enterprise system: A review," International Journal of Scientific & Technology Research, vol. 8, pp. 655-660, 2019.

[13] A. B. Sallow, M. Sadeeq, R. R. Zebari, M. B. Abdulrazzaq, M. R. Mahmood, H. M. Shukur, et al., "An investigation for mobile malware behavioral and detection techniques based on android platform," IOSR Journal of Computer Engineering (IOSR-JCE), vol. 22, pp. 14-20, 2020.

[14] N. M. Salih and K. Jacksi, "State of the art document clustering algorithms based on semantic similarity," Jurnal Informatika, vol. 14, pp. 58-75, 2020.

[15] R. Ibrahim, S. Zeebaree, and K. Jacksi, "Survey on semantic similarity based on document clustering," Adv. sci. technol. eng. syst. j, vol. 4, pp. 115-122, 2019.

[16] O. H. Jader, S. Zeebaree, and R. R. Zebari, "A state of art survey for web server performance measurement and load balancing mechanisms," International Journal of Scientific & Technology Research, vol. 8, pp. 535-543, 2019.

[17] S. Zeebaree, R. R. Zebari, K. Jacksi, and D. A. Hasan, "Security Approaches For Integrated Enterprise Systems Performance: A Review," Int. J. Sci. Technol. Res, vol. 8, 2019.

[18] K. Jacksi, R. K. Ibrahim, S. R. Zeebaree, R. R. Zebari, and M. A. Sadeeq, "Clustering documents based on semantic similarity using HAC and K-mean algorithms," in 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020, pp. 205-210.

[19] S. R. Zeebaree, H. M. Shukur, L. M. Haji, R. R. Zebari, K. Jacksi, and S. M. Abas, "Characteristics and analysis of hadoop distributed systems," Technology Reports of Kansai University, vol. 62, pp. 1555-1564, 2020.

[20] P. Y. Abdullah, S. Zeebaree, K. Jacksi, and R. R. Zeabri, "An hrm system for small and medium enterprises (sme) s based on cloud computing technology," International Journal of Research-GRANTHAALAYAH, vol. 8, pp. 56-64, 2020.

[21] H. Dino, M. B. Abdulrazzaq, S. Zeebaree, A. B. Sallow, R. R. Zebari, H. M. Shukur, et al., "Facial expression recognition based on hybrid feature extraction techniques with different classifiers," TEST Engineering & Management, vol. 83, pp. 22319-22329, 2020.

[22] H. Malallah, S. Zeebaree, R. R. Zebari, M. Sadeeq, Z. S. Ageed, I. M. Ibrahim, et al., "A comprehensive study of kernel (issues and concepts) in different operating systems," Asian Journal of Research in Computer Science, vol. 8, pp. 16-31, 2021.

[23] S. Zeebaree, R. R. Zebari, and K. Jacksi, "Performance analysis of IIS10. 0 and Apache2 Cluster-based Web Servers under SYN DDoS Attack," TEST Engineering & Management, vol. 83, pp. 5854-5863, 2020.

[24] M. B. Abdulrazaq, M. R. Mahmood, S. R. Zeebaree, M. H. Abdulwahab, R. R. Zebari, and A. B. Sallow, "An analytical appraisal for supervised classifiers' performance on facial expression recognition based on relief-F feature selection," in Journal of Physics: Conference Series, 2021, p. 012055.

[25] H. Shukur, S. Zeebaree, R. Zebari, O. Ahmed, L. Haji, and D. Abdulqader, "Cache coherence protocols in distributed systems," Journal of Applied Science and Technology Trends, vol. 1, pp. 92-97, 2020.

[26] S. Fahad and W. Yafooz, "Review on semantic document clustering," International Journal of Contemporary Computer Research, vol. 1, pp. 14-30, 2017.

[27] D. A. Hasan, S. R. Zeebaree, M. A. Sadeeq, H. M. Shukur, R. R. Zebari, and A. H. Alkhayyat, "Machine Learning-based Diabetic Retinopathy Early Detection and Classification Systems-A Survey," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021, pp. 16-21.

[28] M. R. Mahmood, M. B. Abdulrazzaq, S. Zeebaree, A. K. Ibrahim, R. R. Zebari, and H. I. Dino, "Classification techniques' performance evaluation for facial expression recognition," Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, pp. 176-1184, 2021.

[29] S. Zeebaree, L. M. Haji, I. Rashid, R. R. Zebari, O. M. Ahmed, K. Jacksi, et al., "Multicomputer multicore system influence on maximum multi-processes execution time," TEST Engineering & Management, vol. 83, pp. 14921-14931, 2020.

[30] B. R. Ibrahim, F. M. Khalifa, S. R. Zeebaree, N. A. Othman, A. Alkhayyat, R. R. Zebari, et al., "Embedded system for eye blink detection using machine learning technique," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021, pp. 58-62.

[31] H. I. Dino, S. Zeebaree, O. M. Ahmad, H. M. Shukur, R. R. Zebari, and L. M. Haji, "Impact of load sharing on performance of distributed systems computations," International Journal of Multidisciplinary Research and Publications (IJMRAP), vol. 3, pp. 30-37, 2020.

[32] R. R. Zebari, S. R. Zeebaree, A. B. Sallow, H. M. Shukur, O. M. Ahmad, and K. Jacksi, "Distributed denial of service attack mitigation using high availability proxy and network load balancing," in 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020, pp. 174-179.

[33] Z. A. Younis, A. M. Abdulazeez, S. R. Zeebaree, R. R. Zebari, and D. Q. Zeebaree, "Mobile Ad Hoc Network in Disaster Area Network Scenario: A Review on Routing Protocols," International Journal of Online & Biomedical Engineering, vol. 17, 2021.

[34] H. M. Shukur, S. R. Zeebaree, R. R. Zebari, B. K. Hussan, O. H. Jader, and L. M. Haji, "Design and implementation of electronic enterprise university human resource management system," in Journal of Physics: Conference Series, 2021, p. 012058.

[35] H. I. Dino, S. Zeebaree, A. A. Salih, R. R. Zebari, Z. S. Ageed, H. M. Shukur, et al., "Impact of Process Execution and Physical Memory-Spaces on OS Performance," Technology Reports of Kansai University, vol. 62, pp. 2391-2401, 2020.

[36] S. M. Mohammed, K. Jacksi, and S. Zeebaree, "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms," Indonesian Journal of Electrical Engineering and Computer Science, vol. 22, pp. 552-562, 2021.

[37] K. H. Sharif, S. R. Zeebaree, L. M. Haji, and R. R. Zebari, "Performance measurement of processes and threads controlling, tracking and monitoring based on shared-memory parallel processing approach," in 2020 3rd International Conference on Engineering Technology and its Applications (IICETA), 2020, pp. 62-67.

[38] L. Haji, R. Zebari, S. Zeebaree, W. Abduallah, H. Shukur, and O. Ahmed, "GPUs impact on parallel shared memory systems performance," Int. J. Psychosoc. Rehabil, vol. 24, pp. 8030-8038, 2019.

[39] G. M. Zebari, S. Zeebaree, M. M. Sadeeq, and R. Zebari, "Predicting Football Outcomes by Using Poisson Model: Applied to Spanish Primera División," Journal of Applied Science and Technology Trends, vol. 2, pp. 105-112, 2021.

[40] K. P. M. Kumar, J. Mahilraj, D. Swathi, R. Rajavarman, S. R. Zeebaree, R. R. Zebari, et al., "Privacy Preserving Blockchain with Optimal Deep Learning Model for Smart Cities," CMC-COMPUTERS MATERIALS & CONTINUA, vol. 73, pp. 5299-5314, 2022.

[41] R. K. Ibrahim, S. R. Zeebaree, K. Jacksi, S. H. Ahmed, S. M. Mohammed, R. R. Zebari, et al., "Clustering Document based on Semantic Similarity Using Graph Base Spectral Algorithm," in 2022 5th International Conference on Engineering Technology and its Applications (IICETA), 2022, pp. 254-259.

[42] Z. N. Rashid, S. R. Zeebaree, R. R. Zebari, S. H. Ahmed, H. M. Shukur, and A. Alkhayyat, "Distributed and Parallel Computing System Using Single-Client Multi-Hash Multi-Server Multi-Thread," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021, pp. 222-227.

[43] Z. N. Rashid, S. R. Zeebaree, M. A. Sadeeq, R. R. Zebari, H. M. Shukur, and A. Alkhayyat, "Cloud-based Parallel Computing System Via Single-Client Multi-Hash Single-Server Multi-Thread," in 2021 International Conference on Advance of Sustainable Engineering and its Application (ICASEA), 2021, pp. 59-64.

[44] A. B. Sallow, S. R. Zeebaree, R. R. Zebari, M. R. Mahmood, M. B. Abdulrazzaq, and M. A. Sadeeq, "Vaccine tracker/SMS reminder system: design and implementation," ISSN (Online), pp. 2581-6187, 2020.

[45] R. R. Zebari, S. R. Zeebaree, Z. N. Rashid, H. M. Shukur, A. Alkhayyat, and M. A. Sadeeq, "A Review on Automation Artificial Neural Networks based on Evolutionary Algorithms," in 2021 14th International Conference on Developments in eSystems Engineering (DeSE), 2021, pp. 235-240.

[46] O. H. Jader, S. R. Zeebaree, R. R. Zebari, H. M. Shukur, Z. N. Rashid, M. A. Sadeeq, et al., "Ultra-Dense Request Impact on Cluster-Based Web Server Performance," in 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA), 2021, pp. 252-257.

[47] A. S. Abdulraheem, A. I. Abdulla, and S. M. Mohammed, "Enterprise resource planning systems and challenges," Technology Reports of Kansai University, vol. 62, pp. 1885-1894, 2020.

[48] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," Journal of Applied Science and Technology Trends, vol. 1, pp. 56-70, 2020.

[49] H. S. Yahia, S. Zeebaree, M. Sadeeq, N. Salim, S. F. Kak, A. Adel, et al., "Comprehensive survey for cloud computing based nature-inspired algorithms optimization scheduling," Asian Journal of Research in Computer Science, vol. 8, pp. 1-16, 2021.

[50] S. Zeebaree, S. Ameen, and M. Sadeeq, "Social media networks security threats, risks and recommendation: A case study in the kurdistan region," International Journal of Innovation, Creativity and Change, vol. 13, pp. 349-365, 2020.

[51] L. M. Abdulrahman, S. R. Zeebaree, S. F. Kak, M. A. Sadeeq, A. AL-Zebari, B. W. Salim, et al., "A state of art for smart gateways issues and modification," Asian Journal of Research in Computer Science, vol. 7, pp. 1-13, 2021.

[52] F. Q. Kareem, S. Zeebaree, H. I. Dino, M. Sadeeq, Z. N. Rashid, D. A. Hasan, et al., "A survey of optical fiber communications: challenges and processing time influences," Asian Journal of Research in Computer Science, pp. 48-58, 2021.

[53] K. Jacksi, S. R. Zeebaree, and N. Dimililer, "LOD Explorer: Presenting the Web of Data," Int. J. Adv. Comput. Sci. Appl. IJACSA, vol. 9, pp. 1-7, 2018.

[54] A. Zeebaree, A. Adel, K. Jacksi, and A. Selamat, "Designing an ontology of E-learning system for duhok polytechnic university using protégé OWL tool," J Adv Res Dyn Control Syst Vol, vol. 11, pp. 24-37, 2019.

[55] H. M. Yasin, S. R. Zeebaree, and I. M. Zebari, "Arduino based automatic irrigation system: Monitoring and SMS controlling," in 2019 4th Scientific International Conference Najaf (SICN), 2019, pp. 109-114.

[56] S. A. Elavarasi, J. Akilandeswari, and K. Menaga, "A survey on semantic similarity measure," International Journal of Research in Advent Technology, vol. 2, pp. 389-398, 2014.

[57] A. AL-Zebari, S. Zeebaree, K. Jacksi, and A. Selamat, "ELMS–DPU ontology visualization with Protégé VOWL and Web VOWL," Journal of Advanced Research in Dynamic and Control Systems, vol. 11, pp. 478-85, 2019.

[58] D. A. Zebari, H. Haron, S. R. Zeebaree, and D. Q. Zeebaree, "Multi-level of DNA encryption technique based on DNA arithmetic and biological operations," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 312-317.

[59] S. Melasagare and V. Thombre, "Document Classification and Clustering using Feature Extraction for Similarity Measure," 2016.

[60] S. Zebari and N. O. Yaseen, "Effects of parallel processing implementation on balanced load-division depending on distributed memory systems," J. Univ. Anbar Pure Sci, vol. 5, pp. 50-56, 2011.

[61] K. Jacksi, N. Dimililer, and S. R. Zeebaree, "A survey of exploratory search systems based on LOD resources," 2015.

[62] M. P. Naik, H. B. Prajapati, and V. K. Dabhi, "A survey on semantic document clustering," in 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT), 2015, pp. 1-10.

[63] K. B. Obaid, S. Zeebaree, and O. M. Ahmed, "Deep learning models based on image classification: a review," International Journal of Science and Business, vol. 4, pp. 75-81, 2020.

[64] M. A. Omer, S. R. Zeebaree, M. A. Sadeeq, B. W. Salim, Z. N. Rashid, and L. M. Haji, "Efficiency of malware detection in android system: A survey," Asian Journal of Research in Computer Science, vol. 7, pp. 59-69, 2021.

[65] Z. S. Ageed, S. Zeebaree, M. Sadeeq, M. B. Abdulrazzaq, B. W. Salim, A. A. Salih, et al., "A state of art survey for intelligent energy monitoring systems," Asian Journal of Research in Computer Science, vol. 8, pp. 46-61, 2021.

[66] S. R. Zeebaree, A. B. Sallow, B. K. Hussan, and S. M. Ali, "Design and simulation of high-speed parallel/sequential simplified DES code breaking based on FPGA," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 76-81.

[67] A. S. Abdulraheem, S. Zeebaree, and A. M. Abdulazeez, "Design and implementation of electronic human resource management system for duhok polytechnic university," Technology Reports of Kansai University, vol. 62, pp. 1407-1420, 2020.

[68] G. M. O. Zebari, K. Faraj, and S. Zeebaree, "Hand writing code-php or wire shark ready application over tier architecture with windows servers operating systems or linux server operating systems," International Journal of Computer Sciences and Engineering, vol. 4, pp. 142-149, 2016.

[69] S. H. Ahmed and S. Zeebaree, "A survey on security and privacy challenges in smarthome based IoT," International Journal of Contemporary Architecture, vol. 8, pp. 489-510, 2021.

[70] H. H. A. Razaq, A. S. Gaser, M. A. Mohammed, E. T. Yassen, S. A. Mostafad, S. R. Zeebaree, et al., "Designing and implementing an arabic programming language for teaching pupils," Journal of Southwest Jiaotong University, vol. 54, 2019.

[71] Z. S. Ageed, S. R. Zeebaree, M. A. Sadeeq, R. K. Ibrahim, H. M. Shukur, and A. Alkhayyat, "Comprehensive Study of Moving from Grid and Cloud Computing Through Fog and Edge Computing towards Dew Computing," in 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA), 2021, pp. 68-74.

[72] Y. S. Jghef, M. J. M. Jasim, H. M. Ghanimi, A. D. Algarni, N. F. Soliman, W. El-Shafai, et al., "Bio-Inspired Dynamic Trust and Congestion-Aware Zone-Based Secured Internet of Drone Things (SIoDT)," Drones, vol. 6, p. 337, 2022.

[73] S. Chavhan, S. R. Zeebaree, A. Alkhayyat, and S. Kumar, "Design of Space Efficient Electric Vehicle Charging Infrastructure Integration Impact on Power Grid Network," Mathematics, vol. 10, p. 3450, 2022.

[74] N. T. Muhammed, S. R. Zeebaree, and Z. N. Rashid, "Distributed Cloud Computing and Mobile Cloud Computing: A Review," QALAAI ZANIST JOURNAL, vol. 7, pp. 1183-1201, 2022.

[75] A. S. Aljuboury, S. R. Zeebaree, F. Abedi, Z. S. Hashim, R. Q. Malik, I. K. Ibraheem, et al., "A New Nonlinear Controller Design for a TCP/AQM Network Based on Modified Active Disturbance Rejection Control," Complexity, vol. 2022, 2022.

[76] H. B. Abdalla, A. M. Ahmed, S. R. Zeebaree, A. Alkhayyat, and B. Ihnaini, "Rider weed deep residual network-based incremental model for text classification using multidimensional features and MapReduce," PeerJ Computer Science, vol. 8, p. e937, 2022.

[77] V. D. Majety, N. Sharmili, C. R. Pattanaik, E. L. Lydia, S. R. Zeebaree, S. N. Mahmood, et al., "Ensemble of Handcrafted and Deep Learning Model for Histopathological Image Classification," CMC-COMPUTERS MATERIALS & CONTINUA, vol. 73, pp. 4393-4406, 2022.

[78] L. M. Abdulrahman, S. R. Zeebaree, and N. Omar, "State of Art Survey for Designing and Implementing Regional Tourism Web based Systems," Academic Journal of Nawroz University, vol. 11, pp. 100-112, 2022.

[79] H. Abdullah and S. R. Zeebaree, "Android Mobile Applications Vulnerabilities and Prevention Methods: A Review," 2021 2nd Information Technology To Enhance e-learning and Other Application (IT-ELA), pp. 148-153, 2021.

[80] L. Stanchev, "Semantic document clustering using information from WordNet and DBPedia," in 2018 IEEE 12th International Conference on Semantic Computing (ICSC), 2018, pp. 100-107.

[81] S. Zeebaree, N. Cavus, and D. Zebari, "Digital Logic Circuits Reduction: A Binary Decision Diagram Based Approach," LAP LAMBERT Academic Publishing, 2016.

[82] A. Maind, A. Deorankar, and P. Chatur, "Measurement of semantic similarity between words: A survey," International Journal of Computer Science, Engineering and Information Technology, vol. 2, pp. 51-60, 2012.

[83] A. M. Abed, Z. N. Rashid, F. Abedi, S. R. Zeebaree, M. A. Sahib, A. J. a. Mohamad Jawad, et al., "Trajectory tracking of differential drive mobile robots using fractional-order proportional-integral-derivative controller design tuned by an enhanced fruit fly optimization," Measurement and Control, vol. 55, pp. 209-226, 2022.

[84] D. M. Abdulqader and S. R. Zeebaree, "Impact of Distributed-Memory Parallel Processing Approach on Performance Enhancing of Multicomputer-Multicore Systems: A Review," QALAAI ZANIST JOURNAL, vol. 6, pp. 1137-1140, 2021.

[85] L. M. Haji, S. R. Zeebaree, O. M. Ahmed, M. A. Sadeeq, H. M. Shukur, and A. Alkhavvat, "Performance Monitoring for Processes and Threads Execution-Controlling," in 2021 International Conference on

Communication & Information Technology (ICICT), 2021, pp. 161-166.

[86] Z. N. Rashid, S. Zeebaree, and A. Sengur, "Novel remote parallel processing code-breaker system via cloud computing," ed: TRKU, 2020.

[87] N. M. ABDULKAREEM and S. R. ZEEBAREE, "OPTIMIZATION OF LOAD BALANCING ALGORITHMS TO DEAL WITH DDOS ATTACKS USING WHALE OPTIMIZATION ALGORITHM," Journal of Duhok University, vol. 25, pp. 65-85, 2022.

[88] A. A. Yazdeen and S. R. Zeebaree, "Comprehensive Survey for Designing and Implementing Web-based Tourist Resorts and Places Management Systems," Academic Journal of Nawroz University, vol. 11, pp. 113-132, 2022.

[89] S. I. Ahmed, S. Y. Ameen, and S. R. Zeebaree, "5G Mobile Communication System Performance Improvement with Caching: A Review," in 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI), 2021, pp. 1-8.

[90] H. A. Mohammed, S. Zeebaree, V. M. Tiryaki, and M. M. Sadeeq, "Web-Based Land Registration Management System: Iraq/Duhok Case Study," Journal of Applied Science and Technology Trends, vol. 2, pp. 113-119, 2021.

[91] R. K. Ibrahim, S. R. Zeebaree, K. Jacksi, M. A. Sadeeq, H. M. Shukur, and A. Alkhayyat, "Clustering document based semantic similarity system using TFIDF and k-mean," in 2021 International Conference on Advanced Computer Applications (ACA), 2021, pp. 28-33.

[92] A. E. Mehyadin, S. R. Zeebaree, M. A. Sadeeq, H. M. Shukur, A. Alkhayyat, and K. H. Sharif, "State of Art Survey for Deep Learning Effects on Semantic Web Performance," in 2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM), 2021, pp. 93-99.

[93] A. B. Sallow, H. I. Dino, Z. S. Ageed, M. R. Mahmood, and M. B. Abdulrazaq, "Client/Server remote control administration system: design and implementation," Int. J. Multidiscip. Res. Publ, vol. 3, p. 7, 2020.

[94] M. A. Sadeeq and S. R. Zeebaree, "Design and analysis of intelligent energy management system based on multi-agent and distributed iot: Dpu case study," in 2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM), 2021, pp. 48-53.

[95] I. M. Ibrahim, S. R. Zeebaree, H. M. Yasin, M. A. Sadeeq, H. M. Shukur, and A. Alkhayyat, "Hybrid Client/Server Peer to Peer Multitier Video Streaming," in 2021 International Conference on Advanced Computer Applications (ACA), 2021, pp. 84-89.

[96] H. S. Malallah, R. Qashi, L. M. Abdulrahman, M. A. Omer, and A. A. Yazdeen, "Performance Analysis of Enterprise Cloud Computing: A Review," Journal of Applied Science and Technology Trends, vol. 4, pp. 01-12, 2023.

[97] A. F. Jahwar and S. Zeebaree, "A state of the art survey of machine learning algorithms for IoT security," Asian J. Res. Comput. Sci, pp. 12-34, 2021.

[98] T. M. G. Sami, Z. S. Ageed, Z. N. Rashid, and Y. S. Jghef, "Distributed, Cloud, and Fog Computing Motivations on Improving Security and Privacy of Internet of Things," Mathematical Statistician and Engineering Applications, vol. 71, pp. 7630-7660, 2022.

[99] H. A. Hussein, S. R. Zeebaree, M. A. Sadeeq, H. M. Shukur, A. Alkhayyat, and K. H. Sharif, "An investigation on neural spike sorting algorithms," in 2021 International Conference on Communication & Information Technology (ICICT), 2021, pp. 202-207.

[100] N. A. Kako, "DDLS: Distributed Deep Learning Systems: A Review," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, pp. 7395-7407, 2021.

[101] M. A. Omer, A. A. Yazdeen, H. S. Malallah, and L. M. Abdulrahman, "A Survey on Cloud Security: Concepts, Types, Limitations, and Challenges," Journal of Applied Science and Technology Trends, vol. 3, pp. 47-57, 2022.

[102] I. S. Abdulkhaleq and S. R. Zeebaree, "Science and Business," International Journal, vol. 5, pp. 126-136.

[103] Z. S. Hammed, S. Y. Ameen, and S. R. Zeebaree, "Investigation of 5G Wireless Communication with Dust and Sand Storms," Journal of Communications, vol. 18, 2023.

[104] F. Abedi, S. R. Zeebaree, Z. S. Ageed, H. M. Ghanimi, A. Alkhayyat, M. A. Sadeeq, et al., "Severity Based Light-Weight Encryption Model for Secure Medical Information System."

[105] D. M. Abdullah and S. R. Zeebaree, "Comprehensive survey of IoT based arduino applications in healthcare monitoring."

[106] A. A. Yazdeen, R. Qashi, H. S. Malallah, L. M. Abdulrahman, and M. A. Omer, "Internet of Things Impact on Web Technology and Enterprise Systems," Journal of Applied Science and Technology Trends, vol. 4, pp. 19-33, 2023.

[107] D. Mahapatra, C. Maharana, S. P. Panda, J. P. Mohanty, A. Talib, and A. Mangaraj, "A fuzzy-cluster based semantic information retrieval system," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 675-678.

[108] R.-G. Radu, I.-M. Rădulescu, C.-O. Truică, E.-S. Apostol, and M. Mocanu, "Clustering documents using the document to vector model for dimensionality reduction," in 2020 ieee international conference on automation, quality and testing, robotics (aqtr), 2020, pp. 1-6.

[109] S. Fatimi, C. E. Saili, and L. Alaoui, "A framework for semantic text clustering," International Journal of Advanced Computer Science and Applications, vol. 11, 2020.

[110] T. Zheng, Y. Gao, F. Wang, C. Fan, X. Fu, M. Li, et al., "Detection of medical text semantic similarity based on convolutional neural network," BMC medical informatics and decision making, vol. 19, pp. 1-11, 2019.

[111] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, "The k-means clustering algorithm with semantic similarity to estimate the cost of hospitalization," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 13, pp. 313-322, 2019.

[112] W. Lwin, "Impressive approach for documents clustering using semantics relations in feature extraction," in Proceedings of the 2019 the 9th International Workshop on Computer Science and Engineering, WCSE, Changsha, China, 2019, pp. 18-20.

[113] J. Park, C. Park, J. Kim, M. Cho, and S. Park, "ADC: Advanced document clustering using contextualized representations," Expert Systems with Applications, vol. 137, pp. 157-166, 2019.

[114] M. Wrzalik and D. Krechel, "Balanced word clusters for interpretable document representation," in 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2019, pp. 103-109.

[115] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," Scientometrics, vol. 111, pp. 1017-1031, 2017.

[116] S. R. Kolhe and S. Sawarkar, "A concept driven document clustering using WordNet," in 2017 International Conference on Nascent Technologies in Engineering (ICNTE), 2017, pp. 1-5.

[117] K. Sumathy, "A hybrid approach for measuring semantic similarity between documents and its application in mining the knowledge repositories," International Journal of Advanced Computer Science and Applications, vol. 7, 2016.

[118] M. Rafi, M. S. Shaikh, and A. Farooq, "Document clustering based on topic maps," arXiv preprint arXiv:1112.6219, 2011.

[119] D. S. Al-Azzawy and F. M. L. Al-Rufaye, "Arabic words clustering by using K-means algorithm," in 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), 2017, pp. 263-267.

[120] P. Zandieh and E. Shakibapoor, "Clustering data text based on semantic," International Journal of Computer (IJC), vol. 26, pp. 195-202, 2017.

[121] P. Nanayakkara and S. Ranathunga, "Clustering sinhala news articles using corpus-based similarity measures," in 2018 Moratuwa Engineering Research Conference (MERCon), 2018, pp. 437-442.

[122] D. SHABANA AFREEN, "SEMANTIC BASED DOCUMENT CLUSTERING USING LEXICAL CHAINS," 2016.

[123] I. Ali and A. Melton, "Semantic-based text document clustering using cognitive semantic learning and graph theory," in 2018 IEEE 12th

International Conference on Semantic Computing (ICSC), 2018, pp. 243-247.

[124] S. S. Desai and J. Laxminarayana, "WordNet and Semantic similarity based approach for document clustering," in 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2016, pp. 312-317.

[125] Q. Bai and C. Jin, "Text Clustering Algorithm Based on Semantic Graph Structure," in 2016 9th International Symposium on Computational Intelligence and Design (ISCID), 2016, pp. 312-316.

[126] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 61-66.

[127] S. Kumar and K. K. Bhatia, "Semantic similarity and text summarization based novelty detection," SN Applied Sciences, vol. 2, pp. 1-15, 2020.

[128] P. Banik, S. Gaikwad, A. Awate, S. Shaikh, P. Gunjgur, and P. Padiya, "Semantic analysis of wikipedia documents using ontology," in 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), 2018, pp. 1-6.

[129] I. Blokh and V. Alexandrov, "News clustering based on similarity analysis," Procedia computer science, vol. 122, pp. 715-719, 2017.

[130] A. Hotho, A. Maedche, and S. Staab, "Ontology-based text document clustering," KI, vol. 16, pp. 48-54, 2002.

[131] R. P. Honeck, "Semantic similarity between sentences," Journal of Psycholinguistic Research, vol. 2, pp. 137-151, 1973.

[132] J. Duan, Y. Wu, M. Wu, and H. Wang, "Measuring semantic similarity between words based on multiple relational information," IEICE TRANSACTIONS on Information and Systems, vol. 103, pp. 163-169, 2020.

[133] J.-B. Gao, B.-W. Zhang, and X.-H. Chen, "A WordNet-based semantic similarity measurement combining edge-counting and information content theory," Engineering Applications of Artificial Intelligence, vol. 39, pp. 80-88, 2015.

[134] P. Kherwa and P. Bansal, "Latent semantic analysis: an approach to understand semantic of text," in 2017 international conference on current trends in computer, electrical, electronics and communication (CTCEEC), 2017, pp. 870-874.

[135] N. M. Nejad, S. Scerri, and S. Auer, "Semantic similarity based clustering of license excerpts for improved end-user interpretation," in Proceedings of the 13th International Conference on Semantic Systems, 2017, pp. 144-151.

[136] J. Zeng, J. Ge, Y. Zhou, Y. Feng, C. Li, Z. Li, et al., "Statutes recommendation based on text similarity," in 2017 14th Web Information Systems and Applications Conference (WISA), 2017, pp. 201-204.

[137] H. Nie, J. Zhou, Q. Guo, and Z. Huang, "Improved semantic similarity method based on HowNet for text clustering," in 2018 5th International Conference on Information Science and Control Engineering (ICISCE), 2018, pp. 266-269.

[138] B. Berendt, A. Hotho, and G. Stumme, "Towards semantic web mining," in The Semantic Web—ISWC 2002: First International Semantic Web Conference Sardinia, Italy, June 9–12, 2002 Proceedings 1, 2002, pp. 264-278.

[139] K. Yang, H. Zhang, Z. Chu, and L. Sun, "A Text Topic Mining Algorithm Based on Spatial Propagation Similarity Metric," in 2019 Chinese Control And Decision Conference (CCDC), 2019, pp. 4339-4344.