Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon - College of Information Technology
Software Department

INFORMATION DIFFUSION DETECTION USING SEMANTIC SIMILARITY BASED ON ONTOLOGY AND INFLUENTIAL USERS

A Dissertation
Submitted to the Council of the College of Information Technology, University of Babylon in Partial Fulfillment of the Requirements for the Doctor of Philosophy Degree in Information Technology / Software

By
Yasir Abd Alhamed Najem Mansor

Supervised by
Professor Dr. Asaad Sabah Hadi Abass

2022 A.D. **1443 A.H.**

Supervisor Certification

I certify that the dissertation entitled (**INFORMATION DIFFUSION DETECTION USING SEMANTIC SIMILARITY BASED ON ONTOLOGY AND INFLUENTIAL USERS)** was prepared under my supervision at the department of Software / College of Information Technology/the University of Babylon as partial fulfillment of the requirements of the degree of Ph.D. in Information Technology-Software.

Signature:

Supervisor Name Prof Dr. Asaad Sabah Hadi

Date: / 5 /2022

The Head of the Department Certification

In view of the available recommendations, I forward the dissertation entitled "**Ontology Modeling and Influential User Analysis on Twitter to Detect Information Diffusion**" for debate by the examination committee.

Signature:

Name: Assistant Prof. Dr. Ahmed Saleem Abbas

Head of Software Department

Date: / 5 /2022

Certification of the Examination Committee

We, the undersigned, certify that (Yasir abd alhamed najem) candidate for the degree of Doctor of Philosophy in Information Technology - Software, has presented his dissertation of the following title (**INFORMATION DIFFUSION DETECTION USING SEMANTIC SIMILARITY BASED ON ONTOLOGY AND INFLUENTIAL USERS**) as it appears on the title page and front cover of the dissertation that the said dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: (/5 /2022).

Signature:
Name: Dr. Israa Hadi Ali
Title: Professor
Date: / 5 / 2022
(**Chairman**)

Signature:
Name: Dr. Ahmed Hussein Aliwy
Title: Professor
Date: / 5 / 2022
(**Member**)

Signature:
Name: Dr. Mohammed Abdullah Naser
Title: Professor
Date: / 5 / 2022
(**Member**)

Signature:
Name: Dr. Emad I Abdul Kareem
Title: Associate Prof.
Date: / 5 / 2022
(**Member**)

Signature:
Name: Dr**.** Dhiah Eadan Jabor
Title: Associate Prof.
Date: / 5 / 2022
(**Member**)

Signature:
Name: Dr. Asaad Sabah Hadi
Title: Professor
Date: / 5 / 2022
(**Member and Supervisor**)

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:
Name: Dr. Hussein Atiyah Lafta
Title: Professor
Date: / 5 / 2022
(**Dean of Collage of Information Technology**)

Dedication

I hereby declare that this Dissertation, submitted to University of Babylon in partial fulfillment of requirements for the degree of Doctorate of Philosophy in Information Technology-Software has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose sources are appropriately cited in the references.

Signature:

Date: / 5 / 2022

Yasir Abd Alhamed Najem

Acknowledgement

First and forever, all my praise is to my God for His graces that enable me to continue the requirements of my study and overcome the difficulties, which stood in my way during the courses and research.

My thanks are due to the ministry of Iraqi defense and Babylon University for giving me the opportunity for postgraduate studies.

I would like to express my deepest gratitude and appreciation to my supervisor **Prof Dr. Asaad Sabah Hadi ,** for his valuable advices, guidance, and cooperation during this work despite his valuable time and his many works.

Finally, thanks are due to **my family**, **my teachers, my friends**, and to **anyone** who helped me in one way or another**.**

Yasir Abd Alhamed Najem

Abstract

The proliferation of social network media such as Twitter has accelerated the process of sharing information and expressing opinions about global health crises and important events. Twitter is used by many government officials around the world as one of the main communication channels to share political events and news related to Covid-19 on a regular basis to the general public.

Due to the use of different terms for expressing the same topic in a Twitter post, it becomes difficult to build applications such as finding the rate of diffusion-specific news on Twitter, inquiry answering regarding Covid-19 Twitter posts, retrieving certain information, and automatic translation by following previous traditional text-matching techniques. In order to solve this problem, it requires providing a knowledge source that collects the terms that reflect a single meaning in a formal concept with their characteristics and relationships, such as ontology, and a technique to compute the semantic similarity between these concepts.

This dissertation presents a way to discover the spread of information on Twitter based on semantic meaning rather than word-to-sentence matching. Using a semantic similarity technique based on ontology and influencers. Because of the lack of an ontology in this field, a source of knowledge (ontology) was built that explains the concepts used in this field in a semantic way, it included 1203 concepts. It is the most used concepts in Twitter tweets with the spread of the Corona epidemic.

The proposed system was evaluated through a data set evaluated by humans, where the results showed an accuracy of 89.1% compared with the human evaluation, as well as the results were better than the result of

Atish (73.9%), which is one of the latest research in this field using the same data.

Credits or Publication

Some of the works presented in this thesis have been publish or accepted as listed below.

1. Entropy based Weighted Features for Detecting the Influential Users on Twitter.

   Published in: Date Added to IEEE Xplore: 26 April 2021, ISBN Information: INSPEC Accession Number: 20654040, DOI: 10.1109/AiCIS51645.2020.00012

2. Semi-automatic ontology learning for twitter messages based on semantic feature extraction.

   Published in: Springer Nature Switzerland AG 2021 ,A. M. Al-Bakry et al. (Eds.): NTICT 2021, CCIS 1511, pp. 1–14, 2021. https://doi.org/10.1007/978-3-030-93417-0_1

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
ATR	Automatic Term Recognition
CRITIC	CRiteria Importance Through Inter-criteria Correlation
CWWE	Custom Weighted Word Embedding
DL	deep learning
EA	Entropy Method
EWM	Entropy Weight Method
F-Logic	Frame Logic
HITS	Hyperlink-Induced Topic Search
IC	Information Content
IDF	Inverse Document Frequency
INFOcas	information cascades on Twitter about the virus
KIF3	Knowledge Interchange Format
KR	Knowledge Representation
LCS	Least Common Subsumer
LExO	Expressive Learning Ontology
LIWC	Linguistic Inquiry and Word Count
LSTM	Long Short-Term Memory
MCDM	Multi Criteria Decision Making
MW	Mean Weight
NLP	Natural Language Processing
OSN	Online Social Network
OWL	Web Ontology Language
POS	Part Of Speech

RDF	Resource Description Framework
SI	Susceptible Infected
SIR	susceptible (S), infected (I), recovered (R)
SIRemp	SIR model from empirically-validated cases
SIRS	Susceptible Infected Removed Susceptible
SIRsim	susceptible (S), infected (I), recovered (R),simulater(sim)
SIS	Susceptible Infected Susceptible
SPARQL	SPARQL Protocol and RDF Query Language
TF	Term Frequency
TM	Text Mining
UFS	University of the Free State
UGC	User-Generated Content
W3C	World Wide Web Consortium
WSD	Word Sense Disambiguation
XML	Extensible Markup Language

CHAPTER ONE

General Introduction

CHAPTER ONE
GENERAL INTRODUCTION

1.1 Introduction

The massive use of social media like Twitter has accelerated the exchange and spread of information, opinions, and events. The information exchanged has several formats such as reviews, messages, videos, and photos. Especially opinions and events about public crises such as Covid-19. Since the beginning of the epidemic at the end of 2019, the number of people infected with the epidemic is 131 million people and 2.84 million deaths in all countries of the world [1].

A quarantine has been imposed in most countries. The world has become heavily dependent on platforms of social media like Twitter to receive news and express their views. The Twitter platform has a large and wide role in spreading and updating news in real-time. Eysenbach and Chew (2010) [2] has shown that Twitter can be used in real-time "science of information" studies. It became a source for posting guidance from authorities of health in response to public concerns. Twitter is used by many government officials around the world as one of the main communication channels to participate in political events and news related to Covid-19 on a regular basis for the general public [3].

Twitter messages, compared to traditional messages, pose a new challenge in the field of discovering the spread of information. Because Twitter messages are written by anyone, restricted in length, contain

expressions in different languages, and many spelling and grammatical errors, which will affect negatively the algorithms for discovering the spread of information on the Twitter platform.

The first, researcher applied classical text mining theories to discover the spread of information on Twitter. In addition to taking advantage of some of the unique features of Twitter such as hashtags. Where there has been a growing interest in what is called 'Twitter-mining' [4].

After that, the researcher has gone to study the spread of information on Twitter by relying on influential users and opinion leaders in social network media, by analyzing the social links that users can generated by using the allowed actions.

Due to the different terms that are used to express the same topic in a Twitter post, it becomes difficult to build applications to find specific information in Twitter messages using the above methods without providing a source of knowledge that unifies these terms into a single concept such as ontology and a semantic technique to find the similarity between the posts. Ontology construction from texts for a specific domain is a process that involves analyzing those texts and extracting terminology and concepts related to the domain and their relationships. After that, represent the ontology through ontology languages. Where the knowledge is represented as concepts and relations that connect those concepts to be in a machine-understandable form [5] [6].

1.2 Related Work

Many works have been introduce proposing different techniques in identifying the Information spread on Twitter as shown in table (1.1).

Amartya Hatua et al. (2017) [7] described an information spread model on Twitter, the model deals with Twitter data as three dimensions, these dimensions consist of the tweets volume, the tweets sentiment, and the tweets influence. To prophesy parameters for each dimension is used the nonlinear series of time method of Long Short-Term Memory (LSTM), the linear series of time method of Auto-regressive Integrated Average Moving (ARIMA), and used the Recurrent Neural Networks to compare the performance.

Raad Bin et al. (2018) [8] have employed the Hyperlink-Induced Topic Search (HITS) Algorithm and Friends network to find the information diffusion scale and pattern, Through the analysis of the network of friends, information is provided about the influential individuals in the network, by employing the number of retweets and the time between the tweet and the retweet. The method has evaluation based on how particular topic matchs one of the patterns and how to distinguish friends compared to other users.

Jinwen Xu et al. (2020) [9] have conducted a study applying text mining and spatial analyses methods to Twitter data during the Diego of a Winter Storm in December 2018. Many of the previous studies have focused on original tweets, and the study has used retweets to model information dissemination and analyze the geographic information flow distribution on different topics.

Sanjay Kumar et al. (2020) [10] have developed a model for the spread of data in social networks based on a method inspired by nature, which is to simulate forest fires, where a spark of fire can lead to a massive fire. The spread of fires in forests depends on vegetation, weather conditions, and terrain that serve as fuel. In the same way, the current users are a tree and information is the fire, and thus the spread of information across the network depends on the relationships between users and followers, the importance of the topic, and other features.

Ly Dinh et al. (2020) [11] have suggested a model using network analysis and spread modeling to discover the similar between the SIR (SIRsim) model, the SIR that depends on present COVID-19 cases (SIRemp), and information available on Twitter about the epidemic (INFOcas).

Amit K. et al. (2020) [12] have presented a theory that the spread of specific content on Twitter is driven by a series of nouns, adverbs, and adjectives that make up the sentence. Therefore, they suggested Custom Weighted Word Embedding (CWWE) to study the degree of spread of the content, as the method includes first extracting the words and creating a matrix of these words using the sequences in the text of the tweet, then multiplying the basis the weights has been assigned to the present index, where is given a higher weights if the influential class of distinct words such as nouns and adjectives, then they try to predict the possibility of spreading information using the deep neural network architecture for long-term memory, which in turn is improved on accuracy and training

execution time through the architecture of the convolutional neural network.

Ankit K et al. (2020) [13] have proposed a model that implements high-level soft analysis via homophile network-based community discovery techniques. This followed by the application of different diffusion models to study the rate of information diffusion in the ego network.

Junaid A. et al. (2021) [14] have proposed an approach that combines Twitter content features, user features, and LIWC linguistic features with machine learning algorithms to analyze Twitter's information diffusion metric.

Holger H. et al. (2021) [15] have proposed a model based on three components that influence the information dissemination of a tweet: the content, features, and sentiment (positive or negative choice of words) of the tweet, k-means of grouping and manual coding have been used to categorize tweets by subject, sentiment, length and number of emoji’s, images, videos, and links.

Table 1.1 Related Work System

Reference	Data Type	Technique	parameters	Output	Evaluation methods
Amartya H. et al (2017) [7]	Twitter data	Non-linear time series model of LSTM	tweets volume, the tweets sentiment and the tweets influence	information spread on twitter	cluster validity indices (CVIs)
Raad B. et al. (2018) [8]	Twitter data	HITS Algorithm and Friends network	number of retweets , the time between the tweet	information diffusion scale and pattern	How certain a topic matches one of the patterns and compared friends to other users
Jinwen X. et al. (2020) [9]	Twitter data	Applies text mining and spatial analysis methods	Retweets, analyzed the geographical distribution of information flows on various topics	spread Twitter data during Winter Storm Diego in 2018 December	comparison with previous works that based on retweet only
Sanjay K.et al. (2020) [10]	Social data	Models of simulate forest fires	Relationships between users and followers, and the importance of the topic and other features.	spread of data in social networks	comparison with Independent Cascade (IC) and Susceptible-Infected-Recovered (SIR) Model
Ly D. et al. (2020) [11]	COVID-19 tweets	Network analysis and spread modeling	SIRsim model , SIRemp model, INFOcas model	discover the similarities SIRsim and SIR model and INFOcas for spread	comparison with SIRsim model , SIRemp model, and INFOcas model

				COVID-19 information	
Amit K. et al. (2020) [12]	Twitter data	Deep neural network long-term memory, and improve the result by architecture of the convolutional neural network.	Custom Weighted Word Embedding	spread of specific content on Twitter	The results are compared to a pre-trained word embedding
Ankit K et al. (2020) [13]	Twitter data	High-level soft analysis via homophile network-based community discovery techniques and different diffusion models	community discovery techniques	information diffusion in the ego network	comparison with previous works with the respect to time
Junaid A et al. (2021) [14]	Twitter data	Machine learning algorithms	Twitter content features, user features, and LIWC linguistic features	metric the Twitter's information diffusion	achieved 77% accuracy with SVM while classifying the information diffusion categories
Holger H et al. (2021) [15]	Twitter data	K-means of grouping and manual coding	the content, features, and sentiment (positive or negative choice of words) of the tweet	information spread on twitter	The system was evaluated using pre-coding data

1.3 Motivations

Using different terms to express the same topic in Twitter posts, led to difficulty and lack of accuracy in discovering the spread of specific information depending on word-match. Was a motivation to use a Semantic web similarity technique based on ontology to find the spread depending on the meaning of the tweet instead of word matching.

1.4 Problem Statement

Exposing the spread of specific information on Twitter is a difficult and important process. Previously, it has used a framework for text mining, which relies on methods based on matching keywords or influencer user technology only. These methods are inaccurate in the case of searching for certain news. Each Twitter user writes in different styles and terms to express the same topic. Therefore, the main problem of this research is finding the spread of specific information within Twitter tweets by relying on the semantic meaning of the tweet instead of matching words.

1.5 Main Goal and Objectives

The main goal of the dissertation is to develop the semantic web similarity comparative model that used to find the spread of specific information within Twitter tweets, through the objectives are follows:

1. Finding tweets data collector, to collect the data from kaggle repository.
2. Finding COVID-19 ontology constructor, by using layer cake concepts from tweets, and enriching the ontology from other medical ontologies.
3. Finding a user's influencer detector that used to extract the influential users in the network based on user activity, attention obtained and social connectivity.
4. Finding the influencers tweets collector, which collect the influencer's tweets from dataset.

1.6 Contribution

The main contributions of the proposed system are:

1. Build a COVID-19 ontology constructor that generates ontology specified in a Covid-19 tweets domain.
2. Build users influencer detector by using eight features that indicate their ability to disseminate and amplify information. Using the messages of these users only will reduce the comparison process and increase the speed of discovery of the spread of information.

3. Build semantic web similarity comparative model. This model using to finding information diffusion based on ontology through finding the similarity between the target message and the set of tweets by combining two approaches: Edge-counting measure approaches and Feature-based measures approaches.
4. Increasing the speed of similarity discovery by simplifying hierarchical relationships in ontology, by using the technique of hierarchical tree summarization in building the ontology.

1.7 Challenges

Generally, there are wide spectrum of challenges related to information diffusion detection in Twitter such as:

1. The most important challenges that is faced in discovering the spread of information on Twitter involve, unstructured data, processing large amounts of data, collecting data, issues of privacy, untrue sentence structures, and vernacular. In addition, their writings may contain many grammatical or misspellings, abbreviated and informal terms. Therefore, discovering the spread of information in social networks compared to data that has good structured and edited is a difficult task.
2. The length of user-generated content (UGC) on SNs is limited (eg, only 280 characters for a Tweet). Therefore, ideas are briefly represented by an insufficient amount of information, which generates extra challenges.

3. Extracting the relations of concepts from diverse sources is challenge, where semantic lexical and other medical ontologies domain are used.
4. The process of expanding ontology learning to accommodate larger sets of domain concepts is a major challenge, thus , some techniques and tools used to build and enrich the ontology from the corpus and other ontologies.
5. The quality of ontology is affected by human intervention, as it is inversely proportional to human intervention. Therefore, it is preferable to conduct knowledge acquisition in an automatic or semi-automatic manner, which is a challenge and a great burden for ontology acquisition.

1.8 Scope of Dissertation

The study focuses on finding the spread of specific information within the Twitter tweets about the COVID-19 epidemic in the English language. By building specialized techniques in the field of semantic similarity based on ontology. The data used contain tweets that are set From March 29 to April 30, 2020.

1.9 Dissertation layout

Chapter 1: Introduction - Overview of the research, Motivations, problem, objectives, Contribution, Challenges, the methodology followed in the research, and Previous related work that has also been discussed on semantic web technologies and mining techniques to find information diffusion in the social network.

Chapter 2: Theoretical Background- introduces the general introduction about relevant subjects surrounding this work. This chapter begins with a discussion about the information diffusion concept and its models, followed by a description of Online Social Networks, Semantic web similarity, Ontology learning models, Word Sense Disambiguation techniques , Lastly, presents details about Data mining and the Entropy weight method.

Chapter 3: Presents The Architecture Of The Proposed System- ontology-based detection information diffusion on Social Network, which describes the process of discovering the spread of certain information on the Twitter network by utilizing the semantic web similarity which is based on the knowledge source (build COVID-19 Ontology Domain) and influential users. The ontology is developed using layer cake model, Protégé and OWL, find the influential users based in

multi-feature, also, build a hybrid theory in Semantic Web Similarity to find the rate of spread information.

Chapter 4: Experimental Results and System Evaluation- presents a set of diagnosing to test and evaluate the system with a discussion of the results.

Chapter 5: Conclusions and Recommendations for Future Work - gives a list of conclusions drawn from the analysis of empirical results. In addition, in this chapter, some suggestions made to clarify the directions of research that the work may be taken in the future.

CHAPTER TWO

Information Diffusion Detection Based on Semantic Similarity

CHAPTER TWO
INFORMATION DIFFUSION DETECTION BASED ON SEMANTIC SIMILARITY

2.1 Introduction

The massive use of social networks have accelerated the exchange and spread of information, opinions and events about public crises such as Covid-19, as the world has become heavily dependent on platforms of social media like Twitter to receive news and express their views. To discover the spread of certain information on this platform, and due to use different terms for expressing the same topic in Twitter posts, it became difficult to discover this information by following previous mining methods that depend on matching between words or sentences[2].

In order to solve this problem, this study presents a new approach to detect spread specific information on Twitter, depending on knowledge sources such as ontology, influential user detection techniques, and semantic web similarity techniques.

2.2 Information Diffusion

The information diffusion process: is the transmission of information from one person to another or from one person to a community in the network. Also known as information propagation. These users can be consider as nodes in the social network, while the interactions between these users are represented as relationships that are represented by the

edges between two nodes in these networks. A real group of user can be represented as a community in the social network, and specific information can be published by these nodes within the community [16].

Many studies are looking at the spread of information to its significant importance in several applications such as rumor controlling [17][18][19] behavior analysis[20], Measuring public opinion, studying psychological phenomenon[21], and allocating resources in public health care systems[22]. The spread models fall into two classes of explanatory models and predictive models. Predictive models are models that train the system to predict the spread of information and then use it to find the spread, such as the IC model, the LT model, and the Game Theory model. While the explanatory models aim to examine the information diffusion process and elucidate the factors that affect it in an attempt to explain this phenomenon, it includes epidemics models and influence models.

2.2.1 Epidemics Models

The process of spreading information can be represent as the process of the spread of an epidemic. In the epidemic spread, there are users infected with pathogens and users exposed to pathogens. The virus can spread from an infected person to susceptible persons, and information can be spread from communicator persons to recipients in a similar way[23]. The basic models are the SI model (Susceptible Infected)[24], the SIS model (Susceptible Infected Susceptible)[25][26], SIR model (Susceptible Infected Removed)[27], and SIRS model (Susceptible Infected Removed Susceptible)[28].

2.2.2 Influence Models

One of the basics of social networks is the use of Influence analysis to find the spread of information in these networks[29]. Influence analysis is divided into three types: influence of individual, the influence of community, and influence maximization. Where we have used individual influence as a part of proposed method to detect the information diffusion in Twitter [30].

2.2.2.1 Individual Influence

The individual effect refers to researches that are relate to opinion leaders. Opinion leaders are the nodes that can act as a bridge to disseminate information. They have an impact on other individuals of a social network[31]. The influence of opinion leaders in information dissemination research cannot be neglected, research on the influence of opinion leaders includes ways based on the network architecture, alternate information and user features [32].

Previously, has been one indicator or a simple evaluation method the focus of user impact analyses and research, for example, Leavitt [33] has proposed a measured user influence by the number of followers on Twitter. Cha et al. [34]analyzed that whether a high follower count or a retweet would definitely have a significant impact on the user. With the development of research, it has been found that it is not possible to find influential users using one attribute or a simple method. The methods of finding influential users can be split into the following classes:

1) Topological Based Methods: In these ways often utilize topological attributes such as nodes connection relationships and edge weighting, to assess user influence in online social networks[35]. Where the centrality, betweenness and closeness methods are used to measure user impact, also the position of the node can be used to find influential users in social networks[36]. Furthermore, in many researches, the performance has been improved by combined between centrality and other topological attributes to evaluate user impact[37].

2) Topics Based Methods: There are many researches that have dealt with a study evaluating the influence of an individual on a specific topic[38], and the simplest way to find influence is through a user ranking based on the quality of their tweets within a specific topic, and some ways, evaluated the user impact out of using of hashtags[39].Another kind of topic-sensitive method is TwitterRank, which uses the topical similarity between users and a modified PageRank algorithm to evaluate the impact of users[40].

3) Interaction And User Features Based Methods: User interactions are considered as one of the most important functions of all social networks, where many studies are based on user interaction behavior such as retweeting, mention and comment to find the impact of users. Often the simple ways use retweet behavior and PageRank to ranking users impact in OSN. Some methods have evaluated user influence by using URL click behavior[41]. More complex methods involved combining user interaction behaviors with other features such as

network topology to find user impact and yield very good results[42][43].

2.3 Online Social Networks (OSN)

A social network is a group of individuals, organizations or other categories connected by common interests[44]. OSN such as Facebook, Twitter and the other OSNs are essential trends in person's connectivity and correlation habits, where individuals can introduce themselves, establish connections with other accounts and react with them [45].

2.3.1 Twitter

Twitter is a free social networking microblogging service that allows the propagation of short messages called tweets for members. The registered user on Twitter is able to propagate tweets and follow Tweets by other users by using various devices and platforms such as sending by cell phone as a text message, computer client software or by posting at the Twitter.com website[46].

Twitter has default public settings different from LinkedIn or Facebook wherein Facebook and LinkedIn the users need to agree on social connections, while in public twitter any user can follow any other user. Users can add hashtags to a keyword in their tweets to connect tweets to a public topic. The hashtag which acts as a Meta tag is expressed as the #keyword.

Originally Tweets were limited to 140 char and redoubled to 280 in Nov. 2017. but video and audio posts are still limited to 140 sec. Users

react on Twitter by following people who do tweets that are deemed interesting, so Users can post information to their followers by tweeting or retweeting other users' posts. Furthermore, by including the username of the user in the tweet, the users can mention other users. A number of followers for Users is an index of popularity for users and is considered a metric of influence, called followers influence. Instead, the Retweet count measures the user's ability to amplify information, which spreads to other users and is named the Retweet effect. The influence of user mention is the amount of the mention with the name of the user which acts as the name value of the user and metric the ability of that user to engage other users in a subject discussion[47].

2.3.2 Twitter Features Measures

1. The ratio of Follower to Following (Rf): compares between the number of individuals following User (Ff) to the number of individuals that user following them (Fg). If the score is smaller than 1, is it possible that User is a group follower of other users for the purpose of acquiring more people himself. Otherwise, the higher this value, the higher people's interest in User posts, without User needing to reciprocate their interest.

$$Rf = Ff / Fg \qquad (2.1)$$

2. Ratio of Retweet and Reply to user’s tweets (Rrm): It detects the number of user's Tweets from out of his total Tweets that involve other users' reaction. Which is the portion of User's Tweets that

have been amplified by another user or that generate interest in them from the all Tweets posted by user, and the following metric is defined in[48] where Nr is the number of retweets that user obtain, Nm is the number of Reply that user obtain, Nt is the number of user tweets:

$$\mathbf{Rrm=(Nr + Nm) / Nt} \qquad \mathbf{(2.2)}$$

3. The sum of the number of retweets and reply acquired by the user (Urmo). Where Nr is the number of retweets acquired by the user, Nm is the number of reply acquired by the user:

$$\mathbf{Urmo= Nr + Nm} \qquad \mathbf{(2.3)}$$

4. Followers number (Ff): the entire number of followers for the user. In General, user influence increases as the number of followers' increases. We believe that the followers' number is a significant feature of the user's impact.
5. A number of tweets (Nt): the entire number of tweets the user has posted. This measure indicates the productivity of the User. The research by Keeler and Berry [49]suggests five characteristics of influencers, 'activist' is one of which. A user's post to a big number of Tweets signal to a height level of user participation in the communities.
6. Number of total like acquired by the user (Nu_like): Where Nu_like is the Number of like acquired by the user from other users.
7. New Reply and retweet (NeRepl_Rt): the number of new replies and retweet to the user over a period of time, where the reply represents the nominal value of an individual.

8. New Tweets (NeT): The numbers of Tweets a user recently create during a specific period of time. We recognize that the effect is time sensitive because a user's impact in a social network changes during time. If the influencer hasn't posted any Tweets for a while, then their influence will likely start to wane.

2.4 Semantic Web

The Semantic Web is a standard of the WWW set by the World Wide Web Consortium (W3C). The target of the Semantic Web is to make the data of Internet machine-readable[50]. The term “semantic” means the sense or understanding. The essential difference between the technologies of Semantic Web and the other technologies regarding to data (like relational databases) is that the Semantic Web is interested with the meaning of data and not the structure of its[51]: The Semantic web mainly consisting three standards techniques:

[1] Resource Description Framework (RDF): semantic web data model, all Semantic Web information is represented and stored in the RDF model[52].

[2] SPARQL Protocol and RDF Query Language (SPARQL): the Semantic Web query language, it's designed specifically to query the semantic web data such as ontology[53].

[3] Web Ontology Language (OWL): Knowledge representation (KR) language for the semantic web [52].

OWL allow you to create concepts that can be reuse as much and as often as possible. creation means that each concept is defined carefully so

that allow to be selected and assembled in various combinations with other concepts as needed for various applications and purposes.

The way to distinguish between the semantic web applications and the rest of the applications is to use the three techniques mentioned above. However, the Semantic Web is called by many expressions, such as Web 3.0 or the Linked Data Web. Where these techniques are used to represent the data of the Semantic Web, such as ontology that describes concepts and their relationships between entities and categories of things. These semantics provide many advantages such as logical reasoning on data and working with heterogeneous data[54].

2.4.1 Semantic Similarity

It is a method based on calculating the semantic similarity between two terms through information extracted from one or more knowledge sources such as ontology or lexical databases and lexical dictionaries. Where the knowledge basis of this source provides an organized representation of terms or concepts and their relationships in a semantic manner, which provides a semantic scale free of ambiguity[55]. By taking the actual meaning of the term given in sentence.

1. Types of knowledge-based semantic similarity methods:

Depending on the basic principle of measuring semantic similarity between words, semantic similarity methods can be categorized into four types[56]:

[1] Edge-counting methods.

[2] The feature-based methods.

[3] Methods based on information content

[4] Hybrid metrics method, as shown in table (2.1).

Table 2.1 Types of Semantic Similarity Methods Based on Knowledge

Method	**Principle**	**Measure**	**Feature**	**Advantage**
Path Based	*Length of the path linking different word senses*	*Shortest Path*	*Number of edges between the concepts*	*Simple measure*
		Wu & Palmer	*Path length augmented by subsumer path to root*	*Simple measure*
		L & C	*Number of edges between the concepts*	*Simple measure*
Information Content Based	*The concepts Sharing common information are similar*	*Resnik*	*Information content of the lowest common subsumer*	*Simple measure*
		Lin	*Information content of the lowest common subsumer and compared concepts*	*Considers the information content of compared concepts*
Feature Based	*The concepts having common features are similar*	*Tversky*	*Compares features of the concepts*	*Considers features while computing similarity*
Hybrid	*All of above*	*Zhou et al*	*All of above*	*All of above*

2.4.1.1 Edge-Counting Methods

The most obvious way to calculate edges is to look at ontology as a graph, that links words taxonomically and to calculate edges between two terms to measure their similarity. The greater the distance between the terms, the less similar they are. Rada et al. [57] proposed a measure called path, that the similarity is inversely proportional to the shortest path length between two terms. Wu and Palmer [58]suggest WuP measure, since the depth of words in the ontology is an important feature. The wup measure counts the number of edges between each term and their Least Common Subsumer (LCS). LCS is the common ancestor shared by both terms in the given ontology.

Where the two terms are indicate as 't1', 't2', their LCS indicated as tlcs, which are explain in figure 2.1, and the shortest path length between them indicated as wup is measured as:

$$Sim_{wup}(t_{1,}t_2) = \frac{2depth\ (t_{lcs})}{depth\ (t_1) + depth\ (t_2)} \qquad (2.4)$$

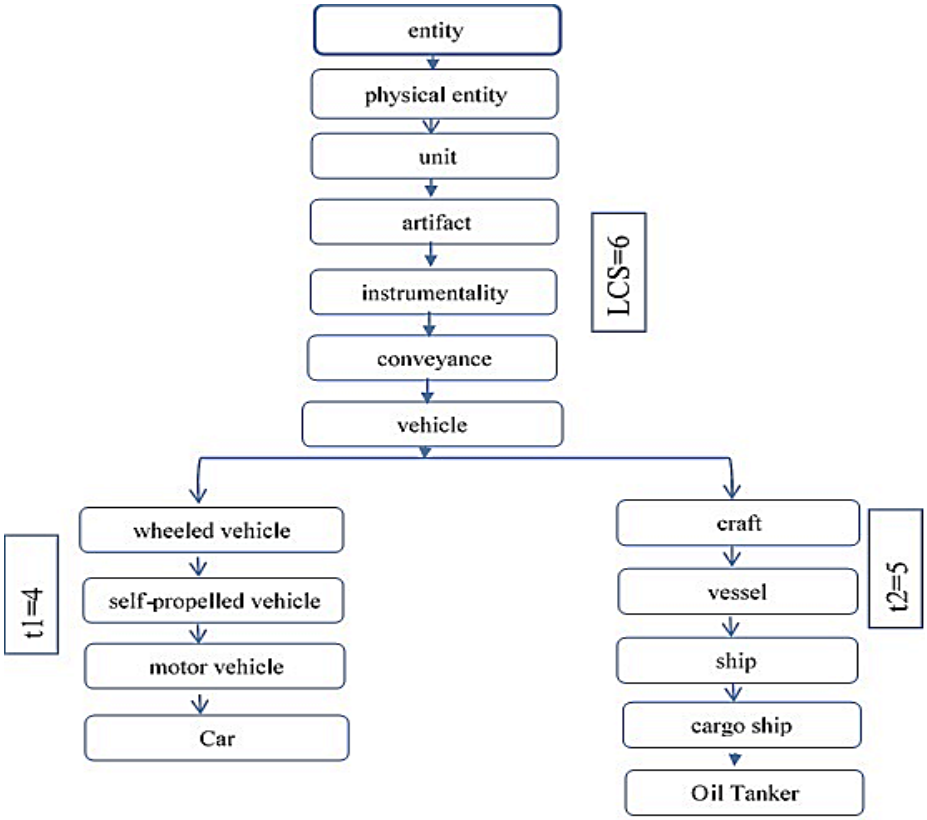

Figure 2.1 the Edge-counting measure

2.4.1.2 Feature-Based Methods

In this method, the semantic similarity was calculated based on properties of the words like gloss, concepts of neighboring, and non-taxonomic information modelled in ontology [59]. The meaning of Gloss is the semantic sense of a term in a lexical dictionary, where a set of glosses is called a glossary. There are many semantic similarity methods suggested depend on the senses of words. Similarity measures based on gloss take advantage of the fact that words with similar meanings have more common words in their gloss.

Semantic similarity is measured as the extent to which the gloss of the words under consideration overlaps. The Lesk [60]assigns the value of association between two words based on the overlap of the words in their meaning and the texts of the associated concepts in the ontology.

Jiang et al. [61] suggest a method based on feature where measure the semantic similarity by using the glossary of concepts found in Wikipedia. Most methods that are based on features take into account the common and unshared features of two words. Shared features increase the similarity value and non-shared features decrease the similarity value [59].

2.4.1.3 Methods Based On Information Content

information content (IC) for the concept is meaning the information extracted from the concept that appears in context[62]. A high value of IC indicates that the word is more specific and that it clearly describes a less ambiguous concept, while a low value of IC indicates that the word is more abstract in meaning[63]. The specificity of words is determined using Inverse Document Frequency (IDF), which is based on the principle that the more specific a word is, the less it occurs in the document. IC-value methods measure similarity between terms using their associated IC value.

2.4.1.4 Hybrid measures:

Are a method that combines between measures for various methods in order to take benefit of the advantages of all measures.

2.5 Ontology Learning

Ontology is the core of the Semantic Web, the success of the Semantic Web depends largely on the spread of ontology learning, and the engineering learning process required in ontology in order to gain knowledge efficiently. Ontology is a formal description of knowledge as a set of classes that are category of entities represent objects or types of things in ontology, while individuals represent the ontology community, which are instance of classes. Classes or objects are associated with properties called attributes. It allows inserting statements about data types and their data values, which explain the assignment of conceptual nodes to strings, numbers, and other data types.

Relationships are connects which identify the relation between entities [64]. Which are mostly split into two kind: hierarchical and non-hierarchical relationships. hierarchical relationships are shown as is-a or sub-class of relations, non-hierarchical relationships are represent other types of links, Which share in the enrichment of the ontology without changing its structure the axioms are the official definitions of knowledge for ontology , their aim is to describe the characteristics of perception[65][66]. Extremely expressed using descriptive logic and first order logic, mostly, axioms are mostly classified according to the function they express as follows [67][68]:

[1] Instantiation: an axiom instantiation that specifies an instance to a class.

[2] Assertion: The assert axiom assigns a value to an instance via a property, which is the components of the attributes described above.

[3] Subsumption: a subsumption axiom for two classes states that any instance of the subsumed class is also an instance of the subsuming class, while for two prop ties it states that any two instances connected by the subsumed property are also connected by the subsuming one[67].

[4] Domain: The Domain axiom specifies the domain class for a property. In other words, any connection to an instance and a value by that property means that the instance is from the domain class [66].

[5] Range: The range axiom assigns the scope class (integer, string, etc.) to the values of a property [66].

[6] Disjointness: this means that if a class have an instance another class cannot have the same instance, so, the classes are Disjoint.

[7] Arbitrarily complex axioms: most ontology languages permit for the statement of different axioms that is not part of any type given above[68].

Ontology provides not only knowledge representation that can be reused and shared as shown in figure (1), but it can also add new knowledge about the field[69].

The ontology learning vision contains a numbering of complementary Specialties, which feed on various kinds of structured, semi structured and

unstructured data to support the process of ontology engineering (Natural Language Processing), deep learning (DL), and mining [70].

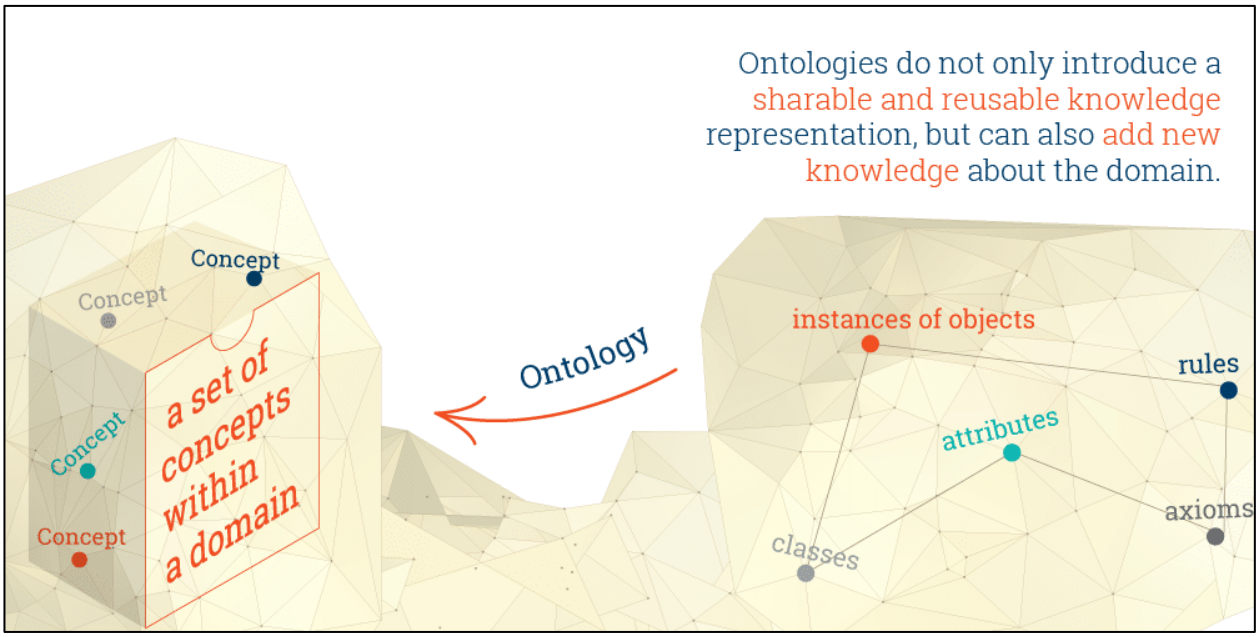

Figure 2.2 Knowledge Representation as Ontology

2.5.1 Ontology Specification Languages:

There are two classify of ontology specification languages, the web languages and traditional ones:

[1] Web Languages: A number of languages appeared for support semantic web, it is also used in the tasks of identifying ontology. For example, Extensible Markup Language (XML) [71]and Resource Description Framework (RDF). In the literature, can note that Web Ontology Language (OWL) is the generality used to perform ontology. OWL language relies on descriptive logic and

has several versions[72], that provide a balance between expression, decision-making and cost.

[2] Traditional Languages: unlike web languages, these classes depend on first order logic and Frame, Which gives a strong representation and is more expressive, However, given the trade-off among computational power and expressiveness, those languages overwhelmingly create non-determinable systems, making the reasoning very difficult. Among those languages, like Ontolingua [73] which is based on KIF3 (Knowledge Interchange Format)[74] and Frame Logic (F-Logic)[75].

2.5.2 Ontology Learning Method:

The techniques of Ontology learning seek to construct ontology in an automatical way or semi automatical. Where all the elements of the ontology are generated at once or by utilizing various tools to learn each component of the ontology. The ontology is made up of various elements. Concepts and relationships represent the main elements of the ontology. It is also said that ontology is expressive in a state that is well and sufficiently enriched with axioms, and there are many ways to learn ontology classified according to the techniques used in its construction.

2.5.2.1 Linguistic and Statistical Approaches

This method of learning ontology depends on the use of techniques depend on linguistic characteristics and methods of mining terms and their patterns in the data set. The learning steps are as follows:

[1] First, the concepts of ontology are collected through applying linguistic analysis to the data set, and the patterns of terms extracted differ in different languages and the domain of ontology, as an example, When the topic is general and in English, the forms of terms will be as follows ([Adj Noun] or [Noun Noun])). After that, linguistic filtering can be used, such as using stop-lists to remove duplicate results. The results are then categorized using mining techniques such as TFIDF [76]or C-value/Nc-value [77]Where terms are immediately considered concepts or are assembled into synonyms based on the Harris hypothesis, which shows that terms that appear in the same context and tend to have the same meaning are synonyms[78].

[2] Taxonomic Relation construction: Taxonomy or concept hierarchy is an important part of the ontology. Hierarchy relations are relations that provide tree visions of the ontology and determine the inheriting between concepts. WordNet is used, which is organized hierarchically lexical system motivated by actual linguistic, psychological, and computational theories lexical memory of humans. WordNet [79]is more like a thesaurus than a dictionary because it organizes lexical information in terms of word meanings (or senses) instead of word forms. The Hypernyms or the IS-a relation is extracted as a triple (concept, IS-a, concept) to generate the tree, A can be defined as “a hypernym of B if all B is a (kind of) A". All levels of the hypernym word can be organized into a hierarchy in which the sensing is structured from the most specified at the low levels to the most generic meaning at the top.

For example, “motor vehicle” is a hypernym of “self-propelled vehicle”[80].

[3] Non Taxonomic Relation construction: To obtain nonhierarchical relations such as Synonyms, antonym, meronymy, some technique is used to collect the above relation from BabelNet [81]and WordNet. To exemplify, the term car has the synonyms (car, auto, automobile, machine, motorcar) and the meronyms (accelerator, airbag, automobile engine, car horn, car door, car seat, etc.).

[4] Uses Web Ontology Language (OWL) to build the ontology axioms, which is the most commonly used to implement ontology. Where can build OWL based on Protégé 5 [82].

2.5.2.2 Machine Learning Approaches

In this method, machine learning is combine with statistical and linguistic approaches to improve results. For example, using the association rule algorithm to extract important links between words and using the hierarchical clustering algorithm to group words using similarity measures. As well as, classification algorithms, that are using to classify new concepts. It also uses inductive logic programming to discover new concepts from extended data. Conceptual clustering for learning concepts, and hierarchy of concepts [82].

2.5.2.3 Frameworks and Systems

To simplify the process of learning ontology for researchers in the domain of ontology, especially when building ontology depends on different techniques such as statistics techniques or linguistic techniques. Several ready-made methods have been built to create ontology. These methods differ according to the formal methods and methodology of building the ontology used, the algorithms used, the type of language and the ontology to be building. Such as GATE12[83], [84], LExO (Expressive Learning Ontology)[85], OntoLearn[86], OntoCmaps[87], OntoGain[88].

2.6 Word Sense Disambiguation:

In natural language processing (NLP), word sense disambiguation (WSD) is an automatic process by a machine for sensing the appropriate meaning of a word in a given context or in discourse. Natural language is ambiguous, so many words can be interpreted in multiple ways depending on the context in which they occur. The process of determining the meanings of words in context is known as word sense disambiguation (WSD)[89].

One of the types of ambiguity is lexical Ambiguity. This ambiguity occurs when a word carrying several meanings appears in the context. e.g.: bank. Bank can be sensed as a riverbank or as a place where people keeps their money and other valuables[90]. There are two approaches to

building a WSD algorithm, ranging from those which do not require training data, to models which are data-driven[91].

2.6.1 Knowledge-Based WSD

Knowledge-based approaches make use of computational dictionaries, such as WordNet or BabelNet, and particularly their graph, where nodes are represented by synsets and the edges represented by the relationships between the synsets. a graph algorithm is a Successful approach of this type, such as random walks (Agirre et al.) [92]and approximation of group (Moro et al.) [93]or theory of game (Tripodi and Navigli) [94] and a modified Lesk algorithm (Manish K. et al.) [95].

The knowledge-based method uses WordNet or BabelNet lexical knowledge. The process of demystifying the word includes three stages, starting with query, then pre-processing, and then WSD classifier.

The query process is represented by taking the query from the user as an input and sending it to the preprocessing unit, then the preprocessing unit converts the query into a regular query by adding the features of the part of speech marks and grammatical definitions, and then the classifier classifies the meaning of the word based on the lexical evidence base and the speech context, and the most important lexical databases are[95].

a) WordNet (Miller et al.) [96]: It is a lexical online database set up under the Supervised of Dr. George Miller in the Laboratory of Cognitive Science at the University of Princeton. WordNet is the more significant resource available to re-searchers in Linguistics Computational, text analyzes, and numerous related fields. Its layout

is inspired by actual lingual, psychological, and computa-tional theories lexical memory of human. The English nouns, verbs, adjectives, and adverbs are regular into synonym group; each one represents a basic lexical concept. Various relationships link sets of synonyms including hypernyms, antonyms, hyponymy, meronyms, holonymy, synonymy, tro-ponymy etc(McCrae et al.)[97].

b) BabelNet (Navigli and Ponzetto) [98] : is a multi-lingual encyclopedic dictionary with wide encyclopedic and lexicographic covering of terms, as well as a semantic network which relates named entities and concepts in a very large network of semantic relations. It has about 13 million Named Entities. Babel Net is designed within the Sapienza NLP collection and servicing by Babelscape. BabelNet simulates the WordNet where it is based on the notion of synset, but it expanded to contain multilingual lexicalizations. Each synset group of BabelNet represents a specific meaning and includes all synonyms that express that meaning in a group of several languages (Navigli et al.) [99].

2.6.1.1 Adapted Lesk Algorithm

The algorithm works to find the meaning of the target word from the Wordnet lexicon through information about the target word and the words surrounding it, which can derive meanings from the Wordnet lexicon. The algorithm uses a small window of context around the target word, and the context is a window of n-word symbols. To the left and anther to the right, to form a group of 2n words around the target word [100].

also, the target word will be included in context, So the context length is 2n + 1-word, If the target word is at the beginning of the sentence or at the end of the sentence, words from the other direction are added to the sentence, and this depends on lesk[100] suggestion where the same number of words must be provided for each window.

The algorithm compares the meanings of each pair of words within the context window. There is a set of relationships that determine the concurrency of pairs, an example of this, compare the definition of a word with the definition of hypernym of the other word, other semantic relationships are also compared, such as meronym, holonym, hyponym, troponym, and Properties of each word in pairs[100].

If the POS of the target word is defined, then the synsets and the relations are limited to that part of speech, or if the part of speech of the target word is unknown, we use all possible relations.

Since there are 7 possible relationships, there are at most 49 possible relationship pairs to consider for a given pair of words. However, if we know the POS of the word, or if the word is used only in a subset of the possible POS, then the number of pairs of relationships considered is less, When comparing two words, we calculate the overlap between them for each of the relationships, and once all comparisons of meanings are completed, we add all comparison scores to get the aggregate score for this group and take is the definition from WordNet[95].

2.6.2 Supervised WSD

The supervised WSD follows a machine–learning technique where it manually introduces a classifier from sense-annotated data sets. Basically, a classifier is concerned to process a single word and extract the proper sense of each instance of the word[101].

2.7 Data Mining

Due to the advancement of information technology and increasing of computer computation power and communication tools, the amount of data has been increasing continuously. This rapid growth in data required developing fast and modern methods to extract and generate helpful knowledge that can be used for supporting decision making. Therefore, data mining recently become one of the important techniques to achieve this goal[102].

Data mining is the framework to generate or extract unknown in advance useful information by exploiting a large structured or unstructured, huge and incomplete database. In other words, data mining is the process of generating and extracting previously undiscovered patterns and relationships among elements of a database. Data mining combines several research fields involving database technology, statistics, machine learning, neural networks, artificial intelligence, etc., which are incorporated to produce beneficial patterns of information[103].

For producing beneficial information from huge databases using data mining techniques, it is important to provide requirements and face some

challenges and solve them. First of all, it is necessary to determine what type of data will be handled by data mining techniques as well as what output we need to obtain.

Text Mining (TM) represents an essential branch of data mining that plays a primary role in several research fields like information retrieval, natural language processing. It is extracting beneficial information from a huge amount of data by following the same steps used by data mining[104].

2.7.1 TF-IDF

Term Frequency–Inverse Document Frequency (tfidf): It is a numeral statistic that reflects the importance of a particular word in a document within a group of documents and is widely used as a weighting factor in information retrieval, information discovery and other applications. tfidf value Increases in proportion to the number of times the word appears in the document but is counteracted by the frequency of the word in the documents[105].

This contributes to the fact that some words are more common in a particular field than others. tfidf It can be used efficiently to classify stop words and find important words in many topics such as text simplification and classification, tfidf is calculated by multiplying two statistical values, namely termed frequency and inverse document frequency, where TF represents the number of times a particular term appears in a document, and it is calculated by adding the number of times that term is intended to appear in each document[106].

$$Tf(t,d) = \frac{count\ \ of\ t\ in\ d}{number\ of\ word\ in\ d} \qquad (2.5)$$

Inverse Document Frequency (IDF) it is a statistical measure used to measure the importance of a particular term in a set of text documents, it is a feature that has been incorporated to reduce the weight of terms that occur most frequently in the document collection and raise the weight of less frequent terms. Where n is the total number of documents in the document set and df(t) is the document frequency of t.

$$IDF(t,d) = log\frac{N}{df+1} \qquad (2.6)$$

for each word in a document will calculate TF and IDF by using equations (5) and (6).Where d expresses the number of occurrences of the term in document d after that will calculate TF-IDF for each term in the document by using Term Frequency (Tft, d) and Inverse Document Frequency (idft, d)[105].

$$TfIDF(t,d) = Tf(t,d) * IDF(t,d) \qquad (2.7)$$

The algorithm TF-IDF is a method used to give each word in the document a value that represents the relevance of that word to the document as explained in Algorithm 2.1, it is very important in:

[1] Information retrieval: It is used to retrieve the information most related to the topic of the search, for example, searching for word

LeBron using search engines, the results will be displayed in order of relevance. This means that the high relevant articles of sports will be ranked higher due to TF-IDF giving the word LeBron a higher rank[105].

[2] Keyword Extraction: it's used to extract the keywords from a document, where the word with a high score is more relevant to the document [103].

Algorithm 2.1: ***Domain single Term Extraction***

Input: Tweets list /*list of tweets
Output: Single term score; /*list of term and its Score

1 Foreach Tweet in Tweets list do
2 d ← next tweet;
3 Foreach w in tweet do
4 w ← next word;
5 $Tf(w,d) \leftarrow \frac{count\ \ of\ w\ in\ d}{number\ of\ word\ in\ d}$
6 $IDF(w,d) \leftarrow log\frac{N}{df+1}$ */* n total number of tweets,*
/ df number of the tweet has w*
7 $TfIDF(w,d) \leftarrow Tf(w,d) * IDF(w,d)$
8 Return TF.IDF (w)
9 ENDFOR
10 ENDFOR
11 END

2.7.2 The C-Value / NC-Value Method For Automatic Recognition Of Multi-Word Terms:

The C-value/NC-value algorithm relies on combining statistical and linguistic information:

- The first part C-value, improves the statistical measures that depend on the repetition of the term to extract the main terms in the text, making it able to extract terms that consist of more than one word and overlapping terms. The second part includes:
  - ❖ A method for extracting terms that tend to appear in a particular context from words.
  - ❖ Use context word information to extract multi-word terms [107].

    This algorithm uses Automatic Term Recognition (ATR) to extract multi-word terms that are not dependent on a specific domain, the aim of this algorithm is to get better the extraction of overlapping terms, in the beginning, a set of documents is entered and then a list of candidate multi-word terms is produced, then they are arranged according to their termhood which is called C-value part, after that, the linguistic information for each term is extracted, which is POS tag, the type of extracted terms is constrained by linguistic filter and stop-list [87].

1) Part Of Linguestic:

The part of linguistic includes the following:

a) Information of Part-of-speech for the document.

b) It is applied the linguistic-filter to the tagged document to delete strings that do not need to be extracted.

c) The stop-list.

POS tagging is the assigning of a grammatical tagging (determiner, noun , adjective, preposition, verb) for all words in the document. It is necessary by the lingual filter that only allows extracting certain strings [107].

In the case that all kinds of terms such as descriptive phrases, noun phrases, verbal phrases, etc. were extracted, it does not need to use the linguistic filter, this case is not desirable, and is not used by any researcher in the field of ATR, the reason is that the resulting statistical information without the linguistic filter does not give good results. Where unwanted strings will be extracted since the majority of terms consist of names and adjectives[108], and sometimes preposition [109].We utilize a linguistic-filter that accepts these kind of terms.

The stop list, a stop-list for ATR is a words list which are not expected to occur as important or major terms in a specific domain where it is used to prevent them from being extracted as strings of terms, that leads to improve the accuracy of the output.

b. The Statistical Part

In this part, the algorithm uses a C-value metric which extracts the termhoods for the candidate string and categorizes it in a list of candidate strings, this scale is depend on the statistical properties of the candidate string which are:

a) The total frequency of the candidate string in the document.
b) Iteration of the candidate string as a part from a longer candidate string.
c) Number of candidate strings longer than them.
d) The length of the candidate string, (number of words for the candidate string). The termhood of (a) string is the number of frequency in documents.

$$\text{Termhood (a)} = \text{f (a)} \qquad (2.8)$$

Where a string is the candidate, f(a) is the acount of candidate string appearance in the document. And for computing the termhood for string should subtract from the number of its occurrences as part of a longer series of them.

$$Termhood\ (a) = f\ (a) - \sum_{b \in T_a} f\ (b) \qquad (2.9)$$

Where f(a) is the count of candidate string appearance in the document, The set of terms (a), that is, (b) is a terms candidate for (a), f(b) is the appearance number of candidate string (b) that contains a, T_a is the set of candidate terms that contain a.

The length of the candidate string is the last property in the C-value scale, the longest string in the document may not appear more times than the shortest string, where the appearance of the longest string is more important than the appearance of the shortest string, many times, so we include in the measurement the length of the candidate string [105].

Because a longer term cannot overlap with another term, there are two cases:

[1] If (a) is the longest string or is not nested with a longer string, then termhood will be its total frequency and the length.

[2] If (a) is a shorter string length, we should consider if they are a portion of any longer string terms, if it is a portion of other longer string terms, its termhood will also take into consideration its frequency as a nested string and the frequency of the longer string terms. Although her appearance as part of longer terms of candidacy negatively affects her candidacy, the higher these terms of candidacy, the greater her independence. This last number modifies the negative effect of nested candidate strings on longer candidate strings. The termhood scale, which is called the C value, is given as:

$$C\text{-}value = \begin{cases} log_2|a|.f(a) & a\ is\ not\ nested \\ log_2|a|.f(a) - \frac{1}{p(Ta)}\sum_{b\in(Ta)} f(b) & otherwise \end{cases} \quad (2.10)$$

Where a is the candidate string, f(a) is the acount of candidate string appearance in the document, Ta is the group of candidate terms that contain the term a. P(Ta) is the number of candidate terms that contain a, as it was explained previously that the scale of C-value depends on the frequency of occurrence of a. The negative effect is reflected on the

candidate string because it is a sub-series of a longer candidate term by the negative sign '–' in front of the

$$\sum_{b \in Ta} f(b) \qquad (2.11)$$

Where autonomy is given to the candidate from a term that has longer length terms by P(Ta), Thus, the higher the value, the greater the unrestraint, reflect by using P (Ta) as the denominator of a negatively signed fraction. The favorable effect for the length of the candidate is adjusted by applying a logarithm to it.

- NC-value: NC-value: after calculate the C-value, the algorithm will select the Context information such as noun ['NN','NNS','NNP','NNPS'] or adjective ['jj'], the NC-value will incorporation of context Information with C-value output list to extract multi-word terms, the nc-value will re-ranked C-value output list by using context information so that the most important terms appear at the top of the list [105]. The algorithm will create a list of important term context words from a set of terms extracted from a corpus. These will be ranked according to their `importance' when appearing with terms. The context words we treat are adjectives, nouns and verbs that either precede or follow the candidate term. The criterion for the extraction of a word as a term context word is the number of terms it appears with. The assumption is that the higher this number, the higher the likelihood that the word is `related' to terms, and that it will occur with other terms in the same corpus. can express the above criterion more formally with the measure:

$$weight(w) = \frac{t(w)}{n} \ldots\ldots\ldots (2.12)$$

Where w is the context word (noun, verb or adjective) to be assigned a weight as a term context word, Weight (w) the assigned weight to the word w, t(w) the number of terms the word w appears with, n the total number of terms considered. After that the NC-value measure is calculate as:

$$\text{NC} - \text{value (a)} = 0.8\ \text{C} - \text{value (a)} + \ 0.2 \sum_{b \in C_a}^{n} f_a(\text{b})\text{weight(b)} \quad (2.13)$$

Where a is the candidate term, C_a is the set of distinct context words of a, b is a word from C_a, $f_a(\text{b})$ is the frequency of b as a term context word of a, weight (b) is the weight of b as a term context word.

Regarding the weights, 0.8 and 0.2 that assigned to C-value and the context factor in the NC- value measure; these were chosen among others after a series of experiments. The combination 0.8, 0.2 gave the best distribution in the precision of extracted terms.

Algorithm 2.2: ***Domain multi-word Term Extraction***

Input: Tweets list /*list of tweets
Output: multi-word term score; /*list of multi-word term and its Score

1 Foreach d in Tweets list do /* d is a tweet
2 d ← next tweet;
3 candidate list ← $find$ candidate terms from d
4 Foreach w in candidate list do /* w is a candidate term
5 w ← candidate terms;
6 find Part-of-speech (w)
7 Stop-list Remove (w).
8 Find Number of times the (w) appears in a list.
9 Find Number of times of the (w) as part of other longer Candidate terms in the list.
10 Find the number of longer candidate terms.
11 Find the length of the (w).
12 Termhood (w) ← f (w)
13 Termhood (w) ←f (w)- $\sum_{b\in T_a} f(b)$
14 IF (w is nested) THEN
15 C-value= $log_2|w|.f(w)$
16 END IF
17 IF (w is not nested) THEN
18 C-value= $log_2|w|.f(w) - \frac{1}{p(Tw)}\sum_{b\in(Tw)} f(b)$
19 END IF
20 weight(w) $= \frac{\mathrm{t(w)}}{n}$ /*find weight for context words
21 NC − value (a) = 0.8 C − value (a) + 0.2 $\sum_{b\in C_a}^{n} f_a$(b)weight(b)
22 ENDFOR
23 ENDFOR
24 END

2.8 Entropy Weight Method

Entropy Method (EA) is a branch of the theory of information, it is a metric of the degree of disturbance in a system, a greater entropy value indicates a higher degree of perturbation. Interactions between factors and indicators can be calculated as a score for each scale, and can calculate the weight of the factors through indicators in the dataset by following the next step[110]:

[1] dataset Standardization: The first step in finding the weight of factors using entropy is to standardize the data because the data is not uniform[111][112]. Where the Normalize value of i_{th} in the j_{th} sample is denoted by p_{ij}, and the method for calculating it is as equation (2.14):

$$p_{ij} = \frac{xij}{\sum_{j=1}^{n} xij} \qquad (2.14)$$

[2] Information entropy Calculation: Information Entropy is an important factor to measure the weight of the evaluation metric, the great entropy of Information reflects higher weight.

[3] Weight calculation: After calculating the information entropy, the weight of each scale is determined using entropy theory, which reflects the scale's importance in the evaluation system. Where in entropy weight method, the value of entropy E_i of the i_{ij} index is calculated as equation (2.15) [113][114].

$$Ei = \frac{\sum_{j=1}^{n} pij * ln\, pij}{ln(n)} \qquad (2.15)$$

In effective evaluation with entropy weight method, it is generally determined when p_{ij} = 0 for ease of computation. The scope of the entropy value E_i is [0, 1]. The higher the E_i, the higher degree of differentiation of index i, and more information can be derived. Hence, more weight must be given to the index. Therefore, in EWM (entropy weight method) the metrics weights compute by equation (2.16).

$$wi = \frac{1 - Ei}{\sum_{i=1}^{m}(1 - Ei)} \qquad (2.16)$$

2.9 Criteria Importance Through Inter-Criteria Correlation Weighting Method (CRITIC).

The CRITIC Weighting Method is used to determine the weights of the criteria of relative importance in multi-criteria decision making problems (MCDM). It is proposed by Diakoulaki et al. [115], that have depend On the analysis of the evaluation matrix to extract the information contained in the evaluation criteria, In other words, it is found by deriving the objective weights to determine the amount of key information for each of the evaluation criteria, the method uses the standard deviation of the criterion and the correlation between the criteria to determine the weights.

Start by constructing a decision matrix $X = \left[x_{ij}\right]_{m*n}$, which includes m alternatives and n criteria, wherex_{ij} is the performance measure of alternative i with respect to criterion j, to get the weight

(w_j) of criterion j_{th}, the following notes are used: $\mathbf{C}_j$, is the amount of information in the criterion j_{th}, $\boldsymbol{\sigma}_j$ is the standard deviation for criterion j_{th}, ρ_{jk} is the correlation coefficient between criteria j_{th} and k_{th}. Based on these symbols, the calculation steps for the CRITIC method is explained in algorithm 2.3 are presented as follows Jahan et al. [116]:

Step 1. Normalize the initial resolution matrix through Eq. (2.17) For Utility Criteria:

$$r_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \qquad (2.17)$$

Step 2. Calculate the correlation between pairs of criteria through Eq. (18) and eq. (19):

$$\boldsymbol{\sigma}_j = std(V) = \sqrt{\frac{1}{m-1} * \sum_{i=1}^{m} (r_{ij} - \bar{r}_j)^2} \qquad (2.18)$$

$$\rho_{jk} = corr(V) \frac{\sum_{i=1}^{m} (r_{ij} - \bar{r}_j)(r_{ik} - \bar{r}_k)}{\sqrt{\sum_{i=1}^{m} (r_{ij} - \bar{r}_j)^2 \sum_{i=1}^{m} (r_{ik} - \bar{r}_k)^2}} \qquad (2.19)$$

Step 3. Calculate the weights of the criteria through Eq. (20) and eq. (21).

$$C_j = \boldsymbol{\sigma}_j * \sum_{k=1}^{n} (1 - \rho_{jk}) \quad j = 1, \ldots, n \qquad (2.20)$$

$$w_j = C_j / \sum_{k=1}^{n} C_k \qquad j = 1, \ldots, n \qquad (2.21)$$

Step 4. Calculate Mean Weight (MW) Deng et al.[117] through Eq. (22). MW is used when there is no information from the decision-maker or information is not sufficient to reach a decision [116]et al.

$$w_j = \frac{\mathbf{1}}{\mathbf{n}} \qquad (2.22)$$

Algorithm 2.3: Compute weight of features by CRITIC Method

Input: sheet of feature metric /* Excel sheet contains details of the user features metrics: Output: features weights;

1 For each value in Feature_value do

2 User_feature matrix ← Build (User * Feature) Matrix

3 ENDFOR

4 Determine Max, Min for all features column

5 Normalize the User_feature matrix by using equation (2.17)

6 $std(\text{matrix}) \leftarrow \sqrt{\frac{1}{m-1} * \sum_{i=1}^{m}(r_{ij} - \bar{r}_j)^2}$ /* Determine standard Division for User_feature matrix

7 $corr(V) \leftarrow \frac{\sum_{i=1}^{m}(r_{ij}-\bar{r}_j)(r_{ik}-\bar{r}_k)}{\sqrt{\sum_{i=1}^{m}(r_{ij}-\bar{r}_j)^2 \sum_{i=1}^{m}(r_{ik}-\bar{r}_k)^2}}$ / *Calculate the correlation Between pairs of criteria

8 $C_j \leftarrow \boldsymbol{\sigma}_j * \sum_{k=1}^{n}(1 - \rho_{jk})$ Calculate the key indicator

9 $w_j \leftarrow C_j \,/\, \sum_{k=1}^{n} C_k$ Calculate weight of feature

10 ENDFOR

11 Retrun Features Weights

12 END

CHAPTER THREE
The Proposed Information Diffusion Detection System

CHAPTER THREE
THE PROPOSED INFORMATION DIFFUSION DETECTION SYSTEM

4.1 Introduction

The proposed system allows discovering the extent of the spread of certain information about the Corona epidemic in Twitter, by using of the semantic web similarity based on an ontology specialized in COVID-19 concepts for Twitter to find the similarity between tweets.

In addition to using techniques to find influential users to reduce the volume and speed of data processing.

This system consists of four main stages: constructing the Ontology of COVID-19, influential users' detection, using the tweet of influential users, and applying semantic similarity theories to find the matching.

4.2 General Proposed Framework

The general Framework for finding the rate of spread of particular news within a group of tweets includes five main stapes as explained in figure (3.1).

1. Collecting data tweets from the Kaggle google repository.
2. Building the Covid-19 ontology constructor, which generates ontology specializing in the most frequent terms in tweets under the hashtag COVID-19 using the data collected. Where the concepts are generated with their properties and relationships.

3. Influential Users Detector, The system analyzes the Twitter users' features to study the influence degree of users within the network through three activities: social connectivity, attention obtained, and event activity.
4. Collects the Influencer's Tweets based on Influencers.
5. Using the hybrid semantic similarity algorithm is used to compute the similarity and detect the diffusion.

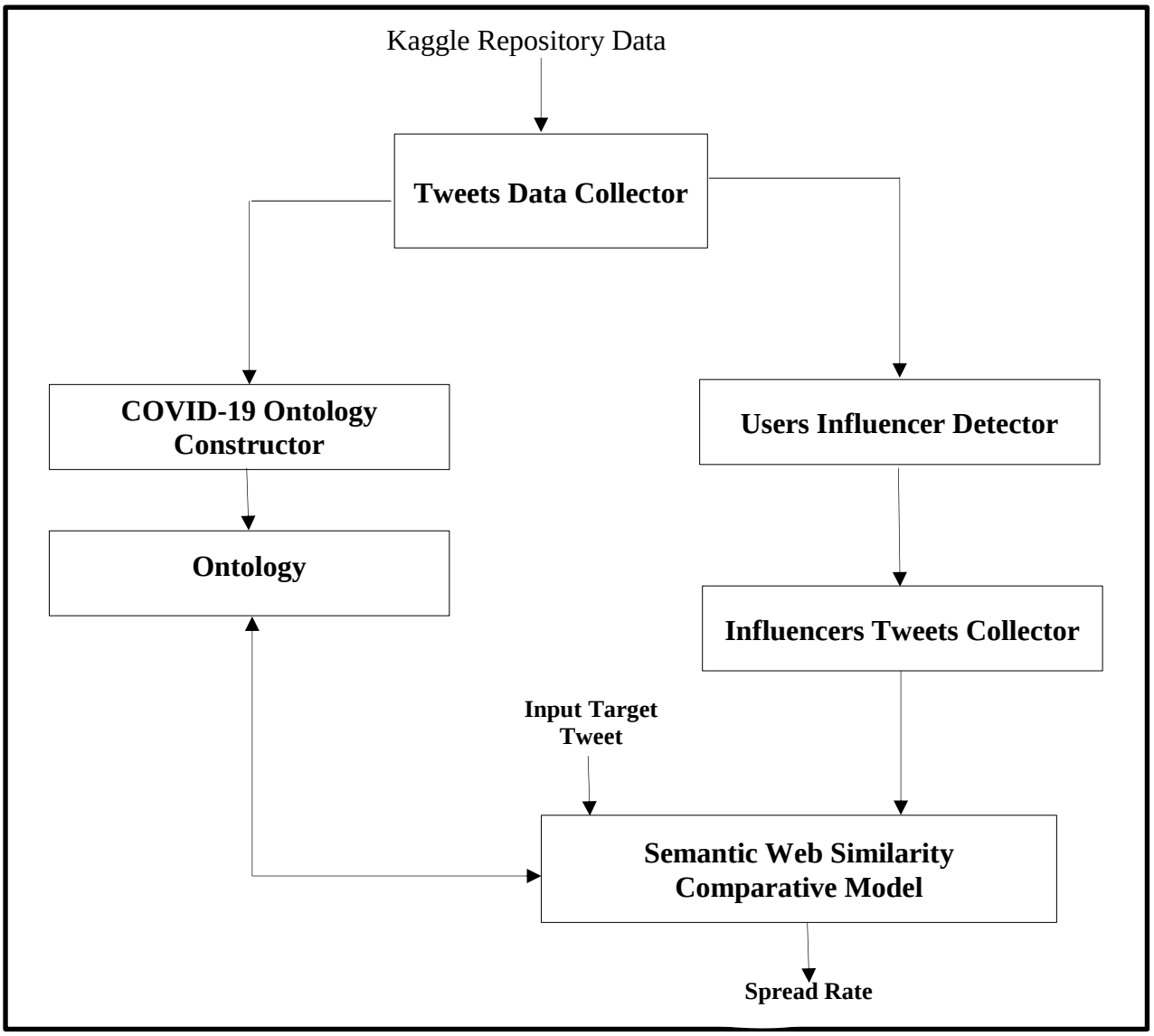

Figure 3.1 General Proposed Framework

4.3 Tweets Data Collector

The dataset used in this research was collecte from Kaggle that is a Google company, which acts as a gathering of researchers and data developers. Researchers interested in machine learning can join this gathering of more than one million individuals and display their data and development models. Kaggle offers a huge repository of community-published data and code. The data contain tweets that are set From March 29 to April 30, 2020.

The Tweets data will divide two-part. One is use to construct the ontology and another part is for finding information spread detection.

4.4 COVID-19 Ontology Constructor

Ontologies play a vital function in many semantic web applications. The manual way to construct ontology is an expensive and time-consuming process. While this work presents a system for ontology building via automatically learning. The ontology is construct with two complementary approaches. The first approach utilizes the COVID-19 tweets to extract the most frequent concepts in its, while the second uses medical ontology to enriching ontology with the medical covid-19 concepts. Figure (3.2) shows an architecture diagram of the proposed system for building COVID-19 ontology. The system contains six steps: (1) pre-processing the raw data, (2) Extracting the domain terms, (3) extracting the concept and Synonyms, (4) building taxonomic relationships, (5) building non-taxonomic relationships, (6) build Axioms and create OWL ontology by protégé.

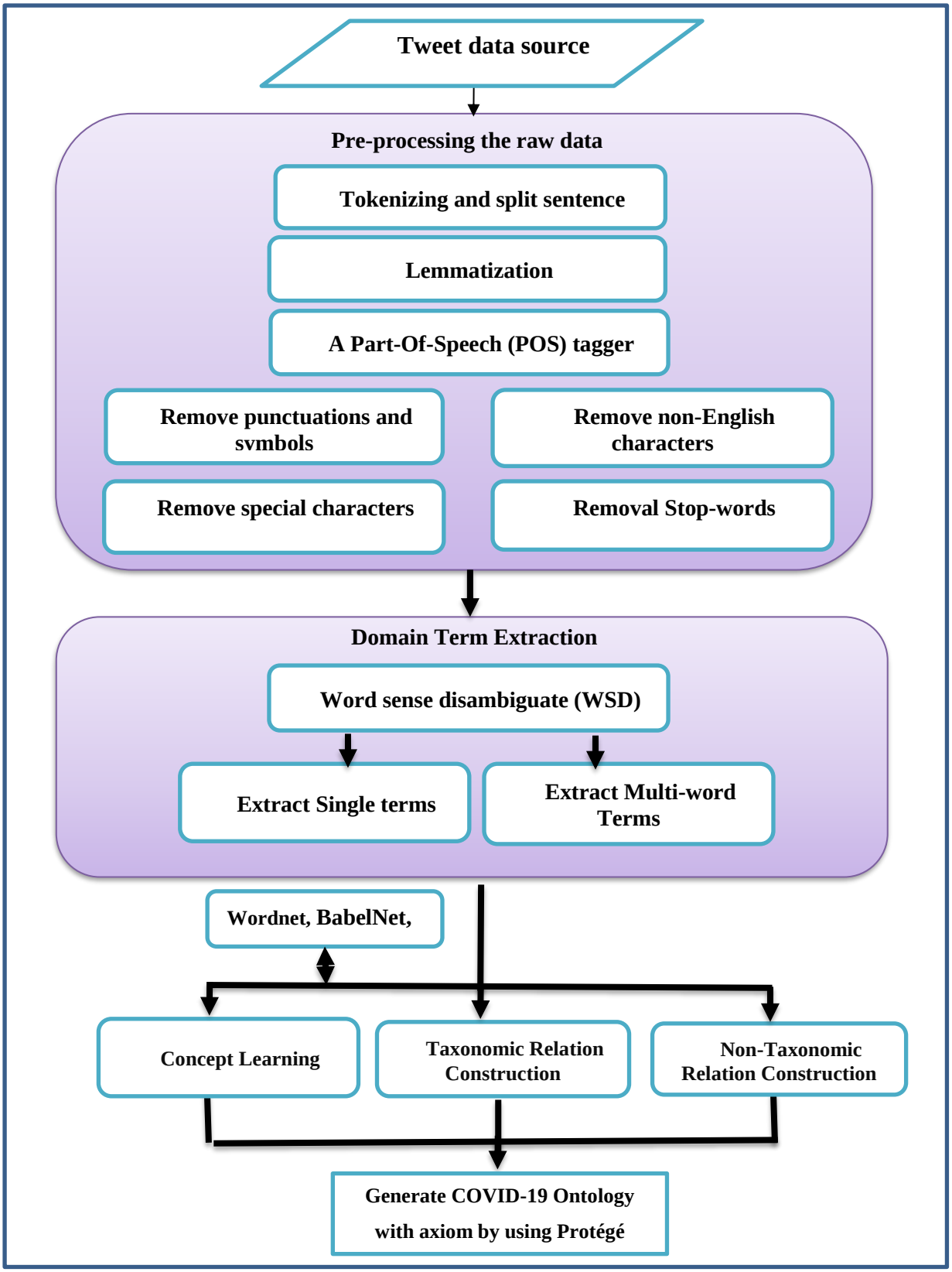

Fig. 3.2 COVID-19 Ontology Constructor

4.4.1 Data Preprocessing

The system used the dataset from Kaggle, which offers a huge repository of data; the data was collecte From March 29 to April 30, 2020. Tweets, by nature, hold a high amount of grammatical, abbreviations, and orthographical blunders. The 280 characters restrain of Twitter makes people use and shorten words (write "u" instead of "you") and abbreviations. Hence, this sort of content information needs intensive examination and pre-processing.

In the beginning, each of the hashtag symbols, @users, and link URLs are eliminate from the tweets, after that, the non-English words are deleted because the present study focuses on English tweets only. Any special characters, abbreviations, punctuation marks and stop-words are also remove because these do not contribute to the semantic sense of the message. Algorithms 3.1, 3.2 respectively, present algorithms for stopwords and punctuation Removing.

Algorithm 3.1: Stopword removing

```
Input: Tweets list, list of stopwords /*list of tweets and list of stopword
Output: Tweets_list_Wstopword   /*Tweets list without stopwords ;
 1   Begin
 2   Foreach Tweet in Tweets list do
 3       Foreach W in Tweet do    /* W is one word
 4           If  W not in the list of stopwords
 5               text  ←  text + W
 6           ENDIF
 7           Else
 8               Delete W
 9       ENDFOR
 10      Tweets_list_Wstopword ← add text
 11  ENDFOR
 12  Return Tweets_list_Wstopword
 13  END
```

Algorithm 3.2: Punctuation removing

```
Input: Tweets list, punctuation list /*list of tweets and punctuation
Output: Tweets_list_Wpunctuation   /*Tweets list without Punctuation;
 1   Begin
 2   Foreach Tweet in Tweets list do
 3       Foreach W in Tweet do    /* W is one word
 4          If  W not in the list of punctuation
 5             text  ←  text + W
 6          ENDIF
 7          Else
 8             Delete W
 9       ENDFOR
 10      Tweets_list_ Wpunctuation ← add text
 11  ENDFOR
 12  Return Tweets_list_ Wpunctuation
 13  END
```

4.4.2 Domain Term Extraction

Terminology extraction is a prerequisite for building the ontology of Tweets. The term may be a single word or multiple words, which refers to a specific meaning in a specific domain. The following steps explain the process of extracting the term:

1) Tokenizing and splitting sentences to reveal sentences and words boundaries.
2) Shorthand words to their basis by lemmatization: Several word/root retrieval techniques have been tried. Some algorithms rely on derivation by cutting the ending or start of a word, given the list of common suffixes and prefixes that can be found in the conjugated word. While the other algorithms depend in their work on

dictionaries to return the word to its source, such as Wordnet Lemmatize that used in this work, as shown in table (3.1).

Table (3.1) the description of lemmatize techniques	
word	Wordnet Lemmatize
Kites	kite
babies	baby
flying	flying
feet	foot

3) Use Part of Speech tagger (POS): to annotate the words with its grammatical category in context, thus specifying if it is a noun, verb, adverb, adjective, etc. as shown in table 3.2.

Table (3.2) the description of POS						
sentence	Ahmed drove a car to school.					
Tokens	Ahmed	drove	the	car	to	school
Lemma	Ahmed	drive	the	car	to	school
POS	‘NN’	‘VBD’	‘DT’	‘NN’	‘IN ‘	‘NP’

Where NN is a noun, IN is a preposition, NP is a proper noun and VBD is a verb, DT is a determinate.

4) Word Sense Disambiguate: In natural language, the word is polysemous when they have multiple meanings for example the word "bass" in "it has a bass sound" and “the grilled bass tastes delicious " has a different meaning. When a person reads these two sentences, the person knows the meaning of the bass through the words surrounding them, while the machine cannot distinguish the meaning. To remove that ambiguity of the words meaning, an adaptive lesk algorithm is use that depends on semantic relation as shown in table (3.3) and algorithm (3.3).

Table (3.3) the description of WSD

action	**Result**
Tweet	We took a step to save lives from the early moment of the start of the crisis in China. Now we must take similar action with Europe.
Preprocessing sentence	Take step save live early moment start crisis china. take similar action Europe.
WSD	take.v.08: take into one's possession step.v.06 : move with one's feet in a specific manner save.v.09:spend sparingly, avoid the waste of live.v.02: lead a certain kind of life; live in a certain style early.a.01: Before a certain time or before the expected time. moment.n.01: a particular point in time beginning.n.05: the act of starting something crisis.n.01: an unstable situation of extreme danger or difficulty china.n.01 : People's Republic of China, mainland China, Communist China, Red China, PRC. take.v.08: take into one's possession. similar.n.01: marked by correspondence or resemblance. natural_process.n.01: A process found in or produced by nature. Europe. n.01: the 2nd smallest continent.

Algorithm 3.3: Word sense disambiguate

Input: sentence /*string of words
Output: meaning of each word in a sentence

```
1   Begin
2   Foreach word in sentence do
3       create context window (2 words left, 2 word right the target word)
4       compare the target word senses with the context words sensing
5       select target word sense with the high value
6   ENDIF
7  Return meanings of the words of the sentence
8  END
```

5) Extract Single Terms:

The aims of the proposed system are to define the term characteristic of this domain. To extract the related terms, a simple mechanism may refer to concepts in that area by counting the term's frequency in a specific set of tweets.

In common, this approach depends on the premise that a term, which is repeat in a group of domain-specific texts refers to the concept, which belongs to this field. Information retrieval research has shown that there are more active ways to weight the term than simple frequency calculation, and this technique is based on weighing scales such as "TF.IDF", which is used in the present work to extract the individual terms as shown in the Algorithm (2.1). Total of (840) single-word terms found, Table (3.4) has samples of a single term.

Table (3.4) Single term Extraction			
Single term	**Score**	**Single term**	**Score**
death	0.139968	vaccine	0.205523
china	0.139086	symptom	0.204598
pandemic	0.121542	number	0.204359
unitedstates	0.117527	hydroxychloroquine	0.202992
mask	0.198004	affect	0.221822
emergency	0.196095	isolate	0.224954
infect	0.195579	dysprosium	0.227317

6) Extract Multi-Word Terms:

To find the terms that consist of more than one word, the C-value/NC-value measure is used, which is an important measure for finding multi-word terms that are presented. It does not depend only on the repetition of the term, but also on its overlap with each other, as shown in Algorithm (2.2). Moreover, this measure also uses contextual clues which are powerful signals of the sequence of some words.

Wherein this algorithm takes tweets as input and produces an output list containing candidate terms consisting of more than one word. These terms are ordered by their termhood. Candidate terms are arranged according to their termhood, the lists start from the top (high score) to the bottom (low score). The proposed system examined each of these parameters and extract (363) MultiWord terms, the Table (3.5) Shows examples of MultiWord terms.

Table (3.5) MultiWord term Extraction			
MultiWord term	**Score**	**MultiWord term**	**Score**
Cases deaths	1263405	Health ministry	336126
Covid patients	857392	Spread corona virus	1001382
Covid stay home save lives	119116	Vaccine covid	123206
Covid tests	251980	COVID-19 DNA vaccine	653391
Corona virus symptoms	112168	COVID 19 mRNA vaccine	254392
Health workers	499262	VRP-SARS-N vaccine	657252
Risk cases	311429	Pfizer BioNTech COVID-19 vaccine	437743
Crisis covid	366162	Chemical element	457362

4.4.3 Concept Learning:

The learning concepts are extracted through the use of lexical knowledge bases to extract the semantic definition of the term (the intentional concept description in the form of natural language description) and synonyms of the term.

The Proposed system does that through the use of two semantic dictionaries as shown in Algorithm (3.4).

- WordNet: It is a lexical online database, it's the more significant resource available to researchers in Linguistics Computational.
- BabelNet is a multi-lingual encyclopedic dictionary with a wide encyclopedic and lexicographic covering of terms, as well as a semantic network that relates named entities and concepts in a very large network of semantic relations. The Concept will be generate by adding uniquely a textual definition called gloss with their synonyms extracted from the above lexical knowledge bases.

Algorithm 3.4: Concept Learning

Input: single term, multi-word term /* the term extract by Alg. (2.1, 2.2)
Output: Concept glosses, Concept Synonym:

```
1   Foreach S_term in Single_term list do
2       W ← next S_term;
3       Defining (w) ← Fetch the official definition from
                       Semantic lexical
4       Synonyms (w) ← Fetch the Synonyms from
                       Semantic lexical
5       Concept (w) ← Build Concept from term,
                       Official definition, Synonyms
6   ENDFOR
7   Foreach M_term in Multi-Word term list do
8       W ← next Multi-Word term;
9       Defining (w) ← Fetch the official definition from
                       Semantic lexical
10      Synonyms (w) ← Fetch the Synonyms from
                       Semantic lexical
11      Concept (w) ← Build Concept from term,
                       Official definition, Synonyms
12  ENDFOR
13  return Concept glosses, Concept Synonym
14  END
```

For example, for the term ‘Covid-19’, the first statement in the algorithm will bring the official definition for Covid-19 from semantic lexical which is **" a viral disease caused by SARS-CoV-2 that caused a global pandemic in 2020".**after that the synonym list will be extracted ‘**corona', 'Covid-19', 'coronavirus disease 2019''Wuhan coronavirus', '2019-nCoV acute respiratory disease', 'Wuhan virus', 'Wuhan flu', 'novel coronavirus pneumonia', 'coronavirus', 'Wuhan pneumonia'** .the third part of the algorithm is to build an RDF triple **(subject, relation, object).**

4.4.4 Taxonomic Relation Construction

Taxonomy or concept hierarchy is an important part of the ontology. Hierarchy relations are relations that provide tree visions of the ontology and determine the inheriting between concepts. The proposed system used Semantic lexical to extract the Hypernyms, or the IS-a relation as a triple (concept, IS-a, concept) between concepts to generate the ontology graph.

The system concludes that, the concept A can be define as a hypernym of concept B if all B is a kind of A, and organized the hierarchy from the most specified at the low levels to the most generic meaning at the top. For example, “motor vehicle” is a hypernym of “self-propelled vehicle”.

After that, the graph is compressed by removing parent nodes with less than 2 children as shown in Figure (3.3). In case, there are three concepts (Oil Tanker, Train and car) and after extracting the hierarchical relationships (IS-a), a graph is constructed for each term (tree (a),tree (b),tree(c) respectively). The sub-trees are combined to build a hierarchical graph that includes all the concepts as shown in graph (d),

and then the graph is summarized to find the final hierarchical graph as shown in graph (e), this process done by algorithm (3.5).

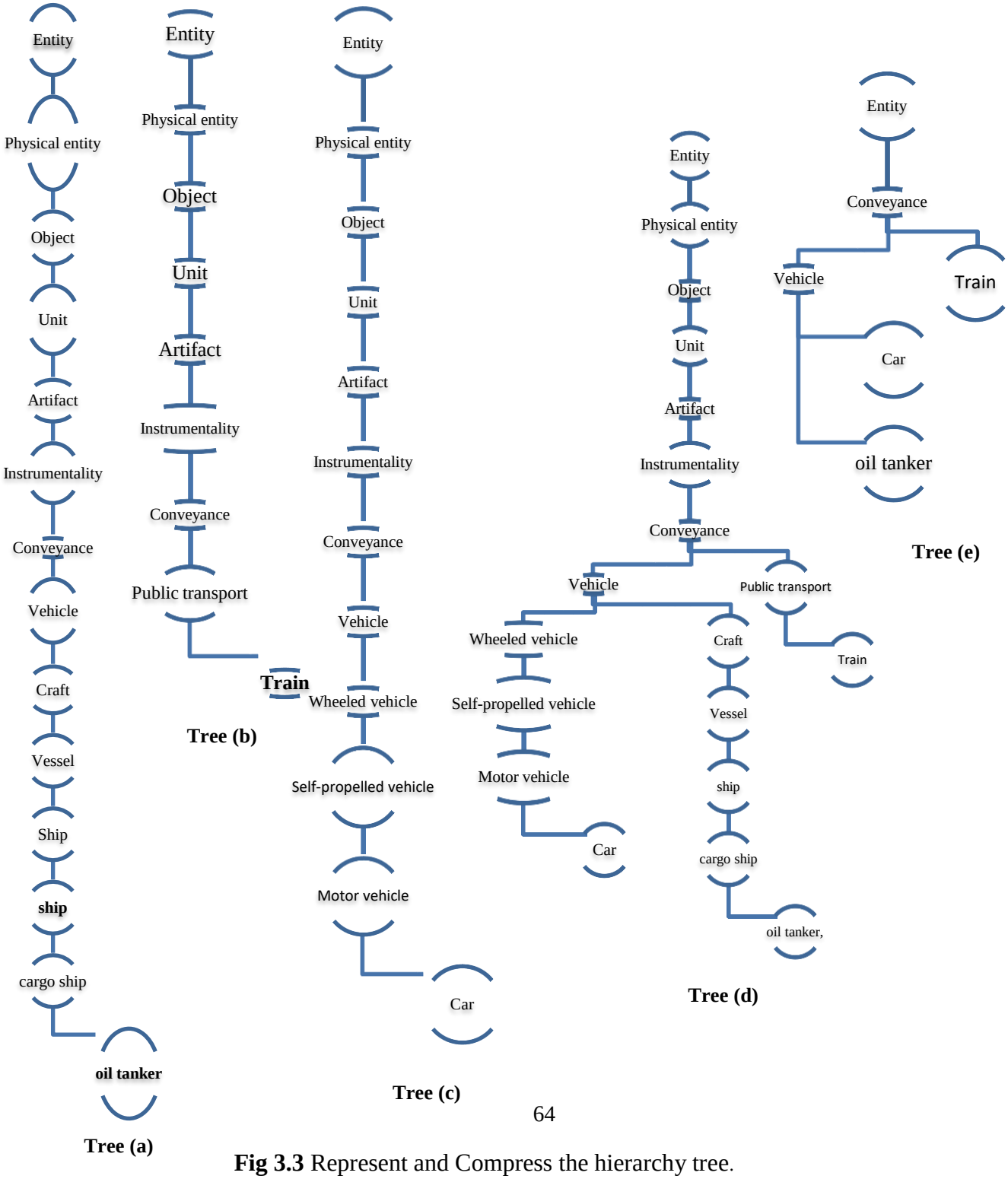

Fig 3.3 Represent and Compress the hierarchy tree.

Algorithm 3.5: Taxonomic Relation construction

Input: term;

Output: ontology graph;

```
1  Begin
2  For each term in term list do
3        W ← next term;
4        tree_hypernym (w) ← Fetch the IS-a relation from
                              Semantic lexical
5  ENDFOR
6  For each tree in tree list do
7        graph ← merge all tree at a point Least Common
                              Subsumer (LCS)
8  ENDFOR
9  FOR each Concept in Graph DO
10       If Concept is not leaf concept
11            If concept has one child
12                 Delete this concept
13            END IF
14       END IF
15 ENDFOR
16 Return ontology graph
17 END
```

4.4.5 Non-Taxonomic Relation Construction:

To obtain nonhierarchical relations such as antonym, meronymy, derivationally, has instance the API is using to collect the above relation from BabelNet and WordNet. To exemplify, the term car has the derivationally (automobilist, automobile, machinist) and the meronyms (accelerator, airbag, automobile engine, car horn, car door, car seat, etc.). As shown in figure (3.4).

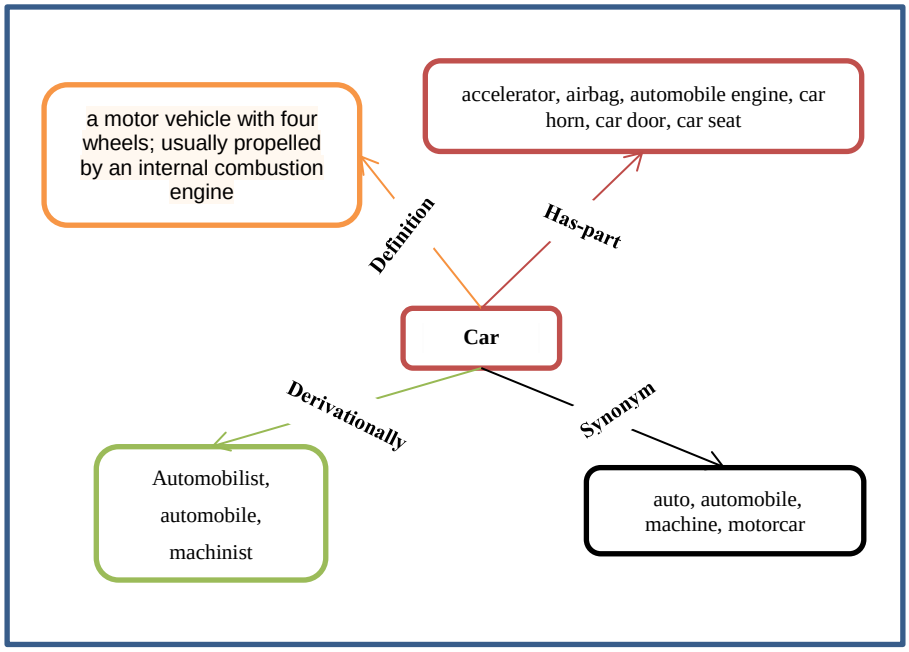

Fig. 3.4 Non-Taxonomic Relation

4.4.6 Build Axioms And OWL Ontology By Protégé.

The definitions of the OWL ontology will be created based on the meta-data above by using Protégé version 5.5.1, as shown in figures (3.5), (3.6) and (3.7).

Protégé is a knowledge base and ontology editor generated by the University of Stanford. Where it is used to create the classes, class

hierarchies, variable-value restrictions, variables, and the properties of the relationships between classes.

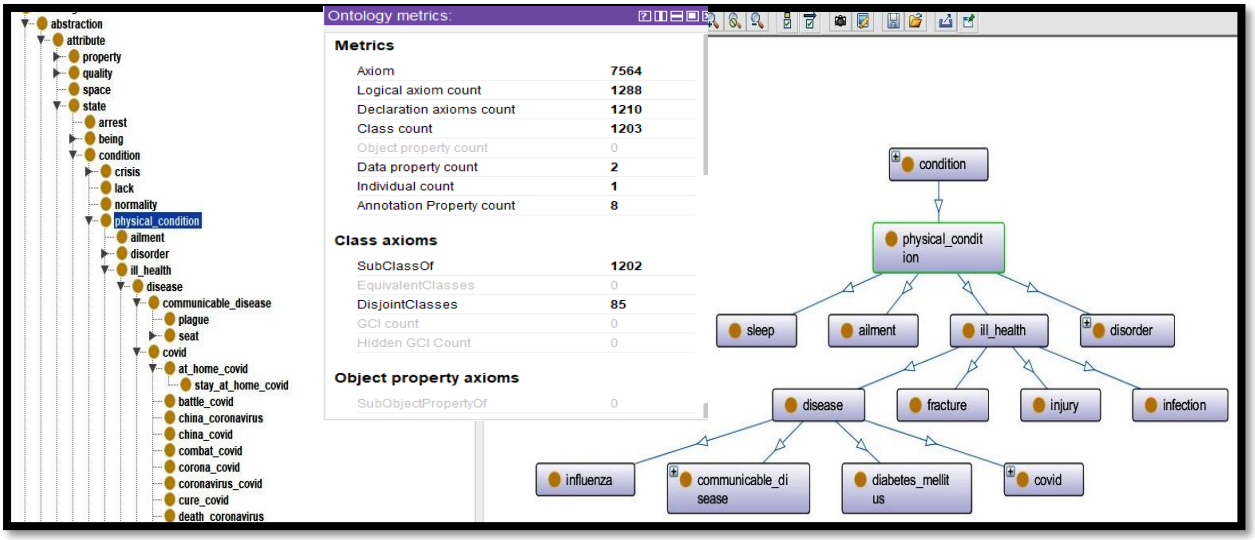

Fig. 3.5 Protégé API & ontology Hierarchy

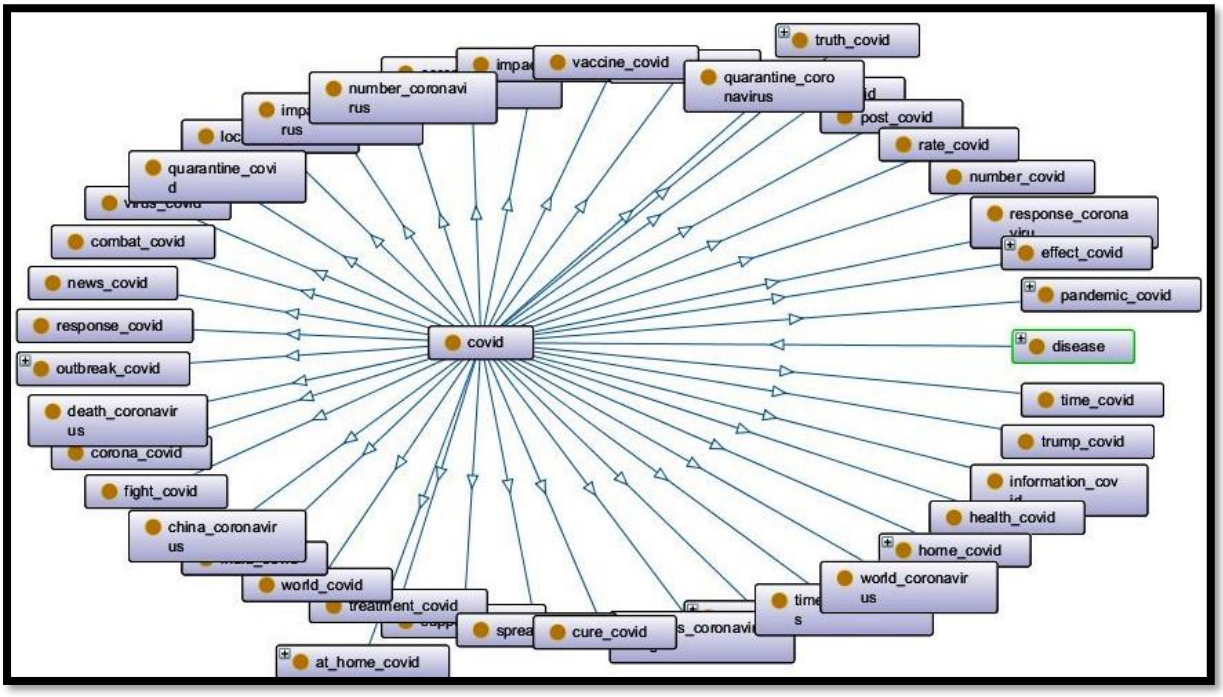

Fig. 3.6 Covid-19 concept Hierarchy

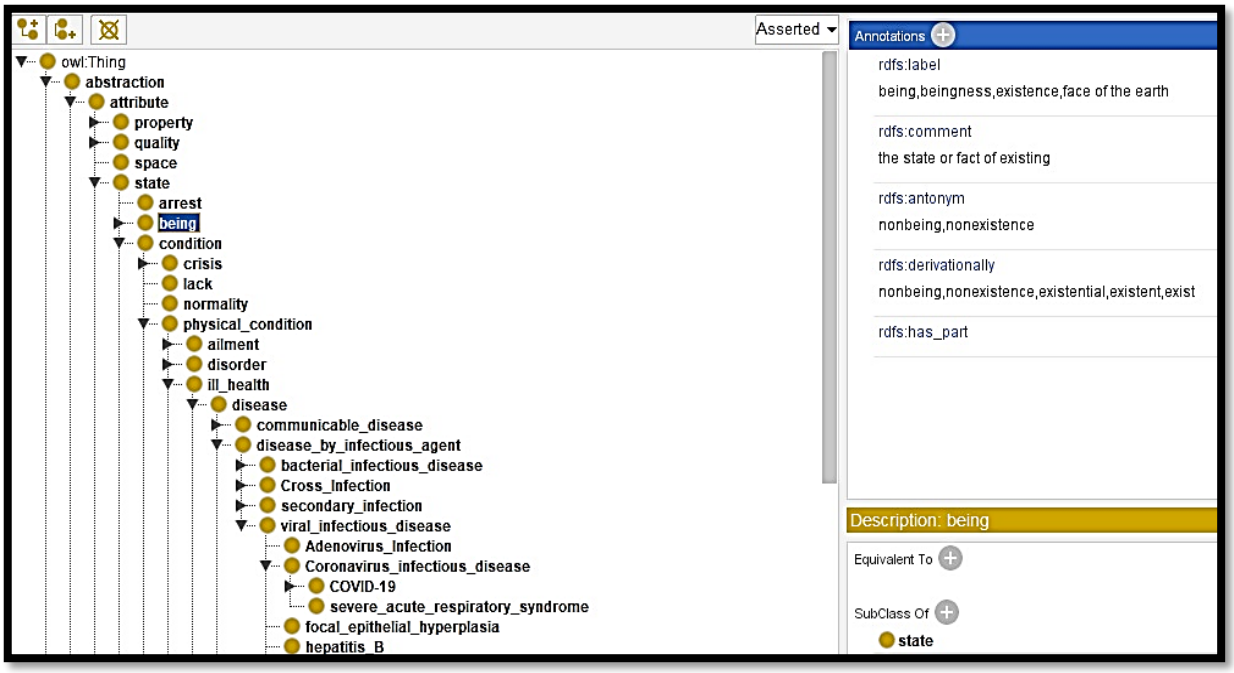

Fig. 3.7 Ontology relation and annotation

4.5 Users Influencer Detector

The work to extract influential users are based on three Characterized features: event activity, the attention acquired and the topology of society. Where it is difficult to measure a user's influence on Twitter using only the number of followers. After that, the system will use the messages of influential users to reduce the data and the speed of finding the spread of information increased.

The methodology to find influential users includes six stages as shown in fig 3.8:

1) COVID-19 Tweet preprocessing.
2) Compute the Features Value for Each User.
3) Compute weight of features by CRITIC Method.
4) Compute the weight of the features for each user.
5) Compute Users weight.
6) Extracting the Influential users and their messages.

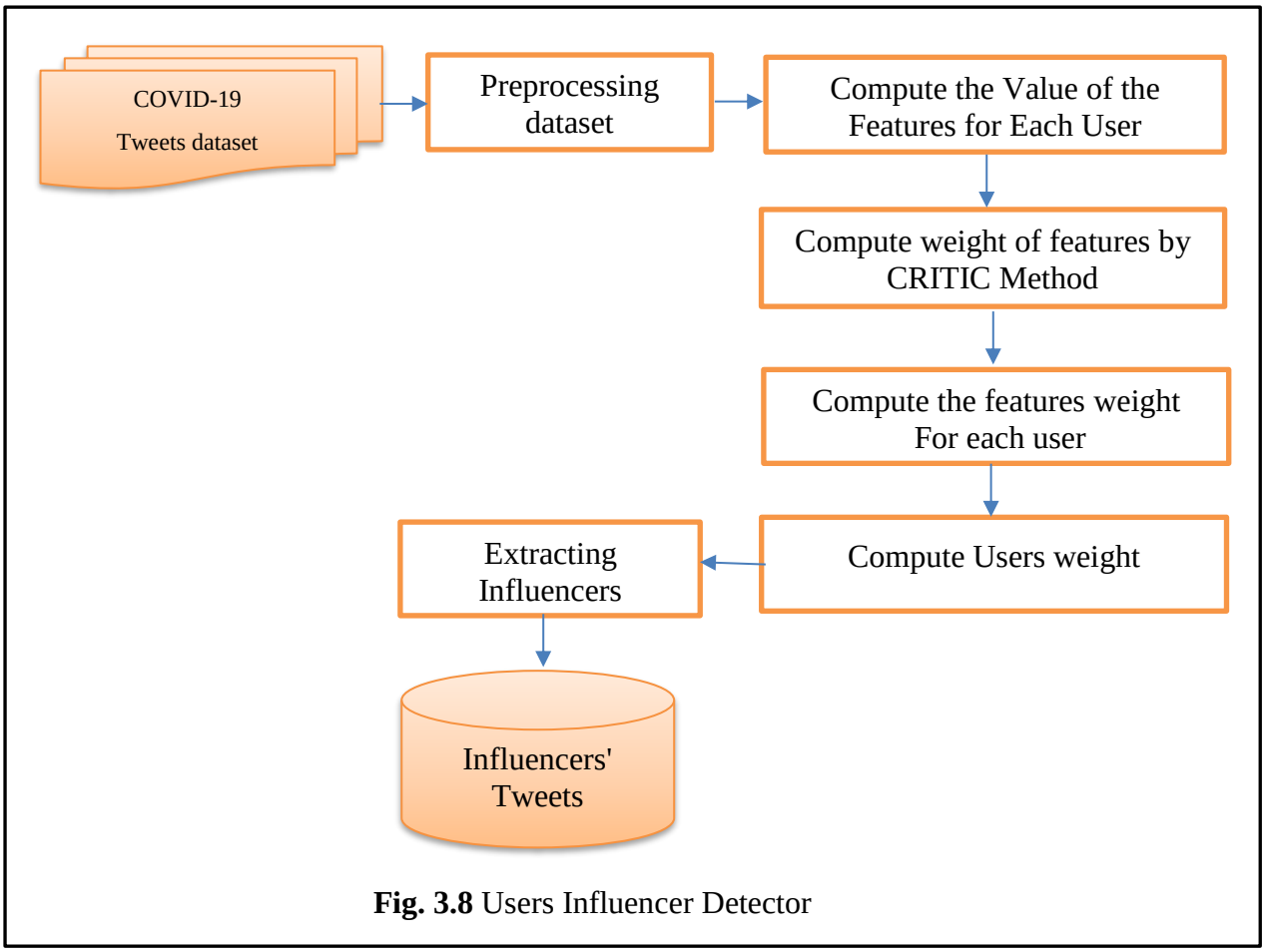

Fig. 3.8 Users Influencer Detector

The process of finding influential users starts by calculating the value of each feature for each user. Then, the weight of each feature is find compared with the rest of the features. The feature weight is multiply by its value to find the final weight and extract the weight of each user.

4.5.1 Covid-19 Tweet Preprocessing

For the experiment of current work, the dataset used is freely available in Kaggle is a Google company that operates as a community of data scientists and developers. Where contain 581,619 accounts and 3,584,764 million tweets as shown in tables (3.6, 3.7) which includes user information and tweet information respectively. The dataset will Pre-processed, which consist of removing unimportant or disturbing elements for the next phases of analysis and in the normalization of some misspelled words, order to provide only significant information, and remove the other.

Table (3.6): Example of User Data

id	text	user_id	in_reply_to_user_id	retweeted_status_id	retweet_count	reply_count	retweeted	num_hashtags	num_mentions	created_at
5.94E+17	How Randolph Hodgson	887281	0	0	0	0		0	0	Fri May 01 09:38:00 +0000 2015
5.94E+17	â€œTwitterâ€™s multi-b	887281	0	0	0	0		0	1	Fri May 01 09:11:07 +0000 2015
5.94E+17	The evolution of advertis	887281	0	0	0	0		0	0	Thu Apr 30 11:30:26 +0000 2015
5.94E+17	RT @rorysutherland: Plar	887281	0	5.93694E+17	14	0		0	4	Thu Apr 30 11:25:10 +0000 2015
5.93E+17	RT @davewiner: Some sa	887281	0	5.93213E+17	3	0		0	1	Wed Apr 29 05:17:36 +0000 2015
5.93E+17	RT @cdixon: "Lemonade	887281	0	5.92111E+17	545	0		0	2	Wed Apr 29 05:10:49 +0000 2015
5.93E+17	RT @benhammersley: I n	887281	0	5.90182E+17	257	0		0	1	Wed Apr 29 04:47:08 +0000 2015
5.93E+17	@profgalloway talks at D	887281	9273802	0	0	0		0	1	Mon Apr 27 10:52:55 +0000 2015
5.92E+17	RT @BI_Graphics: There's	887281	0	5.91711E+17	76	0		0	1	Sun Apr 26 16:45:09 +0000 2015

Table (3.7): Example of Tweet

id	name	followers_count	friends_count	favourites_count	listed_count	url	lang	created_at
488351800	MASC	568	387	15599	1	http://t.co/lzG1h(	en	Fri Feb 10 10:59:44 +0000 2012
75506408	Karan Baweja	208	263	43	2		en	Sat Sep 19 08:28:16 +0000 2009
37407255	à¦…à¦—à§à¦¨à¦¿à	7785	424	1157	217	https://t.co/8Feh	en	Sun May 03 12:17:01 +0000 2009
1664307984	laurel daniel	179	132	1224	0		en	Mon Aug 12 06:35:01 +0000 2013
308236961	King Devin	3325	1327	235	5	http://t.co/bhVTg	en	Tue May 31 02:07:05 +0000 2011
1446131569	Lily Phan	482	439	9909	0		en	Tue May 21 11:29:06 +0000 2013
887281	Paul Youlten	236	300	16	4	http://t.co/GipGL	en	Sat Mar 10 22:25:08 +0000 2007
3013763124	Gabrielle K â™¡	131	322	51	0		en	Sun Feb 08 22:35:05 +0000 2015
2982524516	Karla Albores	88	117	310	0		en	Wed Jan 14 11:50:48 +0000 2015

4.5.2 Compute the features value for Each User

The proposed system use eight measures of features, mentioned in the table (3.8) that described in section 2.3.2, to calculate the value of each feature, the system uses an algorithm 3.6 :

Table (3.8): Twitter Measures of Features

No.	Measures of Features
1.	The ratio of Follower to Following (Rf)
2.	Ratio of Retweet and Reply to user's tweets (Rrm)
3.	The sum of the number of retweets and reply acquired by the user (Urmo)
4.	Followers number (Ff)
5.	A number of tweets (Nt)
6.	Number of total like acquired by the user (Nu_like):
7.	New Reply and retweet (NeRepl_Rt)
8.	New Tweets (NeT)

Algorithm 3.6: Compute the features value for Each User

Input: UserSheet, TweetSheet, /* excel sheet contains users and other sheet contain tweets attribute detail

Output: features value for Each User,

```
1  Begin
2  For each record in UserSheet do
3     UR ← record;    /*User record
4     For each record in TweetSheet do
5        TR ← record; /*Tweet record
6        IF UR _user_ID equal TR_user_ID do
7           Rf ←  The ratio of Follower to Following
8           Rrm ← Ratio of Retweet and Reply per Total tweet
9           Urmo← calculat Number of Ret. and reply acquired by the user
10           Ff ← calculat Followers number
11          Nt ← calculat Total number of tweets
12          Nu_like ← calculat Total Number like acquired by the user
13          NeRepl_Rt ← calculat New Reply and retweet ()
14          NeT ← calculat Total number New Tweets ()
15     ENDIF
16  ENDFOR
17 END
```

The algorithm works to find the value of all measures for each user. At first, it links the user profile table with the tweets profile table through

the user ID. Then it works on calculating the value of each measure through a set of mathematical models and direct calculations.

4.5.3 Compute Weight of Features by CRITIC Method

Instead of directly, comparing the values of the influencing features, due to unknowing the weight of each feature in relation to the rest of the features, and to find the node that has maximum impact in the network, the proposed system used CRiteria Importance through Inter-criteria Correlation method (CRITIC).

In this study, objects refer to users and the indicators refer to features, and our dataset contains all feature values for each user. It is used to specify weights from characteristics of the data itself without any human factors as explain in algorithm (2.3).

4.5.4 Compute the Weight of The Features For Each User

The weights will be compute for specific user through multiplying the feature value by its corresponding value in the CRITIC weight vector, and then find the total user weight by summing all the feature weights for that user.

4.5.5 Extracting the Influential

Finally, the system sorted users by the total weight, the high ranking indicates the most influential user while the lower ranking indicates the minimal influential user.

4.6 Semantic Web Similarity Comparative Model

In order to discover the spread of specific information such as cases, deaths, economic status and other information within a specific topic such as the COVID-19 hashtag.

Our method is based on finding the semantic web similarity based on the source of knowledge (ontology) to find the similarity between the target message and the rest of the messages, the method finds the similarity between various concepts in the semantic space in such a way that the similarity becomes greater as the distance decreases. As shown in figure (3.9) the step of finding Semantic similarity between two tweets consists of:

1) Preprocessing and Remove ambiguity from tweets by using (WSD)
2) Extracted the concept characteristics (semantic formal definition, taxonomy relation, non-taxonomy relation) from COVID-19 OWL ontology by using SPARQL Query Language for OWL.
3) Calculate the entropy vector to give a weight for each measures.
4) Finding similarities using hybrid measures that combine edge-calculation methods and feature-based methods.
5) Calculate the rate information spread.

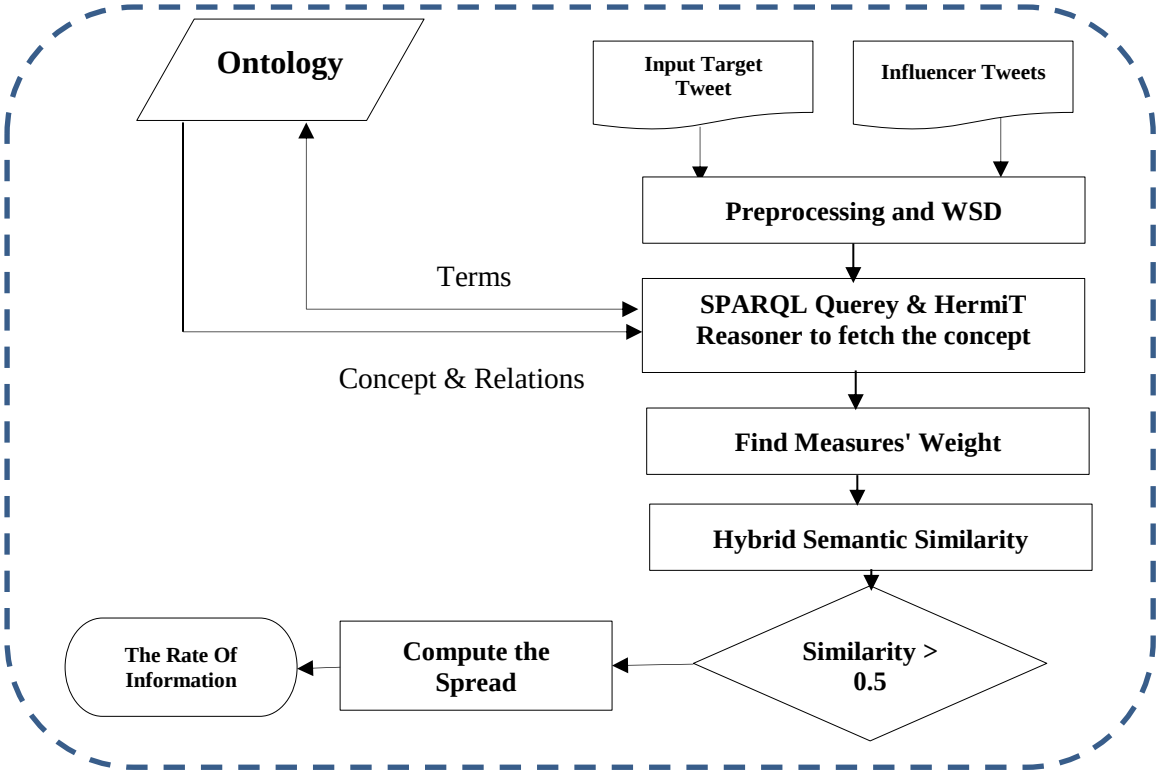

Figure 3.9 Semantic Web Similarity Comparative Model

4.6.1 Preprocessing and Remove Ambiguity from Tweets by Using (WSD)

Tweets, by nature, hold a high amount of grammatical abbreviations, orthographical blunders and stop words. In the beginning, each of the hashtag symbols, @users, and link URLs are eliminated from the tweets, after which the non-English words are deleted because the present study focuses on English tweets only. Any special characters, abbreviations, punctuation marks and stop-words are also removed because these do not contribute to the semantic meaning of the message.

In natural language, the word is polysemous where they have multiple meanings. To remove that ambiguity of the words, we used an adaptive lesk algorithm that depends on semantic relation, will be used to

extract the true meaning of the word. This algorithm relies on comparing the semantic structure (different levels of hypernym) and glosses of the target word with the surrounding words.

4.6.2 Extracted The Concept Characteristics

After the tweets have been purified and ambiguity removed, each tweet is converted into a vector of concepts by bringing the concept of each word from the ontology by using SPARQL Query that was found by World Wide Web Consortium (W3C) it is currently supported by most formats RDF triple. The technique of execution of SPARQL Query depends on subgraph matching, SPARQL provides a technique through which the interaction of users and programs with the ontology.

In this work we used SPARQL Query based on the HermiT reasoner to bringing the concept of each word from the COVID-19 ontology which can come within complex expressions of class or also be associated with class or property names as shown in table 3.9.

Table 3.9 bringing the concept of each word from the COVID-19 ontology by using SPARQL Query based on the HermiT reasoner

Action	Result
Tweet	The number of covid-19 deaths and infection
Pre-processing sentence	number covid-19 death infection
// Concept graph extraction by SPARQL Query PREFIXxs: :http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22# **Select ?object** **where { xs: concept rdfs:subClassOf ?object** **}**	**(Number** IS-a indefinite_quantity IS-a measure IS-a abstraction IS-a entity). ---------- **COVID-19** IS-a Coronavirus_infectious_disease IS-a viral_infectious_disease IS-a disease_by_infectiousagent,disease IS-a ill_health IS-a physical_condition IS-a condition IS-a state IS-a attribute IS-a abstraction IS-a entity).

	---------- **(Death** IS-a happening IS-a event IS-a psychological_feature IS-a ab straction IS-a entity) ---------- **(infection** IS-a ill_health IS-a physical_condition IS-a condition IS-a state IS-a attribute IS-a abstraction IS-a entity)

// Concept non-taxonomy relation extraction by SPARQL Query
 	
PREFIX xs: http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#
select ?label ?comment ?has_part ?antonym ?derivationally
where
{
xs:concept rdfs:label ?label ;
rdfs:comment ?comment;
rdfs:has_part ?has_part ;
rdfs:antonym ?antonym ;
rdfs:derivationally ?derivationally
}
	(Number HasSynonyms number,figure)
	(Number IsDefinedBy a concept of quantity involving zero and units)
	(Number has_Part)
	(Number has_Antonym)
	(Number hasderivationally number,figure)
	---------- 	
	(COVID-19 HasSynonyms corona, COVID-19, coronavirusdisease, coronavirusdisease2019,
	2019-nCoV acute, respiratorydisease, novelcoronaviruspneumonia, Wuhancoronavirus, Wuhanvirus,Wuhanpneumonia, Wuhanflu, coronavirus, covid-19)
	(COVID-19 IsDefinedBy a viral disease caused by SARS-CoV-2 that caused a global pandemic in 2020)
	(COVID-19 has_Part)
	(COVID-19 hasAntonym)
	(COVID-19 hasderivationally)
	(Death HasSynonyms death decease expiry
	(Death IsDefinedBy the event of dying or departure from life
	(Death has_Part)
	(Death has_Antonym brith)
	(Death hasderivationally die,decease)
	----- 	
	(Infection HasSynonyms infection)
	(Infection IsDefinedBy the pathological state resulting from the invasion of the body by pathogenic microorganisms)
	
	(Infection has_Part incubation)
	(Infection hasAntonym;)
	(Infection hasderivationally;)

4.6.3 Applied Hybrid Measures To Find Semantic Similarity.

The proposed method finds the semantic similarity by combining two methods, one based on graph and the other based on Features, which Compare the formal definition, graph and non-taxonomy relationships of each concept in the first sentence with all concept relationships of the second sentence, and selects the best similarity value for each relation, Then, it does the same work between the second and first sentence, to produce six similarity values (graph (IS-a), Has_Definition, Has_Synonyms, has_ Antony, Has_part and Has_Derivationally), after that, multiply the value of each similarity measure by its weight to find the final similarity.

1) Calculate The Edge-Counting Measure:

The most obvious way to calculate edges is to look at the ontology as a graph that links the terms taxonomically, and it uses these edges to calculate the similarity between two terms.

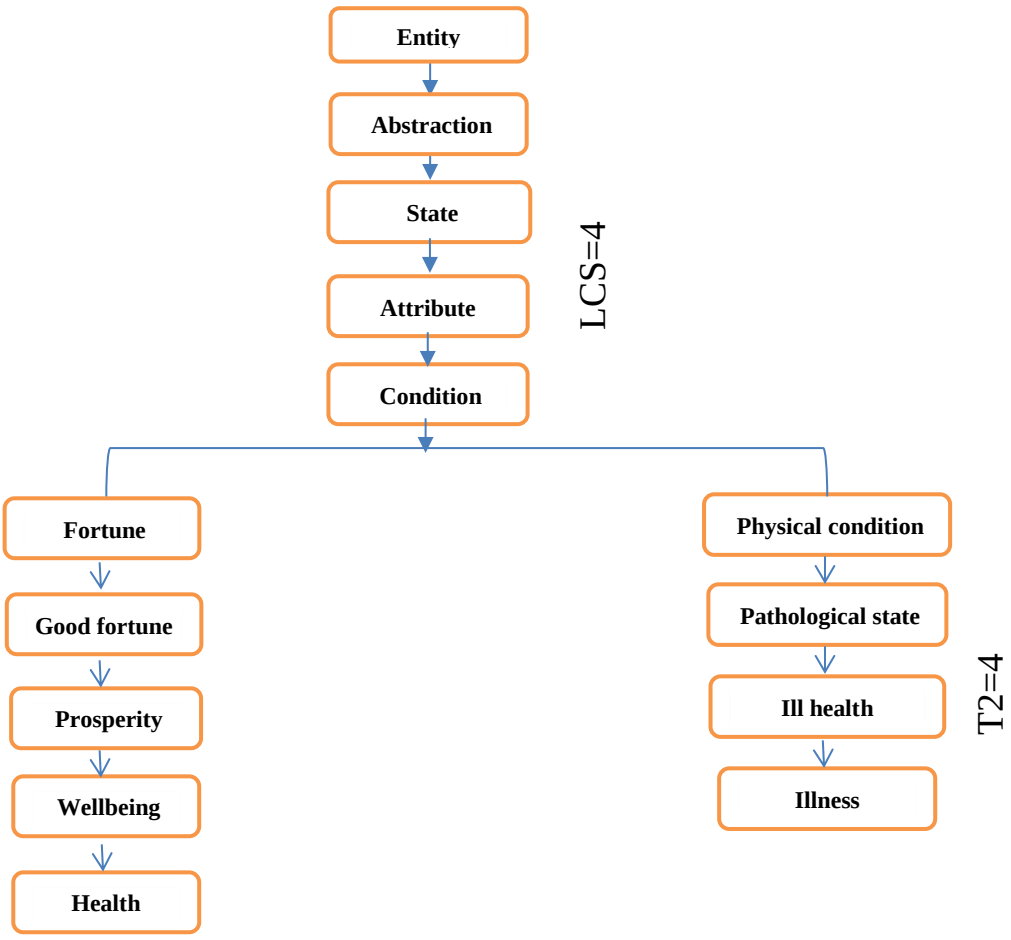

Figure 3.10 the Edge-counting measure

$$\text{Sim}(\text{Health}, \text{Illness}) = \frac{2 * 4}{5 + 4 + 2 * 4} = \frac{8}{17} = 0.47058$$

In this research, used the Wu and Palmer Scale, as the depth of terminology in the ontology is an important advantage. Wu's scale is based on counting the number of edges between two terms and their Least Common Subsumer (LCS), LCS is the mutual ancestor for these two

terms in the specific ontology. Figure 3.10 is an example of calculating the Edge-counting measure.

2) **The Feature-Based Methods:**

In this method, The similarity calculate as a function of concept properties, such as glosses (Has_Definition), Has_Synonyms, has_ Antony, Has_part and Has_Derivationally, Gloss is a meaning of a concept in a COVID-19 ontology, and other features are relations that owned by the concept.

The similarity of glossy and other features is measure by the extent of the interference between the glossiness of the words under consideration, by using a cosine similarity measure as shown in fig 3.11.

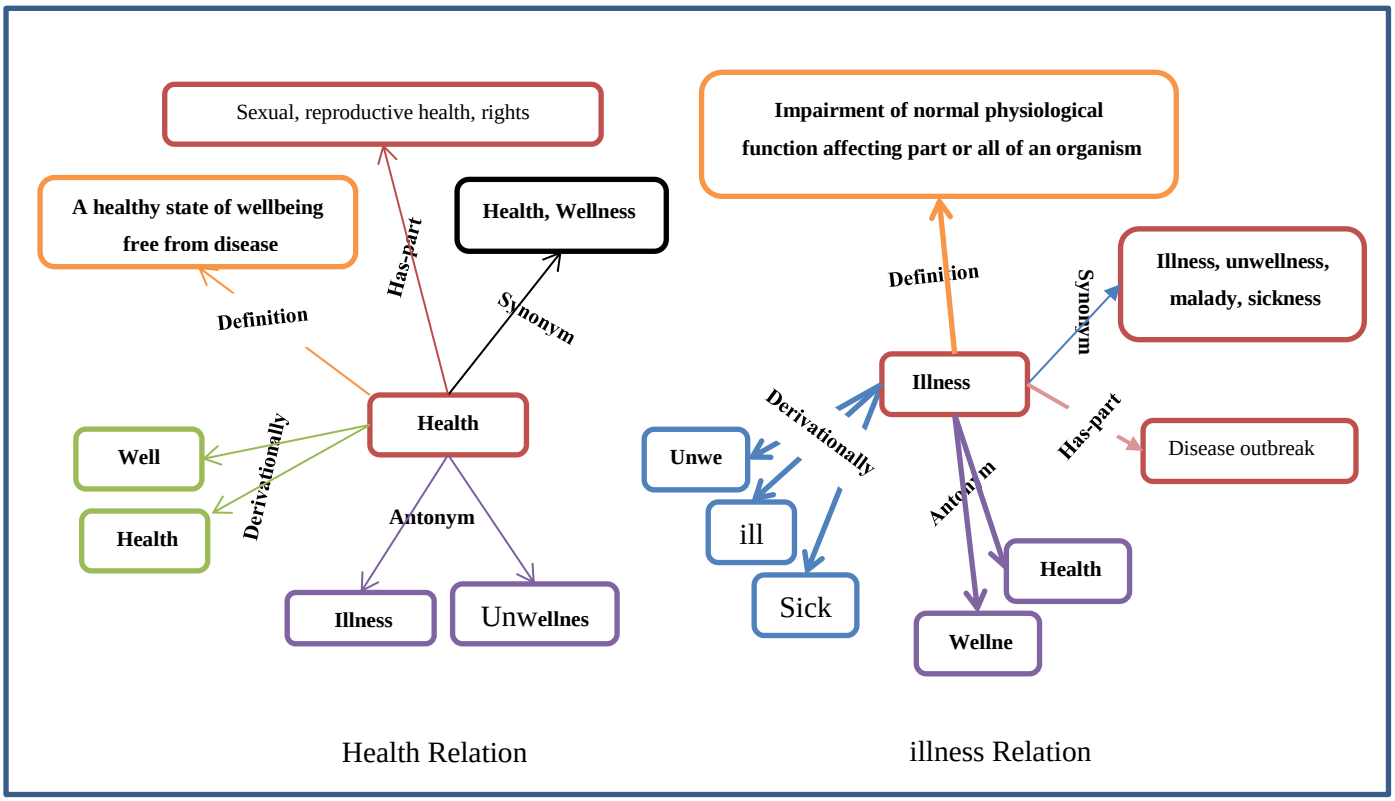

Figure 3.11 Feature based Relation

4.6.4 Calculate the Entropy Weight Vector for Scales

Instead of direct sum of the values of the semantic similarity, and due to unknowing the weight of each measure to the rest of the measures. The system used the entropy method to find the weight of each measure.

The object refers to pair of tweets and the indicator refer to the Similarity scale value, and our result contains all the similarity values of each measure for each concepts pair. The entropy method is an objective enabling method. It is use to specify weights from the characteristics of the data itself without any human factors.

4.6.5 Calculate Final Similarity.

After finding the similarity vector between tweets A and B and the similarity vector between B and A, and in order to find the similarity between the two similarity vectors, multiply the norm value of the first vector with the norm value of the second vector, then normalize the values to make the values within the Range 0-1.

After that, the system multiplies the entropy weight of the scale by the similarity value for that scale and then sum the resulting similarity values from the previous step to find the final similarity value as shown in the equation and algorithm 3.7.

Sim (Tweet A, Tweet B) = α (sim(Edge-counting)) + β(sim(Synonyms))+ £ (sim(definition)+ γ(sim(Antonym)+δ(sim(Has-part)+ η(sim(Derivationally)) (3.1)

After calculating the final value of the similarity between the target tweet and the rest of the tweets. If the similarity value is greater than 0.5, it is considered similar and if it is less than 0.5 then it is neglected.

By dividing the number of tweets similar to the goal tweet by the total number of tweets, as shown in the equation, we deduce the size of the spread.

$$\text{Spread rate} = \frac{\text{Number of similar tweets to the target tweet}}{\text{Total Number of tweet}} \qquad 3.2$$

Algorithm 3.7: Calculate Semantic similarity between two tweets

Input: target Tweet, List of Tweet;

Output: Semantic similarity value

1 Begin

2 Tar_twt ← target Tweet;

3 For each tweet in Tweet_list do

4 twt ← tweet;

5 twt ← Preprocessing and WSD (twt)

6 Tar_twt ← Preprocessing and WSD (Tar_twt)

7 V(twt) ← fetch relation of concepts (twt) from ontology by SPARQL

8 V(Tar_twt)← fetch relation of concepts (Tar_twt) from ontology by SPARQL

9 Sim (V(twt), V(Tar_twt)) ← Calculate the Edge-counting and feature-based measure (V(twt), V(Tar_twt))

10 Sim(V(Tar_twt), V(twt)) ← Calculate the Edge-counting feature-based measure (V(Tar_twt), V(twt))

11 Similart (twt, Tar_twt) ← Sim (V(twt), V(Tar_twt)) **dot prodect** (V(Tar_twt), V(twt))

12 Final_Sim (twt, Tar_twt) ← α (sim(Edge-counting)) + β(sim(Synonyms))+ £ (sim(definition)+ γ(sim(Antonym)+δ(sim(Has-part)+ η(sim(Derivationally))

13 Spread=Number of Similar Tweets to Target Tweet / Total Number of Tweets

14 ENDFOR

15 END

CHAPTER FOUR

Implementation and Results

CHAPTER FOUR
IMPLEMENTATION AND RESULTS

4.1 Introduction

This chapter introduces the experimental results of the proposed system. Various experiments are performed for the proposed techniques, starting from the Knowledge source learning (COVID-19 Ontology learning) steps, extracting influential users and their tweets and using semantic similarity to find the rate of spread of specific information that related to COVID-19 and finally evaluate each step in our proposed system.

4.2 Datasets Description

In this research, have been used three type of dataset. One of them to finding the spread of information, and the other for system evaluation.

4.2.1 The Dataset Used

The data is collected from Kaggle[118]. The Tweets that are set From March 29 to April 30, 2020. The data consist of tweets of users who used the following hashtags: #coronavirus, #corona virus outbreak, #corona virus Pandemic, #Covid19, #Covid_19, #epitwitter, #i have corona, #Stay Home Stay Safe, #Test Trace Isolate. it's consists of about 8,576,636 million tweets posted throughout this period.

After eliminating the non-English tweets, a total of 3,584,764 million Tweets remained. The collected data contain 22 features for each Tweet,

such as user_id, text, followers, and friends, Lang, retweet, replay, etc. The dataset can be found online on the following website:https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april, the dataset features description as shown in Table (4.1).

Table (4.1) the description of Twitter dataset features

No.	Feature Name	Description
1.	status_id	The ID of the actual Tweet.
2.	user_id	The ID of the user account that Tweeted.
3.	created_at	The date and time of the Tweet.
4.	screen_name	The screen name of the account that Tweeted.
5.	text	The text of the Tweet.
6.	source	The type of app used.
7.	reply_to_status_id	The ID of the Tweet to which this was a reply.
8.	reply_to_user_id	The ID of the user to whom this Tweet was a reply.
9.	reply_to_screen_name	The screen name of the user to whom this Tweet was a reply.
10.	is_quote	Whether this Tweet is a quote of another Tweet.
11.	is_retweet	Whether this Tweet is a retweet.
12.	favourites_count	The number of favourites this Tweet has received.
13.	retweet_count	The number of times this Tweet has been retweeted.
14.	country_code	The country code of the account that Tweeted.
15.	place_full_name	The name of the place of the account that Tweeted.
16.	place_type	A description of the type of place corresponding with place_full_name

17.	followers_count	The number of followers of the account that Tweeted.
18.	friends_count	The number of friends of the account that Tweeted.
19.	account_lang	The language of the account that Tweeted
20.	account_created_at	The date and time that the account that Tweeted was created.

4.2.2 Semantic Evaluation Datasets

The proposed system uses two common datasets used to evaluate the competence of algorithms of semantic similarity. The first dataset consists of word pairs and the other includes sentence pairs with its standard similarity values. Semantic similarity efficiency is measured by calculate the correlation between the achieved results and these dataset measures value, the two dataset is:

[1] Goodenough and Rubenstein (G&R)[62]: This data was generated by employing 51 university students whose English is a native language and giving to them a form containing 65 pairs of English nouns, and asked them to rate the degree of similarity between these pairs through a scale ranging from (0) to (4), where the value (0) means that, the words are completely different, but (4) means the two words are identical.

[2] SimLex-999 from UFS Dataset[119], that 900 pairs were similar and 99 not similar but related .where 500 English speakers (native language) persons, they were employed to assess the similarity between word pairs, and the adopted scale was

between zero and six, where the two words are identical in value six. This data contains 666 pairs of nouns, 222 pairs of verbs and 111 pairs of adjectives.

4.3 COVID-19 Ontology Constructor Result

The system implement six stage to generate the COVID-19 Ontology.

4.3.1 Preprocessing Tweets

The tweets contain a lot of noise and punctuation causing ambiguity in the processing of natural language applications and extracting the important concept. Therefore, preprocessing process steps are necessary for getting the best results.

The first step in generating ontology is preprocessing the dataset, the algorithm (3.1, 3.2) was used to remove the hashtag symbols, @users, and link URLs from the tweets, after which the non-English words are deleted because the present study focuses build an ontology for English tweets only. Any special characters, abbreviations, punctuation marks and stop-words are also removed because these do not contribute to ontology learning. Table 4.2 shows a set of tweets after they have been clean.

Table 4.2 Sample of preprocessing

No	User ID	Tweets	Clean Tweets
1.	577	A snapshot **of** inequality **in the** US**, with** stats **to** prove **it.** I'm sure **the** situation **is** similar **in many** countries**. #**coronavirus **#**COVID19 **https://t.co/kcTVkW0HsO**	snapshot inequality us stats prove im sure situation similar many country coronavirus covid

2.	581	**For** #USA **#**Economic Strength data show**s that** locali**ties that** take stronger immediate **#**lockdown measures fare far **better over** time. **There's #**MAGA destroying the US **again on #**covid19 **where** SKorea **will** grow **$tronger while we** literally die. Shame **on the @**GOP **https://t.co/l7qM6oqC7D**	usa economic strength data show locality take strong immediate lockdown measure fare far good time there maga destroy us covid skorea grow tronger literally die shame gop
3.	614	Hereâ€™s a link to #Utahâ€™s #coronavirus dashboard. https://t.co/M9nJUD82OT	link coronavirus dashboard
4.	874	One of my biggest #COVID19 fears since all this started. Iâ€™m from NC, I know how things are in the South. I have a lot of family in NC & in NY. I donâ€™t talk to them that much besides my son whoâ€™s 9. We FaceTime weekly. Not a day goes by I donâ€™t worry for all of them getting sick https://t.co/w7GnZ6qseP	one big covid fear since start nc know thing south lot family nc amp ny talk much besides son facetime weekly day go worry get sick
5.	1081	way before other journalists knew what #Covid19 was, @STATnews ran an article Jan 4th about a â€œgrowing cluster of unexplained pneumonia casesâ€ in China that could be a new #coronavirus https://t.co/y6G4zRINyL	way journalist know covid statnews run article jan th growing cluster unexplained pneumonia china could new coronavirus
6.	1081	yeah, but itâ€™s super unclear how an early-warning #coronavirus detection program in #Wuhan would have been helped in detecting any early warnings about a #coronavirus coming out of #Wuhan.oh... wait.	yeah it super unclear early warning coronavirus detection program wuhan would help detect early warning coronavirus come wuhan oh wait
7.	1153	A tale of two Republicans concerned about numbers during this pandemic thatâ€™s killing the people they swore to protect. #COVID19 #coronavirus https://t.co/YKyHbGLljK	tale two republican concerned number pandemic kill people swear protect covid coronavirus
8.	1186	@Digidave @kouroshbehnam @brianmcc @pkafka I think what Facebook is doing with their #COVID19 resources is highly interesting in terms of acting like a publisher/aggregator. https://t.co/BACTFiqIDl	digidave kouroshbehnam brianmcc pkafka think facebook covid resource highly interest term act like publisher aggregator
9.	1186	@MattHartman @jordanrcrook @joshelman @saraheadler @gregisenberg @jmj Friendly reminder: still aggregating all #coronavirus-related products from @producthunt here: https://t.co/aMK4CzIGeC	matthartman jordanrcrook joshelman saraheadler gregisenberg jmj friendly reminder still aggregate coronavirus relate product producthunt

4.3.2 Domain Terms Extraction Result

The proposed system used a WSD algorithm to identify the intentional meaning of each word in the tweets, in which it extracts the best meaning representation of a word considering the influence of its immediately neighboring words. The results were used later for terms extraction.

To extract the single terms that refer to important concepts in the COVID-19 domain, the TFIDF algorithm was used that assigns a score to each term. The algorithm depends on the premise that a term that is repeated in a group of domain-specific tweets refers to the concept which belongs to this domain.

After applying the algorithm, found 718 single-word terms as shown in Table (4.3). Where the term with high frequency means a High weight in TF-IDF. In addition, we extracted a single 122 term from a medical knowledge source.

Table 4.3 Domain Terms Extraction result

Source	single-word terms	multi-word terms
From tweets	718	211
From medical resource	122	152
Total	**840**	**363**
Total terms	**1203**	

To find the terms that consist of more than one word, the proposed system uses the C-value/NC-value measure, which is an important measure for finding multi-word terms. The measure does not depend only on the repetition of the term, but also on its overlap with each other, also uses contextual clues which are powerful signals of the sequence of some words.

After examining the dataset by the algorithm, its extracted (211) multi-word terms, as shown in Table (4.3). In addition, Extracted 152 multi-word terms from a medical knowledge source.

4.3.3 Extract Concept Definition and Synonyms

In this part, the definition and synonyms of concepts were extracted by using two semantic dictionaries (babilNet, WordNet).

Where the Concept is generated by adding uniquely a textual definition called gloss with their synonyms extracted from the lexical knowledge bases. As shown in table 4.4.

Table 4.4 Extract concept definition and synonyms

NO.	Concept	Definition	Synonyms
1-	Die	pass from physical life and lose all bodily attributes and functions necessary to sustain life	die, decease, perish, go, exit, pass away, expire, pass, kick the bucket, cash in one's chips, buy the farm, conk
2-	COVID-19	a viral disease caused by SARS-CoV-2 that caused a global pandemic in 2020	corona,COVID-9, coronavirus disease, coronavirusdisease2019,2019-nCoVacute, respiratory disease, novel coronavirus pneumonia.

3-	Viral infectious disease	A disease by infectious agent that results in infection has_material_basis_in Viruses.	viral_infectious_disease
4-	disease	an impairment of health or a condition of abnormal functioning	disease
5-	crisis	an unstable situation of extreme danger or difficulty	crisis
6-	symptom	A symptom is a perceived change in function, sensation, loss, disturbance or appearance reported by a patient indicative of a disease.	symptom
7-	Cardiovascular system symptom	Symptoms of cardiovascular system developed based on type of heart disease	cardiovascular_system_symptom
8-	Psychological state	(psychology) a mental condition in which the qualities of a state are relatively constant even though the state itself may be dynamic	psychological state, psychological condition, mental state, mental condition
9-	infection	the pathological state resulting from the invasion of the body by pathogenic microorganisms	infection
10-	injury	any physical damage to the body caused by violence or accident or fracture etc	injury, hurt, harm, trauma
11-	Face mask	a covering to disguise or conceal the face	facemask

4.3.4 Taxonomic and Non Taxonomic Relation Construction

Taxonomy is an important part of ontology and the hierarchy relations are relations that provide graph visions of the ontology and determine the inheriting between concepts.

The proposed system constructs the Hypernyms relation as a triple (concept, Is-a, concept) to generate the graph. By relying on hierarchical relationships in WordNet that is like a thesaurus than a dictionary because it organizes lexical information in terms of word meanings instead of word forms.

After that, the resulted graph was compressed by removing parent nodes that have less than two children to find the final hierarchical graph, where (1202) hierarchical relation was built as shown in table 4.5 and figure 4.1 shows a section of the ontology structure

Table 4.5 Taxonomic and Non Taxonomic Relation

	SubCalssof	**Has-antonym**	**Has-derivationally**	**Has-part**	**Has-synonym**	**Has-definition**
	1202	65	1037	653	4267	1203
Total	**1202**	**7224**				
	Total hierarchical Relation	Total non-hierarchical Relation				

In addition, non-hierarchical relationships were extracted, such as antonym, meronymy, derivationally; Where API has been used to collect the above relation from BabelNet and WordNet. As it appears from the table 4.5 that most of the concepts have one formal definition and more than one synonym, in addition to the derivation relationship.

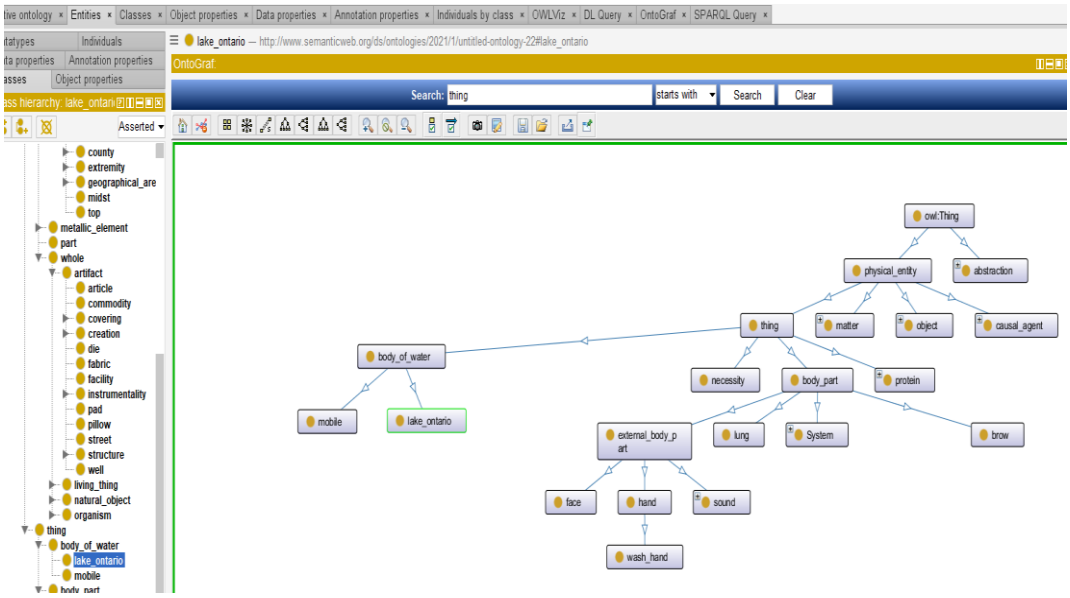

Fig 4.1 part of Ontology Graph

4.3.5 COVID-19 OWL Ontology Construction

After the important terms in the COVID-19 domain were extracted and the hierarchical and non-hierarchical relationships were built for them from the previous steps as shown in Table 4.6, the Protégé uses this data to build ontology by creating the classes, properties and constraints.

Table 4.6 sample of extracted data that used to build ontology

class	Sub-ClassOf	Has-definition	Has-synonym	Has-antonym	Has-derivationally	Has-Part
infection	ill_health	A pathological condition resulting by the infestationof pathogenic	infection, contagion, transmission	----	'infectious,infectious,infect'	'incubation

		microorganisms into the body.				
hospital	building	a health facility where patients receive treatment'	hospital,infirmary	---	hospitalize	burn_center,clinic,hospital_room,ward'
doctor	health_professional	a licensed medical practitioner	doctor,doc,physician,MD,Dr.,medico	---	doctorial,doctor	-----
death	happening	the event of dying or departure from life	death,decease,expiry	brith	Die,decease	
mask	covering	a covering to disguise or conceal the face	mask	----	mask	

The process starts automatically by building the OWL profile, then building the main root, and adding other classes under the root. In addition, it adds the formal definition and the hierarchical and non-hierarchical relationships to classes from the aforementioned table.

After the completion of the ontology building, the HermiT reasoner is activated and run the debugging to ensure that the ontology is correct, coherent and consistent as show in figure 4.2.

Figure 4.3 displays part of the resulting ontology file that includes the definition of the ontology and part of the classes. As for the number of axiom, Class, Data property, Annotation and Annotation Assertion are explained in table 4.7.

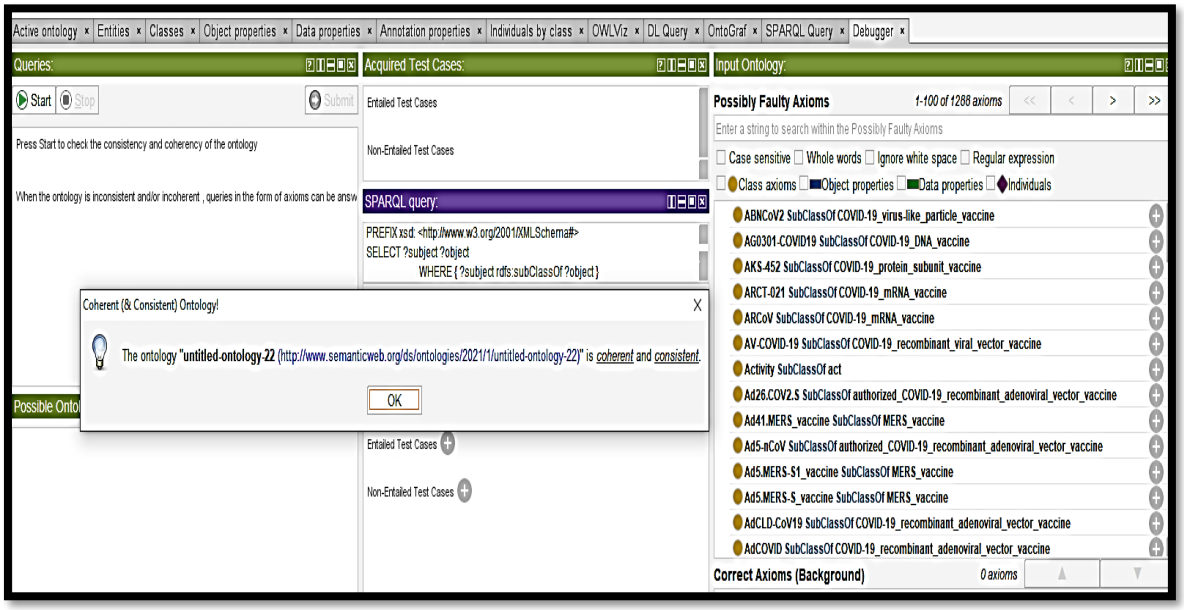

Fig 4.2 ontology debugging

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22"
    xml:base="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xml="http://www.w3.org/XML/1998/namespace"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:rffs="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#rffs:">
   <owl:Ontology rdf:about="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22">
      <owl:imports rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22"/>
   </owl:Ontology>

http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22# infection-->
<owl:Class rdf:about="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#infection">
      <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#ill_health"/>
      <rdfs:antonym></rdfs:antonym>
      <rdfs:comment>the pathological state resulting from the invasion of the body by pathogenic
microorganisms</rdfs:comment>
      <rdfs:derivationally>infectious,infectious,infect</rdfs:derivationally>
      <rdfs:has_part>incubation</rdfs:has_part>
      <rdfs:label>infection</rdfs:label>
   </owl:Class>

http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#hospital -->
   <owl:Class rdf:about="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#hospital">
      <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#building"/>
      <rdfs:antonym></rdfs:antonym>
      <rdfs:comment>a health facility where patients receive treatment</rdfs:comment>
      <rdfs:derivationally>hospitalize</rdfs:derivationally>
      <rdfs:has_part>burn_center,clinic,hospital_room,ward</rdfs:has_part>
      <rdfs:label>hospital,infirmary</rdfs:label>
   </owl:Class>

http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#doctor -->
   <owl:Class rdf:about="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#doctor">
```

```
      <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-
22#health_professional"/>
      <owl:disjointWith rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#nurse"/>
      <rdfs:antonym></rdfs:antonym>
      <rdfs:comment>a licensed medical practitioner</rdfs:comment>
      <rdfs:derivationally>doctorial,doctor</rdfs:derivationally>
      <rdfs:has_part></rdfs:has_part>
      <rdfs:label>doctor,doc,physician,MD,Dr.,medico</rdfs:label>
   </owl:Class>

http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#death -->
   <owl:Class rdf:about="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#death">
      <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#happening"/>
      <owl:disjointWith rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-
22#physical_phenomenon"/>
      <rdfs:antonym>brith</rdfs:antonym>
      <rdfs:comment>the event of dying or departure from life</rdfs:comment>
      <rdfs:derivationally>die,decease</rdfs:derivationally>
      <rdfs:has_part></rdfs:has_part>
      <rdfs:hasinstance></rdfs:hasinstance>
      <rdfs:label>death,decease,expiry</rdfs:label>
      <rdfs:sisterterm>change,alteration,modification</rdfs:sisterterm>
   </owl:Class>

http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#mask -->
   <owl:Class rdf:about="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#mask">
      <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-22#covering"/>
      <owl:disjointWith rdf:resource="http://www.semanticweb.org/ds/ontologies/2021/1/untitled-ontology-
22#protective_covering"/>
      <rdfs:antonym></rdfs:antonym>
      <rdfs:comment>a covering to disguise or conceal the face</rdfs:comment>
      <rdfs:derivationally>mask</rdfs:derivationally>
      <rdfs:has_part></rdfs:has_part>        <rdfs:label>mask</rdfs:label>
   </owl:Class>
```

Fig 4.3 part of Ontology Graph

Table 4.7 OWL Ontology Metrics

Ontology Metrics	Value
Axiom	7564
Logical axiom count	1288
Declaration axioms count	1210
Class count	1203
Data property count	2
Annotation Property count	8
Disjoint Classes	85
Annotation Assertion	5066

Where the axioms represent axiom count that contain the number of logical axioms, the number of declaration axioms, and Annotation Assertion. While the "annotation property count" represent the number of relationships generated in the ontology file.

4.4 Users Influencer Detector Result

The proposed system depends on extracting the messages of influential users that are responsible for amplifying and disseminating information on social networks. By identifying these users, the volume of processed data is reduce and the spread detection speed is increase.

The proposed system for finding the influential users depends on extracting the values of eight features for each user by using Algorithm 3.5 and the result is show in table 4.8.

Table (4.8): example of User feature measure data

users_id	User name	Rf	Rrm	Urmo	Ff	Nt	Nu_like	NeRepl_Rt	NeT
25365536	KimKardashian	496283.1154	282	282	64516805	1	165	0	0
50393960	BillGates	227720.3134	1693	3386	49415308	2	260	0	0
742143	BBCWorld	387953.6761	181	181	27544711	1	13	0	0
18681139	priyankachopra	47723.30926	466.5	933	25770587	2	2846	707	1
1.65E+06	Reuters	19444.20736	69	69	21660847	1	750	0	0
500704345	Pontifex	2283906.125	2282	2282	18271249	1	0	0	0
2.49E+07	ParisHilton	1951.110141	158	158	16988316	1	17439	158	1
14293310	TIME	34176.34343	12	24	16917290	2	1070	11	1
94163409	sonakshisinha	214874.3056	100	100	15470950	1	1076	100	1
56783491	nickjonas	22044.72205	240	240	14593606	1	6771	0	0
1115874631	CGTNOfficial	250652.4643	29.06885246	8866	14036538	305	31400	4683	150
8.19E+17	FLOTUS	1.95E+06	7226.5	43359	13656984	6	3	20363	3
18228898	johnlegend	11470.24209	152	152	13408713	1	25190	0	0
140070953	FIFAcom	15768.41538	13.75	55	13324311	4	6456	18	1

207809313	BJP4India	4337796.333	426	426	13013389	1	0	0	0
487118986	XHNews	197877.0469	33.14859438	8254	12664131	249	7630	3475	122
134758540	timesofindia	23724.4566	15.33044983	8861	12573962	578	53847	3956	300
37034483	ndtv	832180.8	26.76623377	20610	12482712	770	82637	9111	405
14159148	UN	10358.76988	157.3555556	7081	12244066	45	63249	2818	20
68977380	thekiranbedi	41316.84459	251	502	12229786	2	19416	132	1
1.85E+08	BCCI	146861.026	284	284	11308299	1	29	284	1
1.34E+06	WIRED	27000.40933	54.15	1083	10422158	20	65333	552	8
5.80E+07	Riteishd	33683.69381	272	272	10340894	1	1118	272	1
972651	mashable	3520.877023	20.5	41	9791559	2	2732	12	1
6017542	BreakingNews	16916.10873	12	12	9489937	1	6	0	0
240649814	TimesNow	23916.65649	7.166666667	43	9399246	6	27	11	2
39240673	ABPNews	37616.46774	18.83673469	1846	9328884	98	10563	594	42
56304605	sardesairajdeep	15908.21429	339	339	8908600	1	7364	0	0
87818409	guardian	7914.020314	58.33333333	175	8570884	3	415	103	2

Due to each feature having a different impact on users, we calculated the weight of each feature by using CRITIC algorithm 2.3, and the result is show in table 4.9.

Table (4.9) weight of features

Feature	**Rf**	**Rat_nurtrep_Tt**	**Nu_Retweet+Nu_rep**	**followers_count**
Weight	0.11950	0.21758	0.09355	0.15043
Feature	**Ttweet**	**Nu_like**	**new_Rt_Rpl**	**New_Tweet**
Weight	0.12973	0.10037	0.10392	0.08488

We note from the table of weights that the scale ‘Rat_nurtrep_Tt’ and ‘followers_count’ had the greatest influence in calculating the effect of the user, while the scale ‘New_Tweet’ was the least powerful.

Table (4.10) weight of users					
		Kim Kardashian		**BillGates**	
Feature	**Weight feature**	Feature value	Weight of Each feature	Feature value	Weight of Each feature
Rf	0.11950	496283.115	59309.68	227720.31	27214.343
Rat_nurtrep_Tt	0.21758	282	61.35907	1693	368.37199
Nu_Ret+Nu_rep	0.09355	282	26.38156	3386	316.76581
friends_count	0.15043	64516805	9705421	49415308	7433665.6
Ttweet	0.12973	1	0.129735	2	0.2594708
new_Rt_Rpl	0.10037	165	16.56179	260	26.097367
New_Tweet	0.10392	0	0	0	0
Nu_like	0.08488	0	0	0	0
Total user weight			**9764834.8**		**7461591.4**

After that, the weight calculated for each user based on the weights of those properties, through multiply the weight of the feature by the value of the feature to find the weight of each feature for the user, then sums the weights of features for each user to find the final weight of the users. Table 4.10 shows a sample of them.

The users are arranges according to weight to find the influential users who have a weight higher than zero as shown in Figure (4.4).

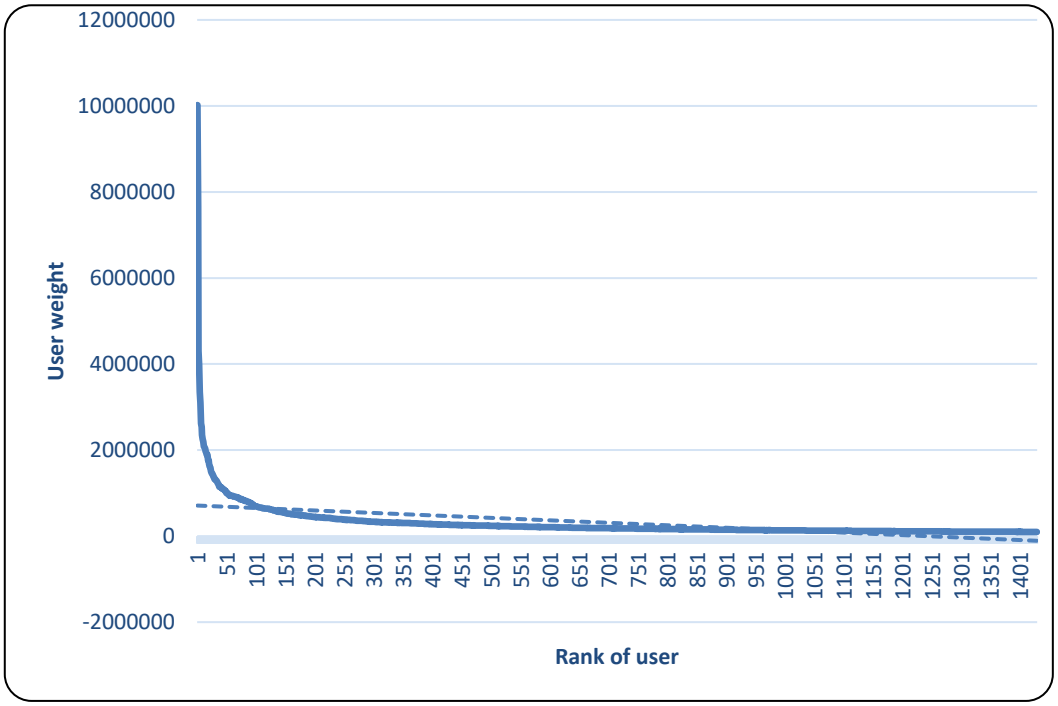

Fig. 4.4 User's influential Weight Value

Table (4.11) shows the results of the experiment in finding the influential users and their tweets that will used in next stage.

Table 4.11 Influential Users result

Total uses	Total Tweets	Influential users	Non Influential users	Number of accepted tweet	Number of reject tweet
581653	992219	346942	234709	709870	282349
Percentage		60 %	40 %	72%	28%

Table (4.11) shows the results of the experiment in finding the influential users and their tweets .Where the results show that the volume of processed data has been reduce to 72% .which will be used in next stage.

4.5 Semantic Web Similarity Comparative Model Result

To calculate the semantic similarity between the target message and the rest of the messages based on ontology. The system relied on the use of two types of semantic similarity measures. The first one based on finding the hierarchical similarity by comparing the location of the two concepts within ontology by using Wu metric. The second based on finding the similarity based on the formal definition and non-hierarchical relationships of concepts by implement cosine similarity metric as explain in section 3.5.3.

The system compute six-similarity value based on number of features that extracted from the ontology, between target tweet and the other tweet as show in result in table 4.12.

Table 4.12 Result semantic similarity for each metric

Target Tweet	Rest Tweets	Edge Met.	Has Definition	Has syn	Has part	Has antonym	Has_derivationally
number of covid-19 deaths and infection conditions increased in the world	institution close due increase number corona infection death world	0.7056	0.7461	0.5614	0.8819	0.2222	0.95
	impressive modern tool assistant hospital general populace commodity understand fix COVID-19 answer	0.3563	0.2263	0.1060	0.7977	0	0
	COVID-19 billion infection case number death COVID-19 exceed COVID-19	0.7973	0.7102	0.5443	0.8819	0.1111	0.95
	Report million case COVID-19 death discovery vaccine COVID-19	0.5762	0.4781	0.3061	0.9354	0.125	0.9499
	movie architectural wonder area light pay tribute health worker area battle global pandemic.	0.2744	0.1650	0.0208	0.7637	0	0
	letter send million family across unify kingdom beginning week	0.3286	0.1750	0	0.8366	0	0

number of covid-19 deaths and infection conditions increased in the world	far increase death COVID-19 confirm lawsuit india include death last hour new case death report	0.5048	0.4033	0.3075	0.9354	0.125	0.9499
	reality government reaction COVID-19 gun_trigger statistically bad economic collapse history upstate modern york	0.3182	0.2266	0.0769	0.7337	0	0
	bad clock_time social distance nice get together friend about watch joke COVID-19 lockdown	0.3152	0.2191	0.0769	0.7337	0	0
	wide margin error public body motivation crystalline COVID-19	0.3146	0.2161	0.125	0.9354	0	0
	think goofy man screen become report hang regretful apology COVID-19	0.2947	0.2583	0.1111	0.8819	0	0
	modern york statewide COVID-19 case march comparison madison square garden field_hockey capacity statewide case march comparison yankee stadium capacity	0.3756	0.1880	0.0526	0.6069	0	0
	modern virus test detect disease five minute	0.37932	0.2167	0	0.95	0	0
	perspective approximately billion people global COVID-19 death bell COVID-19 exceed COVID-19	0.5995	0.4392	0.2811	0.7977	0.0909	0.95
	cattiness tweet early assistant share issue COVID-19 technical_school	0.3923	0.2330	0.1428	0.1	0	0
	mind test positive COVID-19 plastic kit allege test damaging mind damn positive positive plague laid_low remove heavy price mind	0.3184	0.3460	0.0666	0.6831	0	0
	watch news tonight remember virus die die soon deputy head checkup military_officer order possibly calendar_month die united_kingdom alone.s.01 COVID-19	0.3127	0.1881	0.0537	0.6069	0	0.1028

Due to the effect of the previously mentioned measures on semantic similarity is different, therefore, the system calculated the weight of each

measure by using the entropy that explain in section 3.6.4 ,and the result in Table 4.13 .

Table 4.13 metrics weight

Metric	Weight
Graph metric (α)	0.35047
Has_Definition metric (β)	0.24040
Has_Synonyms metric (£)	0.23071
Has-part metric (δ)	0.01719
Has_Antonym metric (γ)	0.00123
Has_Derivationally metric (η)	0.15887

Through Table 4.13, it found that the edge-count metric has the highest effect on the similarity value while the lowest effect is the Has_Antonym metric as shown in the figure 4.5.

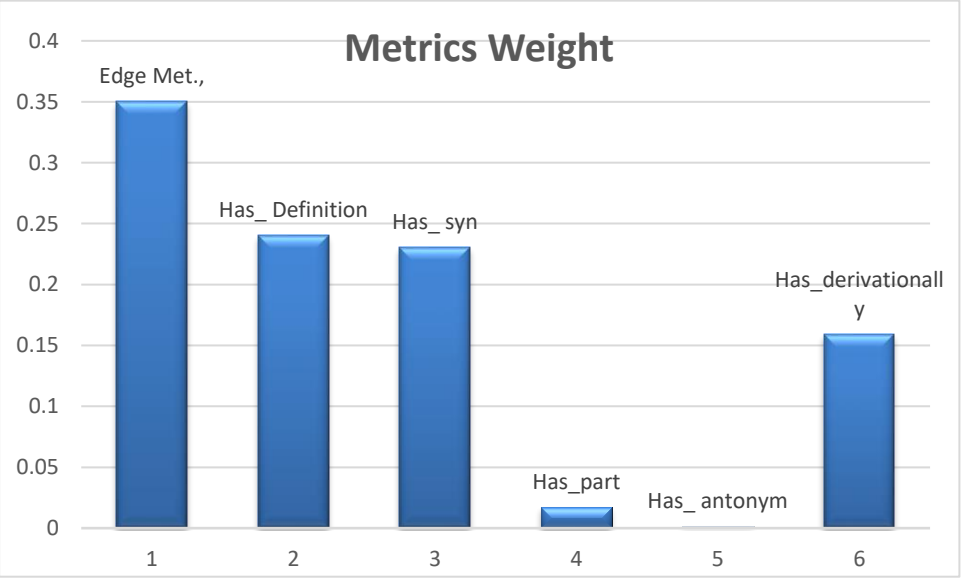

Fig. 4.5 Metrics Weight Value

Finally, the system computes the final semantic similarity between the target tweet and another tweet by multiplying the weight of the scale by its value as mention in section 3.6.5 and then sum the results of the six metric as shown in the table 4.14.

Table 4.14 final semantic similarity

Target tweet	Rest Tweet	Semantic Similarity value
The number of covid-19 deaths and infection conditions increased In the world.	Institutions closed due to the increase in the number of corona infections and deaths in the world	0.72259
	This impressive new tool from @IHME_UW will help hospitals, policymakers, and the general public better understand and prepare for the #COVID19 response in the U.S. https://t.co/tTd5tEtTRc	0.21750
	Coronavirus: #leadingthroughchange 1.3 Billion infection case .https://t.co/H4U8AYEaID"Coronavirus: number of Death from COVID-19 exceeds 30,000 #Coronavirus https://t.co/PXPw8ZjSPo	0.74203
	WHO reported about 1.5 million cases of COVID-19 and 30k deaths . WHO will find a vaccine for covid-19.	0.55473
	In pics | Architectural marvels around the world have lit up t o pay tribute to the health workers in their countries who are battling the global pandemic	0.15382
	The letter will be sent to 30 million households across the Un ited Kingdom starting from next week @BorisJohnson#Coron avirusOutbreak	0.17167
	So far, there has been increase in death where 979 COVID19 confirmed cases in #India, including 25 deaths. In the last 24 hours, 106 new cases 6 deaths have been reported #coronavirusindia https://t.co/qVD4TbGI2h	0.52931
	The reality is that government reaction to the #coronavirus h as triggered what is statistically the worst economic collapse i n the history of upstate New York. https://t.co/bJOeKr0AGB	0.19638
	During these tough times of social distance, itâ€™s nice it get together with friends virtually, watching @TheRoom and laug hing. #Covid_19 #lockdown #StayAtHomeAndStaySafe https: //t.co/VQcCMfQva	0.19353
	Thatâ€™s a wide margin of error. Public bodies need to be m ore transparent. #cdnfoi #cdnpoli #coronavirus https://t.co/J pOzCA1M0f	0.20716

the number of covid-19 deaths and infection conditions increased in the world.	ðŸŽ¶You might think itâ€™s goofy, but the man on the screen is a Newfie,And heâ€™s become the story, after hanging up, is he sorry?ðŸŽ¶ (Apologies to Stompin' TomðŸ¥”) #cdnpoli #coronavirus https://t.co/zUvOiIbt01	0.20618
	New York statewide #coronavirus cases as of March 22: 15,168 (compare with Madison Square Garden hockey capacity 18,006).Statewide cases as of March 29: 59,513 (compare with Yankee Stadium capacity 54,251	0.19946
	This new virus test can detect the disease in five minutes. https://t.co/bYggEkf8SL https://t.co/9HFjqHnS1T	0.20223
	Perspective: @realDonaldTrump there are approximately 7.9 billion people worldwide. Coronavirus: Death toll from COVID-19 exceeds 30,000 #Coronavirus https://t.co/mbppUtMtv1	0.54533
	In spite of what I tweeted earlier. I couldnâ€™t help sharing this emerging #coronavirus tech. https://t.co/fiP1F3lNYM https://t.co/BcdFO8qDv5	0.24367
	My Mind Tested Positive#Covid_19 The plastic kit says my test is negative, but in mind I am damn positive that it is positive. The plague has not only stricken me, but it has taken a heavy toll on my mind	0.22194

4.6 Model Evaluation

Our model pass through three-evaluation stage that explains in figure 4.6.

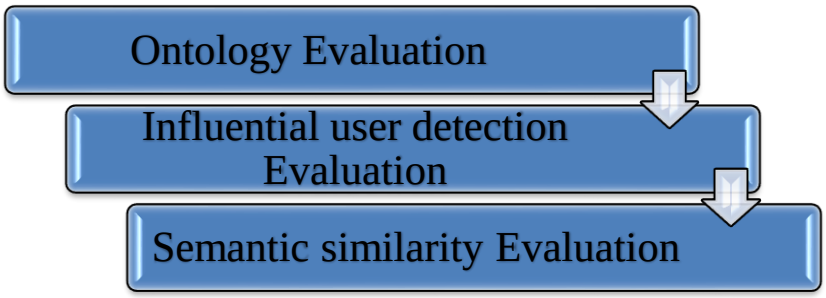

Fig. 4.6 Model Evaluation

4.6.1 COVID-19 Ontology Constructor Evaluation

The ontology evaluation is an important part of learning ontology. It is considered a challenging task as there is no standard evaluation available, as in Retrieval of Information where the precision and recall measure is used. The reason can be trace back to the nature of unsupervised of ontology build method, which makes the evaluation harder than supervised learning like classification. Therefore, the approaches of Ontology evaluation can be classify based on several factors like comparing ontologies, usage and application of ontologies, or human evaluation, in order to meet ontology requirements and compatibilities.

The present work is based on human evaluation by using OntoMetrics [120] which is a web-based tool product by Germany Rostock University that validates and displays statistics about a given ontology. The ontology is evaluate through two schema metrics as show in table (4.15).

1- Inheritance Richness:

It is a metric that describes the information distribution of information over various levels of the inheritage tree of ontology or the prevalence of parental classes. It is a proper indication of how good knowledge is classified into various classes and ontology subcategories. This scale is distinguished by its ability to distinguish the horizontal ontology, for example that color contains a large number of direct branches (red, green, blue, etc.), and while in the vertical ontology it contains a small number of direct branches. An ontologist who has lower inheritance richness will be deep or vertical, and this indicates that the ontologist specializes in a specific domain and contains details related to this domain. This resembles what appears in the results of the present work, as show in table 4.15. On the other hand, the ontology that shows a high value indicates that it is general and not specialized within a specific domain with less detail.

Table 4.15 Ontology Evaluation

Base metrics	**Value**	**Schema metrics**	**Value**
Axioms:	7564	Inheritance richness rang (0-10)	0.999169
Logical axioms count:	1288	Axiom/class ratio rang (0-10)	6.287614
Class count:	1203	Annotation assertion axioms count	5066
Total classes count:	1203		
Object property count:	2		
Total object properties count:	2		
Data property count:	0		

2- Axiom and Axiom/class ratio:

This represents the main component in ontology and they state what is true in a domain. Generated ontology has high value for this measure, which indicates that the results are correct as show in table 4.15.

4.6.2 Users Influencer Detector Evaluation

Because of the publicly available Twitter API, there are several ranking services that on the one hand, calculate the score for individual users, and on the other hand, compare scores of twitter users to create a ranking.

A common and widely used online classification service is SparkScore [121]which determines user performance on Twitter through some features such as Tweets, Following, Followers, Engagement, Avg Likes Per Tweet, Avg Retweets Per, Tweets are retweets as shown in Figure 4.7.

Fig. 4.7 SparkScore dashboard

have been used this service on the dataset, with comparing the service outputs with system results, where the results appeared close ,where the higher score in SparkScore meet rank one in system result as explain in the table (4.16), also the value of correlation between system result and SparkScore value appear high correlation which equal 0.971 as in figure 4.8.

The results also showed that the degree of user influence increases as the user's activity increases in making a tweet or responding to a specific tweet, as well as the degree of user's importance increases with the user's ability to attract the attention of other users through gets a retweet and a mention for his / her tweet, in addition to the number of Following and followers for that user.

Table 4.16 comparing system results with the SparkScore results for the same data

No.	User name	proposed Algorithm	proposed Algorithm Ranking	SparkScore
1.	KimKardashian	10019504.79	1	100
2.	BillGates	7656325.395	2	100
3.	priyankachopra	3983924.531	3	100
4.	Reuters	3346121.006	4	100
5.	Pontifex	3100056.181	5	100
6.	ParisHilton	2622571.934	6	96
7.	BJP4India	2538787.889	7	98
8.	FLOTUS	2352705.397	8	98
9.	johnlegend	2076474.768	9	95
10.	FIFAcom	2058675.6	10	96

11.	ndtv	2030883.188	11	85
12.	XHNews	1979960.314	12	77
13.	timesofindia	1944882.744	13	66
14.	thekiranbedi	1892965.375	14	65
15.	UN	1892077.184	15	72
16.	WIRED	1612193.033	16	61
17.	mashable	1511865.466	17	60
18.	TimesNow	1453796.203	18	57
19.	ABPNews	1444823.31	19	59
20.	TimeOutSG	8717.234193	20	47
21.	irEnriqueCortes	8651.250357	21	33
22.	stevewil94	8649.600199	22	29
23.	pioneersaad	8645.920982	23	14
24.	simcim	8642.565237	24	13
25.	jrutmusic	2072.267127	25	24
26.	agriculturallaw	2071.076028	26	24
27.	Taltool11	566.4408576	27	10
28.	mhsutton	566.339431	28	10
29.	messagebubble	566.2624131	29	10
30.	RobertWKYT	563.1370045	30	16
31.	medtek	343.5956416	31	7
32.	EKitamirike	343.5922451	32	7
33.	TrophyClubGov	343.1892846	33	6
34.	JosephTeoSG	71.58812968	34	4
35.	TheReal_UB	28.22889918	35	1
36.	UnfilteredCraft	28.21334841	36	1
37.	loomah	5.649855775	37	1
38.	JBDMjournal	5.646493353	38	1
39.	RuchaADS	5.582768938	39	1

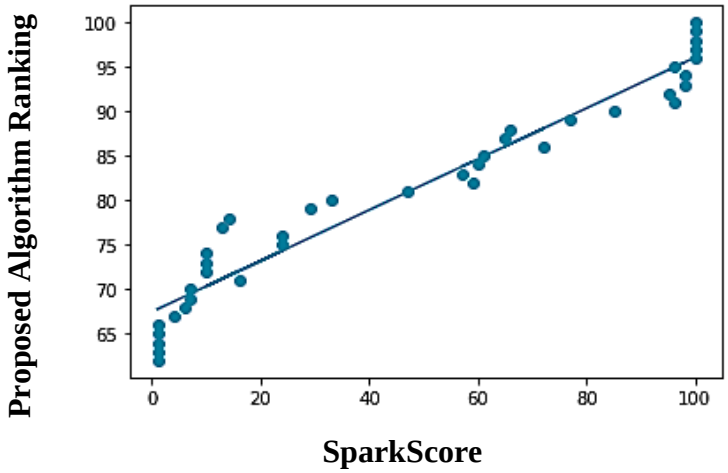

Fig. 4.8 Comparison between proposed algorithm and SparkScore metric For finding influential users

4.6.3 Semantic Web Similarity Comparative Model Evaluation

The Semantic similarity measure Evaluation between texts is one of the challenging and open research problems in the field of Natural Language Processing (NLP). Due to the Difficulties to define rule-based methods for determining semantic similarity measures. The researchers used datasets based on human judgments to evaluate results .The proposed system is being evaluated in two level, first level used a standard dataset SimLex-999 which is a gold standard resource for the evaluation of models that learn the meaning of words and concept and the second level use a standard dataset generated by Rubenstein and Goodenough (R&G) to evaluate the sentences, these datasets are mentioned in 4.2.2.

Table 4.17 comparison between proposed system result and SimLex-999 result

No.	Word1	Word2	SimLex-result	propseed system result	No.	Word1	Word2	SimLex-result	proposed system result
1.	Book	Text	6.35	5.5	32.	hand	thumb	3.88	3.5648
2.	groom	bride	3.17	3.5911	33.	mud	dirt	7.32	5
3.	night	day	1.88	1.9427	34.	way	manner	7.62	8.6601
4.	plane	airport	3.65	3.572421	35.	mouse	cat	4.12	4.946
5.	uncle	aunt	5.8	6	36.	death	burial	4.93	2.298
6.	marrow	blood	3.4	3.0620	37.	mouth	lip	6.1	4.1801
7.	student	pupil	9.35	9.98905	38.	storm	hurricane	6.38	5.8169
8.	leg	arm	2.88	4	39.	tax	income	2.38	3.050
9.	plane	jet	8.1	5.7756	40.	flower	violet	6.95	4.226
10.	woman	man	3.33	4.42695	41.	paper	cardboard	5.38	4.5827
11.	actress	actor	7.12	6.9331	42.	rod	curtain	3.03	2.10459
12.	teacher	instructor	9.25	9.98	43.	hound	fox	2.38	2.091
13.	movie	film	8.87	8.60155	44.	street	alley	5.48	5.5875
14.	dog	cat	1.75	3.1039	45.	boat	deck	4.28	3.9732
15.	area	region	9.47	9.599	46.	car	horn	2.57	2.0247
16.	navy	army	6.43	6.7826	47.	friend	guest	4.25	4.0894
17.	clothes	closet	3.27	3.525	48.	employer	employee	3.65	4.38943
18.	sunset	sunrise	4.47	5.6703	49.	hand	wrist	3.97	3.2709
19.	child	adult	2.98	0.29192	50.	door	doorway	5.4	5.8956
20.	winter	summer	4.38	4.9517	51.	winter	season	6.27	5.8312
21.	taxi	cab	9.2	9.98	52.	decade	century	3.48	3.2274
22.	tree	maple	5.53	5.2264	53.	take	leave	2.47	3.132
23.	bed	bedroom	3.4	4.1	54.	please	plead	2.98	2.1575
24.	arm	shoulder	4.85	4.609	55.	ask	plead	6.47	5.0910
25.	lady	gentleman	3.42	4.2	56.	window	door	3.33	2.7247
26.	boat	anchor	2.25	2.5534	57.	arm	wrist	3.57	2.03659
27.	toe	finger	4.68	4.6344	58.	wine	brandy	5.15	5.356
28.	river	stream	7.3	5.6777	59.	dinner	breakfast	3.33	4.9085
29.	date	calendar	4.42	3.6	60.	hose	garden	1.67	2.6492
30.	sea	ocean	8.27	9.9	61.	child	kid	9.5	9.989

31.	second	minute	4.62	4.5119	62.	parent	adult	5.37	4.53324

A set of words was taken randomly from SimLex-999 database for the purpose of evaluating the proposed system and the results are shown in the Table (4.17) a high Pearson correlation degree with human judgments equal (0.890), also our system is more line regression with human judgments, as shown in figure (4.9).

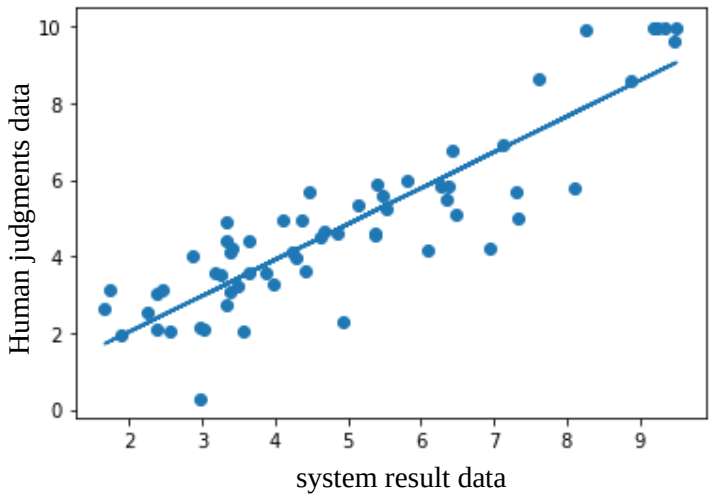

Fig. 4.9 Linear regression model Human Similarity against proposed system Word Similarity result

Also on the sentence level the results as shown in Table (4.18) appear high Pearson correlation degree for proposed algorithm with human judgments.

also, comparing with Atish system [122]that used these dataset to finding the similarity of the semantic sentences , the system results were better than Atish result ,where the degree of correlation of its results with the human evaluation was 0.739, while the system result correlation with human judgement is 0.849, as shown in figure (4.10, 4.11,4.12).

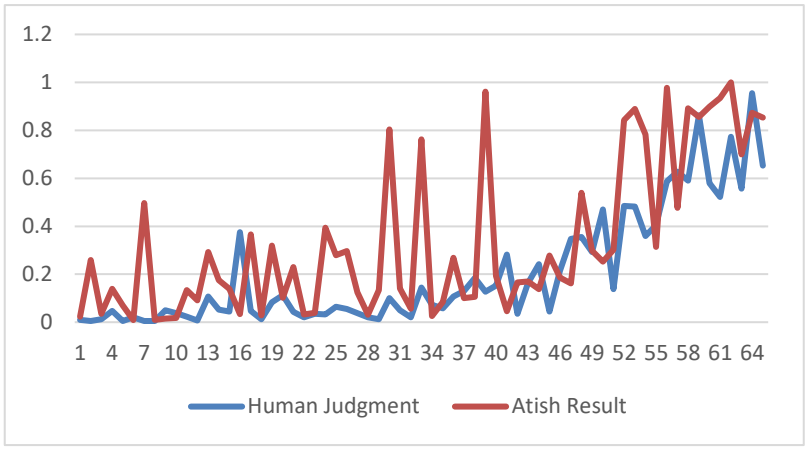

Fig. 4.10 Comparison between Human Similarity against Atish Algorithm Similarity for sentence

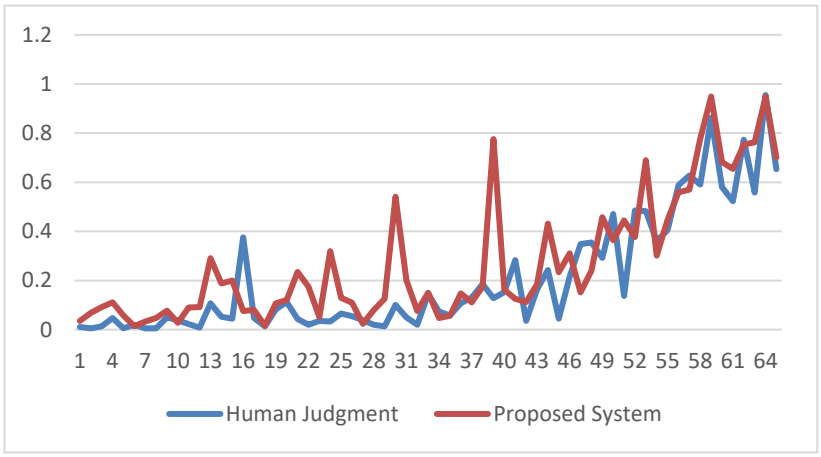

Fig. 4.11 Comparison between Human Similarity against proposed Algorithm for sentence

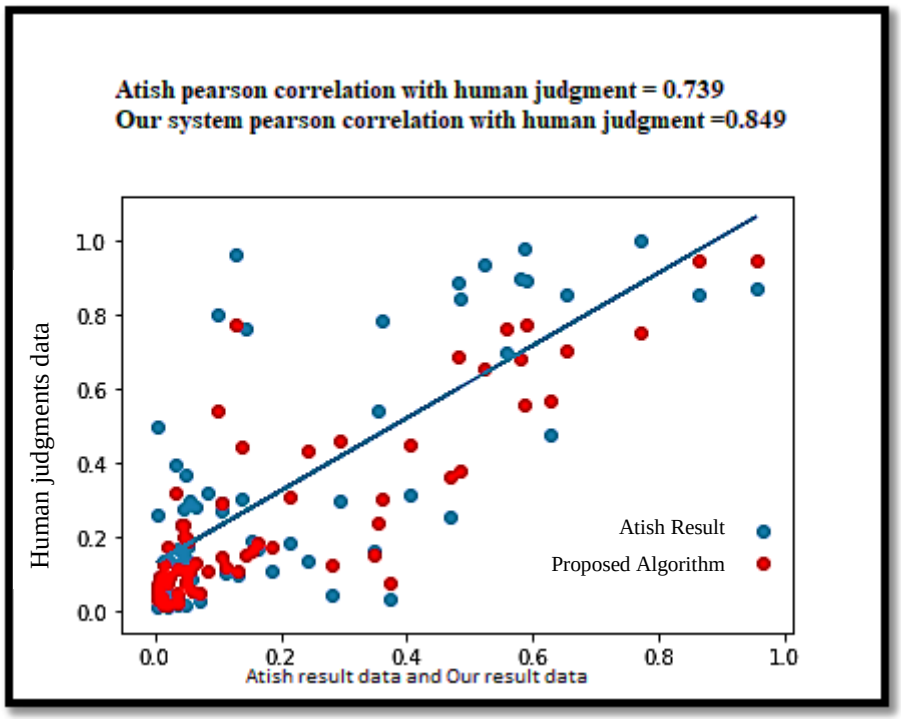

Fig. 4.12 Linear regression model Human Similarity against proposed Algorithm in red and Atish in blue Similarity

CHAPTER FIVE
CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORK

CHAPTER FIVE
CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORK

5.1 Conclusions

In the proposed system, we have examined the problem of disclosure of the dissemination of information in social networks based on ontology and suggested an approach that combines users feature and ontology aspects. The process of detecting the spread of information on Twitter by matching the words of the message which is difficult because many users use many phrases to express the same topic. Therefore, the research was based on the use of semantic metrics based on ontology to search for similarities between the target message and the rest of Twitter's messages.

According to the experiments that is performed and result obtained from this research, a number of conclusions have been obtained; some of them are:

1. The preprocessing stage improves the process of constructing Ontology and detecting information diffusion by removing unuseful words and noises in huge volumes of tweets.
2. Using the features method to find the influential users in the social network (Twitter), decreased the amount of information and increased the processing speed.
3. By using combining of machine learning, linguistic and statistical information metrics, the technique show high efficiency in extracting terms, which extract 840 single terms and 363 multi-word terms as shown in table 4.2.

4. The use of ontology-based semantic similarity techniques to find the spread of information in social networks is an efficient way, especially in the case of using different terminology to express a particular topic.
5. Using the summarize technique of the ontology graph has given a high speed in calculating the degree of similarity of a scale edge-counting.
6. The OWL is the right syntax for representing Twitter information supported via RDF and RDFS by providing many vocabularies and formal semantics.
7. Protégé editor is the most straightforward framework for ontology developers. Moreover, it is rich with several plugins for visualization, inferring, and querying, making it flexible and powerful.
8. Ontological aspects will represent a significant role in semantic web implementations in the future web.

5.2 Recommendations for Future Work

1. Building a model that can detect the spread of specific information within different languages based on multilingual ontology.
2. Building a model that can detect the spread of information within different social networks, such as Facebook and LinkedIn.
3. Building an online model combining Twitter API and the ontology that can offer a flexible and viable model for online checking spread of specific information.
4- Linking the COVID-19 ontology with Linkdata, DBpedia and other sources of knowledge to enrich the ontology with additional relationships and information.

REFERENCES

[1] "WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data." https://covid19.who.int/ (accessed Dec. 20, 2021).

[2] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLoS One*, vol. 5, no. 11, p. e14118, 2010, doi: 10.1371/JOURNAL.PONE.0014118.

[3] S. R. Rufai and C. Bunce, "World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis," *Public Heal. |*, vol. 42, no. 3, pp. 510–516, 2020, doi: 10.1093/pubmed/fdaa049.

[4] D. E. O'Leary, "Twitter Mining for Discovery, Prediction and Causality: Applications and Methodologies," *Intell. Syst. Accounting, Financ. Manag.*, vol. 22, no. 3, pp. 227–247, 2015, doi: 10.1002/isaf.1376.

[5] M. Bravo, L. F. H. Reyes, and J. A. Reyes Ortiz, "Methodology for ontology design and construction," *Contaduria y Adm.*, vol. 64, no. 4, pp. 1–24, 2019, doi: 10.22201/FCA.24488410E.2020.2368.

[6] A. Maedche and S. Staab, "Learning Ontologies for the Semantic Web," *IEEE Intell. Syst.*, vol. 16, no. 2, pp. 72–79, 2001.

[7] A. Hatua, T. T. Nguyen, and A. H. Sung, "Information diffusion on Twitter: Pattern recognition and prediction of volume, sentiment, and influence," *BDCAT 2017 - Proc. 4th IEEE/ACM Int. Conf. Big Data Comput. Appl. Technol.*, pp. 157–167, 2017, doi: 10.1145/3148055.3148078.

[8] R. Bin Tareaf, "Information Propagation Speed and Patterns in Social Networks: A Case Study Analysis of German Tweets," *J. Comput.*, vol. 13, no. 7, pp. 761–770, 2018, doi: 10.17706/jcp.13.7.761-770.

[9] J. Xu and Y. Qiang, "Analysing Information Diffusion in Natural Hazards using Retweets - a Case Study of 2018 Winter Storm Diego," *Ann. GIS*, vol. 00, no. 00, pp. 1–15, 2020, doi: 10.1080/19475683.2021.1954086.

[10] S. Kumar, M. Saini, and M. Goel, "Kumar-2020-Modeling information diffusion in o.pdf," 2020.

[11] L. Dinh and N. Parulian, " COVID -19 pandemic and information diffusion analysis on Twitter ," *Proc. Assoc. Inf. Sci. Technol.*, vol. 57, no. 1, pp. 1–10, 2020, doi: 10.1002/pra2.252.

[12] A. K. Kushwaha, A. K. Kar, and P. Vigneswara Ilavarasan, "Predicting Information Diffusion on Twitter a Deep Learning Neural Network Model Using Custom Weighted Word Features," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12066 LNCS, no. July, pp. 456–468, 2020, doi: 10.1007/978-3-030-44999-5_38.

[13] A. Kumar, D. Chhabra, B. Mendiratta, and A. Sinha, "Analyzing Information Diffusion in Ego-centric Twitter Social Network," *2020 6th Int. Conf. Signal Process. Commun. ICSC 2020*, pp. 363–368, Mar. 2020, doi: 10.1109/ICSC48311.2020.9182719.

[14] J. A. Wahid *et al.*, "Identifying and characterizing the propagation scale of covid-19 situational information on twitter: A hybrid text analytic approach," *Appl.*

Sci., vol. 11, no. 14, 2021, doi: 10.3390/app11146526.
[15] H. Hönings *et al.*, "Health information diffusion on Twitter: The content and design of WHO tweets matter," *Health Info. Libr. J.*, no. May 2020, pp. 1–14, 2021, doi: 10.1111/hir.12361.
[16] X. Han and L. Niu, "On charactering of information propagation in online social networks," *J. Networks*, vol. 8, no. 1, pp. 124–131, 2013, doi: 10.4304/jnw.8.1.124-131.
[17] Z. Zhao, Y. Liu, and K. Wang, "An analysis of rumor propagation based on propagation force," *Phys. A Stat. Mech. its Appl.*, vol. 443, pp. 263–271, 2016, doi: https://doi.org/10.1016/j.physa.2015.09.060.
[18] D. Li, J. Ma, Z. Tian, and H. Zhu, "An evolutionary game for the diffusion of rumor in complex networks," *Phys. A Stat. Mech. its Appl.*, vol. 433, pp. 51–58, 2015, doi: https://doi.org/10.1016/j.physa.2015.03.080.
[19] R. Fallahpour, S. Chakouvari, and H. Askari, "Analytical Solutions for Rumor Spreading Dynamical Model in a Social Network," *Nonlinear Eng.*, vol. 4, no. 1, pp. 23–29, 2015, doi: doi:10.1515/nleng-2014-0025.
[20] J.-E. Lönnqvist and F. große Deters, "Facebook friends, subjective well-being, social support, and personality," *Comput. Human Behav.*, vol. 55, pp. 113–120, 2016, doi: https://doi.org/10.1016/j.chb.2015.09.002.
[21] Y. Dong, H. Chen, W. Qian, and A. Zhou, "Micro-blog social moods and Chinese stock market: the influence of emotional valence and arousal on Shanghai Composite Index volume," *Int. J. Embed. Syst.*, vol. 7, no. 2, pp. 148–155, Jan. 2015, doi: 10.1504/IJES.2015.069987.
[22] Q. Cui, Z. Qiu, W. Liu, and Z. Hu, "Complex dynamics of an SIR epidemic model with nonlinear saturate incidence and recovery rate," *Entropy*, vol. 19, no. 7, pp. 1–16, 2017, doi: 10.3390/e19070305.
[23] C. Liu and Z. K. Zhang, "Information spreading on dynamic social networks," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 19, no. 4, pp. 896–904, Apr. 2014, doi: 10.1016/J.CNSNS.2013.08.028.
[24] R. Pastor-Satorras and A. Vespignani, "Epidemic Spreading in Scale-Free Networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, Apr. 2001, doi: 10.1103/PhysRevLett.86.3200.
[25] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003, doi: 10.1137/S003614450342480.
[26] M. E. J. Newman, "Threshold effects for two pathogens spreading on a network," *Phys. Rev. Lett.*, vol. 95, no. 10, pp. 1–4, 2005, doi: 10.1103/PhysRevLett.95.108701.
[27] L. Da, "Microblog Information Diffusion:Simulation Based on SIR Model," *J. Beijing Univ. Posts Telecommun.*, 2014.
[28] Y. Jin, W. Wang, and S. Xiao, "An SIRS model with a nonlinear incidence rate," *Chaos, Solitons & Fractals*, vol. 34, pp. 1482–1497, Dec. 2007, doi: 10.1016/j.chaos.2006.04.022.
[29] "Social Network Analysis : Methods and Applications by Stanley Wasserman ; Katherine Faust Review by : Christopher Winship Published by : American Statistical Association content in a trusted digital archive . We use information technology and tools to inc," vol. 91, no. 435, pp. 1373–1374, 2014.

[30] H. Li, J.-T. Cui, and J.-F. Ma, “Social Influence Study in Online Networks: A Three-Level Review,” *J. Comput. Sci. Technol.*, vol. 30, pp. 184–199, Jan. 2015, doi: 10.1007/s11390-015-1512-7.

[31] F. Xinghua, J. Zhao, B.-X. Fang, and Y.-X. Li, “Influence Diffusion Probability Model and Utilizing It to Identify Network Opinion Leader,” *Chinese J. Comput.*, vol. 36, pp. 360–367, Feb. 2014, doi: 10.3724/SP.J.1016.2013.00360.

[32] X.-D. Wu, Y. Li, and L. li, “Influence analysis of online social networks,” *Jisuanji Xuebao/Chinese J. Comput.*, vol. 37, pp. 735–752, Apr. 2014, doi: 10.3724/SP.J.1016.2014.00735.

[33] A. Leavitt, E. Burchard, D. Fisher, and S. A. Gilbert, “The Influentials : New Approaches for Analyzing Influence on Twitter,” 2009.

[34] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” *ICWSM 2010 - Proc. 4th Int. AAAI Conf. Weblogs Soc. Media*, pp. 10–17, 2010.

[35] L. Weitzel, P. Quaresma, and J. P. M. De Oliveira, “Measuring node importance on Twitter microblogging,” *ACM Int. Conf. Proceeding Ser.*, 2012, doi: 10.1145/2254129.2254145.

[36] P. Wang, C. Tian, and J. an Lu, “Identifying influential spreaders in artificial complex networks,” *J. Syst. Sci. Complex.*, vol. 27, no. 4, pp. 650–665, 2014, doi: 10.1007/s11424-014-2236-4.

[37] A. Srinivas and R. L. Velusamy, “Identification of influential nodes from social networks based on Enhanced Degree Centrality Measure,” *Souvenir 2015 IEEE Int. Adv. Comput. Conf. IACC 2015*, pp. 1179–1184, 2015, doi: 10.1109/IADCC.2015.7154889.

[38] S. Kong and L. Feng, “A tweet-centric approach for topic-specific author ranking in micro-blog,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7120 LNAI, no. PART 1, pp. 138–151, 2011, doi: 10.1007/978-3-642-25853-4_11.

[39] F. Xiao, T. Noro, and T. Tokuda, “Finding news-topic oriented influential twitter users based on topic related hashtag community detection,” *J. Web Eng.*, vol. 13, no. 5–6, pp. 405–429, 2014.

[40] J. Weng, E. P. Lim, J. Jiang, and Q. He, “TwitterRank: Finding topic-sensitive influential twitterers,” *WSDM 2010 - Proc. 3rd ACM Int. Conf. Web Search Data Min.*, pp. 261–270, 2010, doi: 10.1145/1718487.1718520.

[41] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, “Influence and passivity in social media,” *Proc. 20th Int. Conf. Companion World Wide Web, WWW 2011*, no. January 2011, pp. 113–114, 2011, doi: 10.1145/1963192.1963250.

[42] S. Räbiger and M. Spiliopoulou, “A framework for validating the merit of properties that predict the influence of a twitter user,” *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2824–2834, 2015, doi: 10.1016/j.eswa.2014.11.006.

[43] L. Jing and X. Lizhen, “Identification of microblog opinion leader based on user feature and interaction network,” *Proc. - 11th Web Inf. Syst. Appl. Conf. WISA 2014*, pp. 125–130, 2014, doi: 10.1109/WISA.2014.31.

[44] R. Watermeyer, “Social Networking Sites,” *Encycl. Appl. Ethics*, pp. 152–159, Jan. 2012, doi: 10.1016/B978-0-12-373932-2.00427-0.

[45] B. C. Molokwu and Z. Kobti, “Social network analysis using RLVECN: Representation learning via knowledge-graph embeddings and convolutional neural-network,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2021-Janua, no. ii, pp. 5198–5199, 2020, doi: 10.24963/ijcai.2020/739.
[46] “What is Twitter? - Definition from WhatIs.com.” https://whatis.techtarget.com/definition/Twitter (accessed Dec. 21, 2021).
[47] M. Popović and M. Milosavljević, “Twitter Data Analytics in Education Using IBM Infosphere Biginsights,” no. April, pp. 74–80, 2016, doi: 10.15308/sinteza-2016-74-80.
[48] I. Anger and C. Kittl, “Measuring influence on Twitter,” *ACM Int. Conf. Proceeding Ser.*, 2011, doi: 10.1145/2024288.2024326.
[49] “The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy: Keller, Edward, Berry, Jonathan: 9780743227292: Amazon.com: Books.” https://www.amazon.com/Influentials-American-Tells-Other-Where/dp/0743227298 (accessed Jan. 12, 2022).
[50] “Semantic Web - Wikipedia.” https://en.wikipedia.org/wiki/Semantic_Web (accessed Dec. 28, 2021).
[51] A. Kumar, D. Chhabra, B. Mendiratta, and A. Sinha, “Analyzing Information Diffusion in Ego-centric Twitter Social Network,” *2020 6th Int. Conf. Signal Process. Commun. ICSC 2020*, pp. 363–368, Mar. 2020, doi: 10.1109/ICSC48311.2020.9182719.
[52] V. Sugumaran, “Semantic technologies for enhancing knowledge management systems,” *Successes Fail. Knowl. Manag.*, pp. 203–213, 2016, doi: 10.1016/B978-0-12-805187-0.00014-0.
[53] “Introduction to the Semantic Web.” https://cambridgesemantics.com/blog/semantic-university/intro-semantic-web/ (accessed Jan. 07, 2022).
[54] M. M. Taye, “Understanding Semantic Web and Ontologies: Theory and Applications,” *J. Comput.*, vol. 2, no. 6, 2010.
[55] T. Slimani, “Description and Evaluation of Semantic Similarity Measures Approaches,” *Int. J. Comput. Appl.*, vol. 80, no. 10, pp. 25–33, 2013, doi: 10.5120/13897-1851.
[56] D. Chandrasekaran and V. Mago, “Evolution of Semantic Similarity-A Survey,” *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–35, 2021, doi: 10.1145/3440755.
[57] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and Application of a Metric on Semantic Nets,” *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 1, pp. 17–30, 1989, doi: 10.1109/21.24528.
[58] Z. Wu and M. Palmer, *Verbs Semantics and Lexical Selection*. 1994.
[59] D. Sánchez, M. Batet, D. Isern, and A. Valls, “Ontology-based semantic similarity: A new feature-based approach,” *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, 2012, doi: 10.1016/j.eswa.2012.01.082.
[60] S. Banerjee and T. Pedersen, “Extended gloss overlaps as a measure of semantic relatedness,” *IJCAI Int. Jt. Conf. Artif. Intell.*, no. May 2003, pp. 805–810, 2003.
[61] Y. Jiang, X. Zhang, Y. Tang, and R. Nie, “Feature-based approaches to semantic similarity assessment of concepts using Wikipedia,” *Inf. Process. Manag.*, vol. 51, no. 3, pp. 215–234, 2015, doi: 10.1016/j.ipm.2015.01.001.

[62] D. Sánchez and M. Batet, “A semantic similarity method based on information content exploiting multiple ontologies,” *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1393–1399, 2013, doi: 10.1016/j.eswa.2012.08.049.
[63] G. Zhu and C. A. Iglesias, “Computing Semantic Similarity of Concepts in Knowledge Graphs,” *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 72–85, 2017, doi: 10.1109/TKDE.2016.2610428.
[64] A. C. Khadir, H. Aliane, and A. Guessoum, “Ontology learning: Grand tour and challenges,” *Comput. Sci. Rev.*, vol. 39, p. 100339, Feb. 2021, doi: 10.1016/J.COSREV.2020.100339.
[65] C. Roussey, F. Pinet, M. A. Kang, and O. Corcho, “An Introduction to Ontologies and Ontology Engineering,” *Adv. Inf. Knowl. Process.*, vol. 1, pp. 9–38, 2011, doi: 10.1007/978-0-85729-724-2_2.
[66] A. A. Algosaibi, “Improving semantic properties relationships extraction in ontology evolution,” *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 11, pp. 2659–2668, 2021.
[67] “Handbook of Semantic Web Technologies,” *Handb. Semant. Web Technol.*, 2011, doi: 10.1007/978-3-540-92913-0.
[68] M. Krötzsch, F. Simančík, and I. Horrocks, “A description logic primer,” *Perspect. Ontol. Learn.*, vol. 18, no. June, pp. 3–20, 2014.
[69] “What are Ontologies? | Ontotext Fundamentals Series.” https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/ (accessed Dec. 22, 2021).
[70] M. Somodevilla García, D. Vilariño Ayala, and I. Pineda, “An overview of ontology learning tasks,” *Comput. y Sist.*, vol. 22, no. 1, pp. 137–146, 2018, doi: 10.13053/CyS-22-1-2790.
[71] “Extensible Markup Language (XML) 1.0 (Fifth Edition).” https://www.w3.org/TR/REC-xml/ (accessed Jan. 07, 2022).
[72] S. Bechhofer, “OWL: Web Ontology Language,” *Encycl. Database Syst.*, pp. 2008–2009, 2009, doi: 10.1007/978-0-387-39940-9_1073.
[73] A. Farquhar, R. Fikes, and J. Rice, “The Ontolingua Server: a tool for collaborative ontology construction,” *Int. J. Hum. Comput. Stud.*, vol. 46, pp. 707–727, 1997.
[74] “Knowledge Interchange Format (KIF).” http://www-ksl.stanford.edu/knowledge-sharing/kif/ (accessed Jan. 07, 2022).
[75] M. Kifer, G. Lausen, and J. Wu, “Logical foundations of object-oriented and frame-based languages,” *J. ACM*, vol. 42, no. 4, pp. 741–843, Jul. 1995, doi: 10.1145/210332.210335.
[76] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.
[77] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, “Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction,” *5th Int. Symp. Lang. Biol. Med. LBM’13*, no. i, pp. 45–49, 2013, [Online]. Available: http://www.lirmm.fr/~jonquet/publications/documents/Article_LBM2013_Lossio.pdf.
[78] “Distributional Hypothesis - ACL Wiki.”

https://aclweb.org/aclwiki/Distributional_Hypothesis (accessed Jan. 08, 2022).
[79] "WordNet Search - 3.1." http://wordnetweb.princeton.edu/perl/webwn (accessed Jan. 08, 2022).
[80] D. Tu, L. Chen, and G. Chen, "WordNet based multi-way concept hierarchy construction from text corpus," *Proc. 27th AAAI Conf. Artif. Intell. AAAI 2013*, no. 1647, pp. 1647–1648, 2013.
[81] "BabelNet | The largest multilingual encyclopedic dictionary and semantic network." https://babelnet.org/ (accessed Jan. 08, 2022).
[82] P. Cimiano, M. Alexander, S. Staab, and V. Johanna, "Handbook on Ontologies," *Handb. Ontol.*, pp. 245–267, 2009, doi: 10.1007/978-3-540-92673-3.
[83] "GATE, a General Architecture for Text Engineering on JSTOR." https://www.jstor.org/stable/30204529 (accessed Jan. 08, 2022).
[84] P. Cimiano and J. Völker, "Text2Onto A framework for ontology learning and data-driven change discovery," *Lect. Notes Comput. Sci.*, vol. 3513, pp. 227–238, 2005, doi: 10.1007/11428817_21.
[85] "Learning Expressive Ontologies | Request PDF." https://www.researchgate.net/publication/271131823_Learning_Expressive_Ontologies (accessed Jan. 08, 2022).
[86] P. Velardi, R. Navigli, A. Cucchiarelli, and F. Neri, "Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies," 2003.
[87] "Linguistic patterns for information extraction in OntoCmaps | Request PDF." https://www.researchgate.net/publication/289151317_Linguistic_patterns_for_information_extraction_in_OntoCmaps (accessed Jan. 08, 2022).
[88] E. Drymonas, K. Zervanou, and E. G. M. Petrakis, "Unsupervised ontology acquisition from plain texts: The OntoGain system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6177 LNCS, pp. 277–287, 2010, doi: 10.1007/978-3-642-13881-2_29.
[89] A. Ranjan Pal and D. Saha, "Word Sense Disambiguation: A Survey," *Int. J. Control Theory Comput. Model.*, vol. 5, no. 3, pp. 1–16, 2015, doi: 10.5121/ijctcm.2015.5301.
[90] "Ambiguity (Stanford Encyclopedia of Philosophy)." https://plato.stanford.edu/entries/ambiguity/ (accessed Jan. 08, 2022).
[91] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, "Recent Trends in Word Sense Disambiguation: A Survey," pp. 4330–4338, 2021, doi: 10.24963/ijcai.2021/593.
[92] E. Agirre, O. L. de Lacalle, and A. Soroa, "Random Walks for Knowledge-Based Word Sense Disambiguation," *Comput. Linguist.*, vol. 40, no. 1, pp. 57–84, Mar. 2014, doi: 10.1162/COLI_A_00164.
[93] A. Moro, A. Raganato, R. Navigli, and V. R. Elena, "Entity Linking meets Word Sense Disambiguation," *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 231–244, 2014, [Online]. Available: http://babelfy.org.
[94] R. Tripodi and R. Navigli, "Game Theory Meets Embeddings: a Unified Framework for Word Sense Disambiguation," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 88–99, 2019, doi: 10.18653/V1/D19-1009.

[95] M. Kumar, P. Mukherjee, M. Hendre, M. Godse, and B. Chakraborty, "Adapted lesk algorithm based word sense disambiguation using the context information," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 254–260, 2020, doi: 10.14569/ijacsa.2020.0110330.
[96] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990, doi: 10.1093/ijl/3.4.235.
[97] "(PDF) English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology." https://www.researchgate.net/publication/344298875_English_WordNet_2020_Improving_and_Extending_a_WordNet_for_English_using_an_Open-Source_Methodology (accessed Jan. 08, 2022).
[98] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artif. Intell.*, vol. 193, pp. 217–250, 2012, doi: 10.1016/j.artint.2012.07.001.
[99] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, and F. Cecconi, "Ten Years of BabelNet: A Survey," pp. 4559–4567, 2021, doi: 10.24963/ijcai.2021/620.
[100] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2276, pp. 136–145, 2002, doi: 10.1007/3-540-45715-1_11.
[101] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.*, vol. 41, no. 2, 2009, doi: 10.1145/1459352.1459355.
[102] A. Anshu, "Review Paper on Data Mining TechniquesandApplications," *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 7, no. 2, pp. 22–26, 2019, doi: 10.21276/ijircst.2019.7.2.4.
[103] F. Gorunescu, *Data Mining: Concepts, models and techniques*. 2011.
[104] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," 2017, [Online]. Available: http://arxiv.org/abs/1707.02919.
[105] "Understanding TF-ID: A Simple Introduction." https://monkeylearn.com/blog/what-is-tf-idf/ (accessed Jan. 08, 2022).
[106] M. J. Lavin, "Analyzing Documents with TF-IDF," *Program. Hist.*, no. 8, May 2019, doi: 10.46430/PHEN0082.
[107] K. Frantzi, S. Ananiadou, and H. Mima, "Natural language processing for digital libraries Automatic recognition of multi-word terms: the C-value/NC-value method," *Int J Digit Libr*, vol. 3, no. November 2015, pp. 115–130, 2000, doi: 10.1007/3-540-49653-X.
[108] J. C. Sager, "A Practical Course in Terminology Processing," *A Pract. Course Terminol. Process.*, Jan. 1990, doi: 10.1075/Z.44.
[109] S. M. Katz, "Technical terminology: Some linguistic properties and an algorithm for identification in text," *Nat. Lang. Eng.*, vol. 1, no. 1, pp. 9–27, 1995, doi: 10.1017/S1351324900000048.
[110] R. Kumar *et al.*, "Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: A critical review," *J. Mater. Res. Technol.*, vol. 10, no. 4, pp. 1471–1492, 2021, doi:

10.1016/j.jmrt.2020.12.114.
[111] F. H. Lotfi and R. Fallahnejad, “Imprecise shannon’s entropy and multi attribute decision making,” *Entropy*, vol. 12, no. 1, pp. 53–62, 2010, doi: 10.3390/e12010053.
[112] H. G. Nikunj Agarwal, “Entropy Based Multi-criteria Decision Making Method under Fuzzy Environment and Unknown Attribute Weights,” *Glob. J. Technol. Optim.*, vol. 06, no. 03, 2015, doi: 10.4172/2229-8711.1000182.
[113] K. Dashore *et al.*, “Product evaluation using entropy and multi criteria decision making methods,” *Int. J. Eng. Trends Technol.*, vol. 4, no. 5, pp. 2183–2187, 2013, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.3662&rep=rep1&type=pdf%0Ahttp://www.ijettjournal.org.
[114] A. Mert, “Shannon entropy-based approach for calculating values of WABL parameters,” *J. Taibah Univ. Sci.*, vol. 14, no. 1, pp. 1100–1109, 2020, doi: 10.1080/16583655.2020.1804157.
[115] D. Diakoulaki, G. Mavrotas, and L. Papayannakis, “Determining objective weights in multiple criteria problems: The critic method,” *Comput. Oper. Res.*, vol. 22, no. 7, pp. 763–770, 1995, doi: 10.1016/0305-0548(94)00059-H.
[116] A. Jahan, F. Mustapha, S. M. Sapuan, M. Y. Ismail, and M. Bahraminasab, “A framework for weighting of criteria in ranking stage of material selection process,” *Int. J. Adv. Manuf. Technol. 2011 581*, vol. 58, no. 1, pp. 411–420, May 2011, doi: 10.1007/S00170-011-3366-7.
[117] H. Deng, C. H. Yeh, and R. J. Willis, “Inter-company comparison using modified TOPSIS with objective weights,” *Comput. Oper. Res.*, vol. 27, no. 10, pp. 963–973, Sep. 2000, doi: 10.1016/S0305-0548(99)00069-6.
[118] “Coronavirus (covid19) Tweets - early April | Kaggle.” https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april (accessed Jan. 27, 2022).
[119] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, “The University of South Florida free association, rhyme, and word fragment norms,” *Behav. Res. Methods, Instruments, Comput. 2004 363*, vol. 36, no. 3, pp. 402–407, 2004, doi: 10.3758/BF03195588.
[120] “OntoMetrics.” https://ontometrics.informatik.uni-rostock.de/ontologymetrics/ (accessed Jan. 15, 2022).
[121] “SparkScore from SparkToro.” https://sparktoro.com/tools/sparkscore (accessed Jan. 15, 2022).
[122] A. Pawar and V. Mago, “Calculating the similarity between words and sentences using a lexical database and corpus statistics,” no. February, 2018, [Online]. Available: http://arxiv.org/abs/1802.05667.

Appendix (A)

comparison our system result with Mean Human Similarity and Atish Pawar Algorithm					
No.	**Sentence 1**	**Sentence 2**	**Mean Human Similarity**	**Atish Pawar Algorithm**	**Proposed Algorithm**
1.	Cord is strong, thick string.	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.01 0	0.0225	0.0352
2.	A rooster is an adult male chicken	A voyage is a long journey on a ship or in a spacecraft	0.005	0.2593	0.0679
3.	Noon is 12 o'clock in the middle of the day.	String is thin rope made of twisted threads, used for tying things together or tying up parcels	0.0125	0.03455	0.0917
4.	Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam	0.0475	0.1388	0.1106
5.	famous which is specially written for a fan to keep.	The shores or shore of a sea, lake, or wide river is the land along the edge of it.	0.0050	0.0701	0.05681
6.	An automobile is a car.	In legends and fairy stories, a wizard is a man who has magic powers.	0.0200	0.0088	0.0135
7.	A mound of something is a large rounded pile of it	A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.0050	0.4968	0.0320
8.	A grin is a broad smile.	An implement is a tool or other pieces of equipment	0.0050	0.0099	0.0469
9.	An asylum is a psychiatric hospital.	Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.0500	0.01456	0.0775
10.	An asylum is a psychiatric hospital.	A monk is a member of a male religious community that is usually separated from the outside world.	0.0375	0.0175	0.0273
11.	A graveyard is an area of land, sometimes near a church, where dead people are buried.	If you describe a place or situation as a madhouse,you mean that it is full of confusion and noise.	0.0225	0.1339	0.0878
12.	Glass is a hard transparent substance that is used to make things such as windows and bottles	A magician is a person who entertains people by doing magic tricks.	0.0075	0.0911	0.0906
13.	A boy is a child who will grow up to be a man.	A rooster is an adult male chicken	0.1075	0.2921	0.2971
14.	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable	A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.0525	0.1745	0.1872
15.	mmunity that is usually separated from the outside world.	A slave is someone who is the property of another person and has to work for that person.	0.0450	0.1394	0.1997
16.	An asylum is a psychiatric hospital.	A cemetery is a place where dead peoples bodies or their ashes are buried	0.375	0.03398	0.0761

17.	The coast is an area of land that is next to the sea.	A forest is a large area where trees grow close together.	0.0475	0.3658	0.0783
18.	A grin is a broad smile.	A lad is a young man or boy	0.0125	0.0281	0.0160
19.	The shores or shore of a sea, lake, or wide river is the land along the edge of it.	Woodland is land with a lot of trees.	0.0825	0.3192	0.1288
20.	A monk is a member of a male religious community that is usually separated from the outside world.	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.1125	0.1011	0.1195
21.	A boy is a child who will grow up to be a man.	A sage is a person who is regarded as being very wise.	0.0425	0.2305	0.2292
22.	An automobile is a car.	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable.	0.0200	0.0330	0.1742
23.	A mound of something is a large rounded pile of it.	The shores or shore of a sea, lake, or wide river is the land along the edge of it.	0.0350	0.0386	0.0503
24.	A lad is a young man or boy.	In legends and fairy stories, a wizard is a man who has magic powers	0.0325	0.3939	0.6438
25.	A forest is a large area where trees grow close together.	A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.0650	0.2787	0.1207
26.	Food is what people and animals eat.	A rooster is an adult male chicken.	0.0550	0.2972	0.1093
27.	A cemetery is a place where dead people's bodies or their ashes are buried.	Woodland is land with a lot of trees.	0.0375	0.1240	0.0234
28.	The shores or shore of a sea, lake, or wide river is the land along the edge of it.	A voyage is a long journey on a ship or in a spacecraft.	0.0200	0.0304	0.0796
29.	A bird is a creature with feathers and wings females lay eggs and most birds can fly.	Woodland is land with a lot of trees.	0.0125	0.1334	0.1249
30.	The coast is an area of land that is next to the sea.	A hill is an area of land that is higher than the land that surrounds it.	0.1000	0.8032	0.5378
31.	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.	An implement is a tool or other piece of equipment.	0.0500	0.1408	0.0570
32.	A crane is a large machine that moves heavy things by lifting them in the air.	A rooster is an adult male chicken.	0.0200	0.0564	0.0797
33.	A hill is an area of land that is higher than the land that surrounds it.	Woodland is land with a lot of trees.	0.1450	0.7619	0.1891
34.	A car is a motor vehicle with room for a small number of passengers.	When you make a journey, you travel from one place to another.	0.0725	0.02610	0.0465
35.	A cemetery is a place where dead peoples bodies or their ashes are buried.	A mound of something is a large rounded pile of it.	0.0575	0.0842	0.0558

36.	Glass is a hard transparent substance that is used to make things such as windows and bottles.	A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.1075	0.2692	0.1474
37.	A magician is a person who entertains people by doing magic tricks.	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.1300	0.1000	0.1102
38.	A crane is a large machine that moves heavy things by lifting them in the air.	An implement is a tool or other piece of equipment	0.1850	0.1060	0.1745
39.	Your brother is a boy or a man who has the same parents as you	A lad is a young man or boy.	0.1275	0.9615	0.7685
40.	A sage is a person who is regarded as being very wise.	In legends and fairy stories, a wizard is a man who has magic powers.	0.1525	0.1920	0.1611
41.	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	A sage is a person who is regarded as being very wise	0.2825	0.0452	0.1257
42.	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	A crane is a large machine that moves heavy things by lifting them in the air.	0.0350	0.1660	0.1124
43.	A bird is a creature with feathers and wings females lay eggs, and most birds can fly.	A cock is an adult male chicken	0.1625	0.1704	0.1832
44.	Food is what people and animals eat.	Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.2425	0.1379	0.4368
45.	Your brother is a boy or a man who has the same parents as you.	A monk is a member of a male religious community that is usually separated from the outside world.	0.0450	0.2780	0.2247
46.	An asylum is a psychiatric hospital.	If you describe a place or situation as a madhouse, you mean that it is full of confusion and noise.	0.2150	0.1860	0.3107
47.	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish, or produce steam.	A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.3475	0.1613	0.1508
48.	A magician is a person who entertains people by doing magic tricks.	In legends and fairy stories, a wizard is a man who has magic powers.	0.3550	0.5399	0.2392
49.	A hill is an area of land that is higher than the land that surrounds it.	A mound of something is a large rounded pile of it.	0.2925	0.2986	0.5633
50.	Cord is strong thick string.	String is thin rope made of twisted threads used for tying things together or tying up parcels.	0.4700	0.2530	0.3639
51.	Glass is a hard transparent substance that is used to make things such as windows and bottles.	A tumbler is a drinking glass with straight sides.	0.1375	0.3016	0.4108
52.	A grin is a broad smile.	A smile is the expression that you have on your face when you are pleased or amused or when you are being friendly.	0.4850	0.8419	0.3773

53.	In former times, serfs were a class of people who had to work on a particular persons land and could not leave without that person's permission.	A slave is someone who is the property of another person and has to work for that person.	0.4825	0.8896	0.6892
54.	When you make a journey, you travel from one place to another.	A voyage is a long journey on a ship or in a spacecraft.	0.3600	0.7826	0.3017
55.	An autograph is the signature of someone famous which is specially written for a fan to keep.	Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says.	0.4050	0.3146	0.6514
56.	The coast is an area of land that is next to the sea.	The shores or shore of a sea, lake, or wide river is the land along the edge of it.	0.5875	0.9773	0.5490
57.	A forest is a large area where trees grow close together.	Woodland is land with a lot of trees.	0.6275	0.4770	0.5707
58.	An implement is a tool or other pieces of equipment.	A tool is any instrument or simple piece of equipment that you hold in your hands and use to do a particular kind of work.	0.5900	0.8919	0.7763
59.	A cock is an adult male chicken.	A rooster is an adult male chicken	0.8625	0.8560	0.9489
60.	A boy is a child who will grow up to be a man	A lad is a young man or boy.	0.5800	0.8980	0.6770
61.	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable	A pillow is a rectangular cushion which you rest your head on when you are in bed.	0.5225	0.9340	0.6770
62.	A cemetery is a place where dead peoples bodies or their ashes are buried.	A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.7725	1.0	0.7542
63.	An automobile is a car.	A car is a motor vehicle with room for a small number of passengers	0.5575	0.7001	0.7629
64.	Midday is 12 oclock in the middle of the day.	Noon is 12 oclock in the middle of the day.	0.9550	0.8726	0.9477
65.	A gem is a jewel or stone that is used in jewellery	A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces	0.6525	0.8536	0.7009

الخلاصة :

ادى الانتشار الهائل لوسائل التواصل الاجتماعي مثل تويترإلى تسريع عملية مشاركة المعلومات والتعبير عن الآراء حول الأزمات الصحية العالمية والأحداث المهمة. حيث يستخدم تويتر من قبل العديد من المسؤولين الحكوميين حول العالم كواحدة من قنوات الاتصال الرئيسية للمشاركة في الأحداث السياسية والأخبار المتعلقة بكوفيد بشكل منتظم لعامة الناس .

نظرًا لاستخدام مصطلحات مختلفة للتعبير عن نفس الموضوع في منشور تويتر، يصبح من الصعب إنشاء تطبيقات مثل العثور على نسبة انتشار اخبار معينة على تويتر اوالاستعلام عن تغريدات معينة لكوفيد اواسترداد معلومات معينة و الترجمة الآلية ، باتباع تقنيات مطابقة النص التقليدية السابقة. ومن أجل حل هذه المشكلة ، يتطلب الأمر توفير مصدر معرفة يجمع المصطلحات التي تعكس معنى واحدًا في مفهوم رسمي بخصائصها وعلاقاتها ، مثل الانتولوجي، وتقنية لحساب التشابه الدلالي بين هذه المفاهيم.

تقدم هذه الأطروحة طريقة لاكتشاف انتشار المعلومات في تويتر بناءً على المعنى الدلالي بدلاً من مطابقة الكلمات بين الجملة. باستخدام تقنية التشابه الدلالي المبنية على الأنطولوجيا والمستخدمين المؤثرين. وبسبب عدم وجود انتولوجي في هذا المجال ، تم بناء مصدر معرفة (انتولوجي) يشرح المفاهيم المستخدمة في هذا المجال بطريقة دلالية ، وتضمن 1203 مفاهيم. وهي أكثر المفاهيم المستخدمة في تغريدات تويتر مع انتشار وباء كورونا.

تم تقييم النظام المقترح من خلال مجموعة بيانات مقيمة بواسطة الانسان ، حيث أظهرت النتائج دقة 89.1٪ مقارنه مع التقييم البشرية ، وكذلك كانت النتائج أفضل من نتيجة Atish (73.9%) التي تعد من أحدث الأبحاث في هذا المجال باستخدام نفس ابيانات .

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل - كلية تكنولوجيا المعلومات
قسم البرمجيات

اكتشاف انتشار المعلومات باستخدام التشابه الدلائلي القائم على الانتولوجي والمستخدمين المؤثرين

اطروحة

مقدمة إلى مجلس كلية تكنولوجيا المعلومات للدراسات العليا بجامعة بابل كجزء

لمتطلبات درجة دكتوراه فلسفة في تكنولوجيا المعلومات / البرمجيات

من قبل

ياسر عبدالحميد نجم منصور

بأشراف

أ. د. اسعد صباح هادي عباس

1443 هـ

2022 م