SKEM++: SEMANTIC KEYWORD EXTRACTION MODEL USING COLLECTIVE CENTRALITY MEASURE ON BIG SOCIAL DATA

Devika, R[1], Subramaniyaswamy, V[2,*]

[1,2]School of Computing, SASTRA Deemed University, Thanjavur, India

Email: devika.siva6@gmail.com[1], vsubramaniyaswamy@gmail.com[2*] (corresponding author)

DOI: https://doi.org/10.22452/mjcs.sp2022no1.1

ABSTRACT

In recent times, Online Social Network (OSN) has accumulated a massive volume of user-generated data available in an unstructured format. It consists of user ideas, responses, and opinions on various topics. It extracts essential keywords in OSN, which is endowed with many exciting applications such as information recommendation or viral marketing. This paper emphasizes the importance of semantic graph-based methods for extracting vital keywords experimentally using a novel SKEM++ method. It is an innovative method for keyword extraction from OSN based on centrality measures. It utilizes a distributed computing approach to calculate the network Collective Centrality Measure (CCM) for each node and improve the semantics of keywords. The distributed approach is more scalable and computationally efficient than the conventional system, making it more suitable for large-scale real-time data sets such as the OSN. Experimental outcomes on the real-time Twitter Data set to infer the dominance of the proposed Collective Centrality Measure(CCM) method in evaluation with contemporary schemes in terms of F-score by 81% and recall by 80% and precision by 80% using Semantic Analysis.

Keywords: ***Online Social Network (OSN), Semantic Connectivity, Collective Centrality Measure (CCM), Keyword Extraction***

1.0 INTRODUCTION

The OSN is a graph G (V, E) in which each user is represented by a node (N), and the edges between the nodes represent the connection between two users (E). Several systems are represented by graphs which are a collection of nodes connected by edges. Investigating and analyzing social networks using graph theory is called Social Network Analysis (SNA). Social network analysis is carried out for various reasons such as identifying famous and influential people, patterns of activity, user interest groups, etc. Identifying vital nodes is one of the most significant tasks in social network analysis. Since many of the nodes may reveal different levels of activities, these nodes are active or passive [1]. The intent is to extract keywords using graph-based methods by collective centrality measures. Such an attempt enables identifying and analyzing tendencies or patterns from a large volume of data from diverse [6] data streams. Owing to the dynamic nature of Online Social networks (OSN) such as Twitter, Facebook, etc., exponentially data is generated. This kind of data is often known as Streaming data. It consists of varied pieces of precious information of different sizes. The heterogeneous nature of data is akin to analyzing more complex networks, as shown graphically in Figure 1.

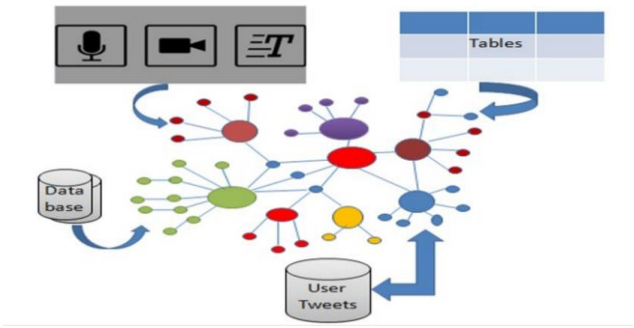


Fig 1: Extracting Semantic graph from Unstructured Data in OSN

Twitter is one of the most popular social networking sites which generates large-scale social data. This social media platform allows users to communicate and remain linked via regular exchanges of messages. The examination and evaluation of Twitter data are valuable because the information mined can be used for various applications. The automatic extraction of keywords while analyzing data on Twitter is one of the main tasks, and it is the paper's primary focus

Various organizations have modeled real-time streaming to estimate tweets to find a suitable business design for them. Twitter is one of the most excessive sources of live and high public connectivity worldwide. On average, 6000 tweets are being tweeted every second on Twitter. Various applications such as industry-wide trends, stock market predictions, alerts for epidemic outbreaks, etc., can be created using this data. Graph-based keyword extraction methods are beneficial to select the most significant nodes. The influential node is defined on the strength of the node, which is calculated by its average centrality measures score. The main contributions of this research are:

- It uses APACHE KAFKA as the main component that allows for real-time event-driven application development and collects real-time streaming tweets from Twitter.
- Using an adaptive LESK algorithm, it pre-processes the tweets as plain words to find the semantic connectivity and its relationship with other words (nodes). It then uses the noun phrasing approach to select nouns as keywords and to remove redundancy.
- It applies the novel Collective Centrality Measure (CCM), a unique and novel measure, to combine nine centrality measures. Considered the first of its kind strategy, it helps find the location of influential nodes in the input graph.
- It illustrates the proficiency of the proposed method for extracting influential nodes (words) from Twitter of online streaming data.
- It establishes that the proposed CCM approach is very efficient on different datasets obtaining high precision, F1 Score, and recall values.

In this algorithm, a lexical database WordNet is employed, which provides a rich hierarchy of semantic meaning. In natural language processing, most of the words are polysemous, which means that each word has multiple meanings. The paper is structured as follows: Section 2.0 offers a complete insight into the required literature. Section 3.0 presents the techniques used in the proposed SKEM++ methods. Section 4.0 describes the proposed methodologies. Section 5.0 presents the experimental evaluation of proposed methods. Finally, section 6.0 widens the scope of the work and specifies some future research directions.

2.0 RELATED WORKS

Various statistics-based approaches use simple statistics like frequency to identify the candidate keywords [11] and [9] spatial distribution of terms [4][8] and reported in the literature. These authors carried out a study by identifying keywords using linguistic form analysis like lexical, semantic, and discourse analysis. A linguistic analysis approach searches exact matches of the query words without understanding the overall meaning of the query. On the other hand, semantic analysis is an approach that verifies the correctness of statements based on the accuracy of their significance, clarity, and consistency [2]. Discourse analysis is used for written or spoken language, specifically in a social context. In recent years [14], machine learning has gained immense popularity in keyword extraction.

Another approach described in [7] [10] represents text as a graph using Graph theory. In this approach, the unique terms consider vertices and links establish relationships among vertices. Applicant terms are graded also using local or overall properties. For analyzing and extracting keywords from the document, a Graph model is always used. In [3] described how edges are recognized and found between the texts in a text graph. The most commonly attributed relation is the term co-occurrence, where two or more words occur within a gap of built-in size from a graph. In addition, a mechanism called term counting those selective activities properties of vertices are used to identify keywords. The Key-Graph method allocates the probability of two text keywords to groups, where each group is related to a concept. Each group has a collection of graded words by an uncertainty parameter that gives the link of each term to its parentages groups.

The reserve words are high-ranking comments. In [12] [15], the team conventional that co-occurrence text graphs exhibit only two nodes linked through a series of nodes. [13] and [5] proposed a key-world scoring method. This method is established on the influence of each node of the graph into the small-world property. Text Rank [7] is among the best practices for graph-based keyword extraction. Another approach is degree-based keyword extraction, which uses the number of edges incident to that nodes. Computationally, it is further proficient than text ranking methods. One more approach is position rank [22], an extension of Text rank. It takes positional information

of terms into account to assign weights to the candidate keywords, favoring words occurring towards the beginning of the text. This method declares the positional importance of the phrase accorded by statistical methods.

In [19], it is assumed that the vertices contributing to the text graph's greatest consistent element are appropriate applicants for keywords. The nodes perform a core-based disintegration for the graph to obtain the keywords. In [23] [20] , truss-based decomposition has been done. To hold vertices from the top-truss as keywords. Subsequently, the quantity of keywords mined through these approaches adjusts towards the building of the graph; these methods are parameter-free.Node selectivity is defined as the average weight distribution on the links of the single node by [7]. In Croatian texts, they applied as well as compared to the number of edges towards the node, number of advantages away from the node, shortest paths between nodes and all other nodes, and betweenness measures. However, it is highly impossible to match their value on English benchmarks only in Croatian texts. Analyzing the text graphs is used to improvise the efficiency of keyword extraction performance. In that technique, by similarity with the idea of significant presenters in social networks, they thought that powerful text in graphs-of-text would act as specific keywords and thus proposed the usage of K-Core and K-Truss.

Other works in the field used variants of PageRank for automatic keyword extraction, in Weighted Page Rank, Biased-PageRank, etc. With prior knowledge, position rank considers the document's positional information to allocate values to the applicant keywords. In [21], single rank is a simple modification of the text rank method. It considers link values with the total co-existence count and extracts key phrases by collecting categorized text. From the multiple documents, co-occurrence is constructed using these methods. Topic rank represents a text as a comprehensive graph in which the nodes represent the theme, and each theme is a collection of related words.

3.0 PRELIMINARIES

Definition 1: Degree (DE): In graph G (V, E) number of edges (E) incident to that node (V).

$$\deg(n) = |\{n'|n' \in V \wedge \exists (n, n') \in E\}| \quad (1)$$

where n and n' are vertices and E is the edge between vertices.

Definition 2: Betweenness (BE): It measures the probability that a random shortest path passes through a given node(v)in graph G.

$$(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

where σst is the total number of shortest paths from node s to node t and
σst(v) is the number of those paths that pass-through v

Definition 3: PageRank (PR**)** is termed over arbitrary walk in the network. It is the variant of eigenvector centrality.

$$\pi_a^t = \sum_b \frac{\pi_b^{t-1}}{g_b} + \frac{(1-\propto)e_a}{m} \quad (3)$$

where $\pi^{\rightarrow t}$ ->Page Rank values for individual vertex step t.

g_b->Degree of neighbor b that have a, vertices with the same probability.

A bounce in the route is spoken by the likelihood α. $\sum_a \pi_a = 1$ for all steps t; typically $\propto = 0.85$.

Definition 4: Closeness (CL) assigns more importance to vertices that are close to all other vertices in the Graph.

$$C(v) = n - 1/ \sum_{v=1}^{n-1} d(x, y) \quad (4)$$

where d (x, y) is the proximity between node x and y; n-1 is the normalized closeness by the sum of minimum possible distance.

Definition 5: Structural Holes (SH): The vertices occupying the bridge between different gatherings in networks.

$$SH = \sum b \left[1 - \sum_{aq} p_{aq} m_{bq}\right] \quad (5)$$

where p_{aq}-> Extent of a' s vitality put resources into a relationship with q,

m_{bq} -> b's interaction to q divided by b's closest relationship with anyone.

The excessive organization of a graph is determined by summarizing this item over all the hubs q. This articulation communicates the non-redundant part of the relationship. The viable size of a's system is characterized as an entirety of the b's non-redundant contacts.

Definition 6: Eigenvector (EV) consists of nodes with same number of incident edges may have dissimilar ranks of position liable on the position of their neighbors. Let A = [aij] be a square Matrix of order n.

$$AX = \lambda X \quad (6)$$

where λ is called as Eigen value of the given matrix A.
X is called a Eigen vector of the given Matrix A corresponding to the eigenvalues λ .

Definition 7: K-Core (KC): The graph is split into cores or shells. The K-Core centrality is calculated as

$$K(a) = G_c \quad (7)$$

where A core (G_c) means that group of vertices with degree h_a are held with G_c core.
Definition 8: Clustering coefficient (CC): The clustering coefficient measures the nearest number of edges (iteration of request cycles) on the system network.

$$CC_{a=} \frac{2e_a}{h_a(h_a - 1)} \quad (8)$$

where 'e_a' -> characterized as the sum of three-way relationship focused on and over its most extreme number of potential associations.
$h_a \in \{0, 1\}$, CC to have value = 0, and $CC_{a=}1$ just if all adjacent of 'a' are intersected.
Definition 9: Eccentricity (EC): The maximum route from one vertex to all other vertices is known as eccentricity.

$$E_a = \max_{a \neq b} \{|j_{ab}|\} \quad (9)$$

where $|j_{ab}|$ -> remoteness of the shortest path between vertices a and b.

4.0 PROPOSED SKEM++ METHODOLOGY

Graph-based approaches are getting immense attention due to the extensive usage of online social networks, and researchers have moved towards graph-based methods. The Vector-based methods cause more complexity compared to graph-based methods. This graph-based method gives a compressed and proficient representation of text. The workflow employed for improvised semantic graph-based keyword extraction on big social data is shown in Fig 2. The first phase preprocesses the real-time tweets extracted from Twitter using Kafka. The next step is applying the adaptive Lesk algorithm to get the semantic keywords. Then these keywords are filtered using the noun phrasing approach, and the construction of the semantic co-occurrence undirected unweighted graph is carried out. In the next phase, for each node, the Collective Centrality Measures are applied to calculate the node weight, and the most influential nodes in the given graph are ranked.

Phase4.1: Preprocess text and Semantic Word Extraction and Graph Construction

Tweets are collected from Twitter using Apache KAFKA, a tool used to read real-time streaming data. For example, one can write the message (publish) and read the same (subscribe), called the publish-subscribe model. Once the letters are consumed, they are deleted from the KAFKA. In Twitter's streaming API, we are sampling data that have previously occurred. Twitter's Streaming API is used to extract Tweets that occur in near real-time. After preprocessing the tweets, the Adaptive Lesk algorithm removes the word ambiguity using dictionary methods. The dictionary method can denote the sense of the term to be disambiguated and used in this work. This method does not rely on corpora-based evidence for disambiguation—it assists in finding the correct meaning for one word at a time. Then, the noun-phrasing approach is applied to filter the keywords.

The major problem faced while constructing a semantic graph is the word ambiguity. Depending upon the context of the phrase, it has a similar meaning and sometimes a different meaning. The Adaptive Lesk algorithm addresses this problem. In the Semantic Graph G (V, E) where V is the set of nodes and E is the link that represents the meaningful relationships between words. If Vm and Vn cooccur together more often in the context, then their semantic tend to increase. For constructing a co-occurrence graph, words taken from the noun phrasing approach are used as vertices. These words, which are initially fetched from Lesk, are selected by word frequency after preprocessing. This is because terms derived from the improvised Lesk algorithm are interpreted using several preprocessing methods.

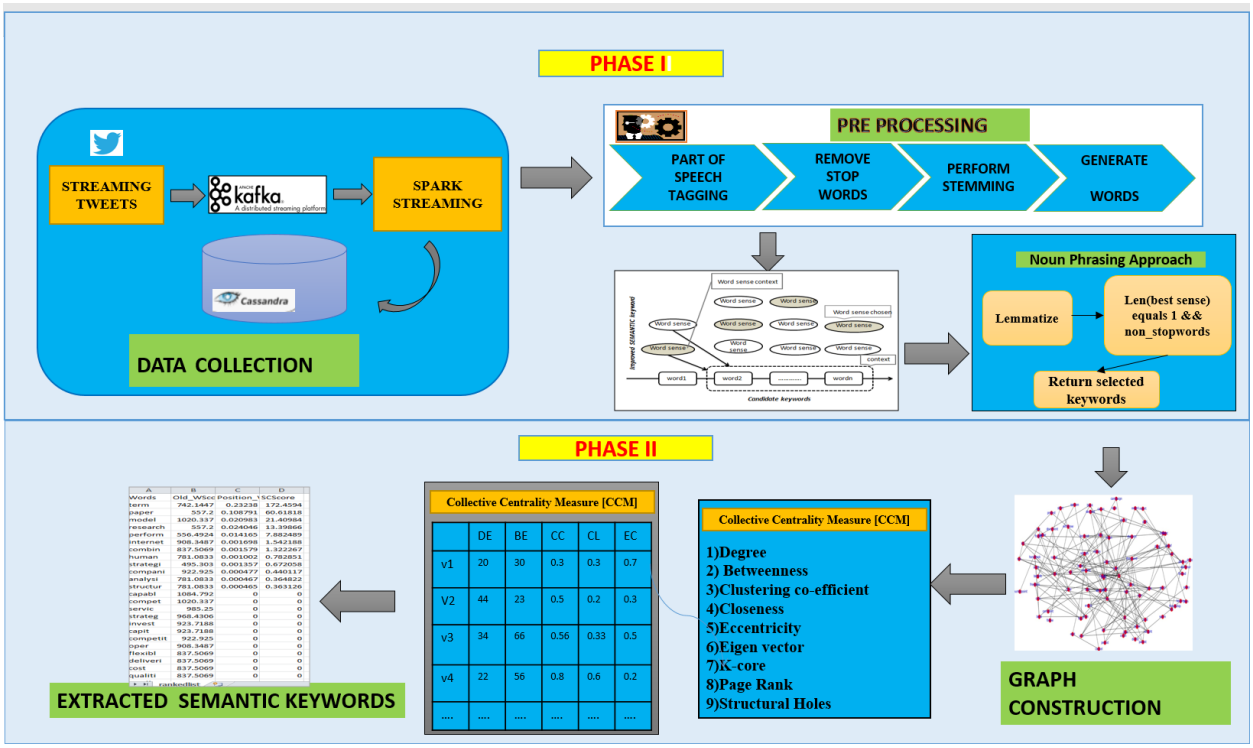


Fig 2: Architecture of Semantic graph-based Keyword Extraction Method

The co-occurrence graph's vertices multiply the values of the following centrality types, i. e., vertex degree that gathers local network information. The hubs and keywords work well. Inverse clustering coefficient extracts knowledge from the network on vertex multiplied by degree. Higher clustering coefficients also function well as the keywords. K-Core collects network knowledge from global outlets. Betweenness centrality is used for finding the shortest distance in the network. Vertices that have strong values, including keywords, often function well. Closeness centrality gathers global knowledge connected to the shortest path lengths and chooses the nearest vertices to the network's other vertices, which serve for keyword classification. Consequently, standard centrality techniques can catch similar concepts, which are represented as keywords. All the tests usually classify top-rated terms as keywords.

Algorithm 1: Improvised Lesk algorithm and noun phrasing for semantic keyword and graph construction

INPUT: Pre-processed tweets, obtained from Twitter's Streaming API
OUTPUT: Returns Semantic graph G

Execute (G, S)

1. Initialize max_overlap = 0
2. Initialize best_sense as None
3. Initialize context with a set of words W in sentence S
4. for all W€S
5. Initialize signature with a set of words in WordNet and examples of sense
6. overlap = COMPUTEOVERLAP (signature, context)
7. end for
8. If overlap > max_overlap then
9. max_overlap = overlap
10. best_sense = sense
11. end if
12. for all best_sense € S
13. normalize(best_sense) // normalize to lowercase and lemmatizes it
14. if len(best_sense) == 1 and not in stopwords then
15. yield keywords k
16. end if
17. end for
18. G: an initially empty graph
19. for all node in k
20. G.add_node(k)
21. end for

22. for all pairs of vertices (v_i , v_j) do
23. G.add_edge(v_i , v_j)
24. end for
25. Return Graph G

Phase 4.2: Node Centrality Calculation

In this phase, nine centrality measures are applied to each node to find the most influential node. The nine centrality measures are Degree, Betweenness, PageRank, Closeness, k-core, EigenVector, Clustering coefficient, Eccentricity, and structural holes. The node with the highest score is considered an influential node, as shown in Algorithm 2. These nine centrality measures give the rank of the most significant node from graph G. This effective node may be a carrier to spread the disease, information or send a signal to a large portion of graph G within a small amount of time [21].

Algorithm 2: Top keyword extraction from the semantic graph

INPUT: Edge List of the graph constructed and centrality = {DE, BE, CL, PR, EV, KC, EC, SH, CT}

OUTPUT: Top keywords extracted

1. Initialize **keyword set X**
2. **for all** keyword i in X do
3. **Compute the node score** of each keyword i

Degree $< - f_a = |\{n'|n' \in V \wedge \exists (n, n') \in E\}|$

Betweenness $< - BE_b = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$

PageRank $< - \pi_a^t = \sum_b \frac{\pi_b^{t-1}}{g_b} + \frac{(1-\propto)e_a}{m}$

Closeness $< - C_a =$ n-1/ $\sum_{v=1}^{n-1} d(x,y)$

K-Core <— K(a)=G_c

Eigen vector < – EA = (AX = λX)

Clustering Coefficient <— $CC_{a=} \frac{2e_a}{h_a(h_a - 1)}$

Eccentricity $\underset{a \neq b}{< -}$ $E_a = \max\{|j_{ab}|\}$

Structure Hole <— SH=$\sum b\,[1 - \sum_q p_{aq} m_{bq}]$, q≠a,b

4. **end for**
5. Use Collective Centrality Measure to extract the keywords

5.0 RESULT AND DISCUSSION

The authors of this paper used Python (version 3.7) for implementation and functions from NLP, Tweepy, Networks, NumPy, codecs packages. We executed the programs on a 64-bit PC with 8GB RAM and Intel Core i7-6700 CPU @ 3.40 GHz 8-core Processor running on Windows10. We used five benchmark datasets for empirical observations. Then, the graph and edge list were generated. Using this, we found nine-centrality measures viz., Degree, Betweenness, clustering coefficient, Closeness, Eccentricity, Eigenvector, K-core, PageRank, and Structural Holes. These are calculated and used to find Collective Centrality Measure. The last step is to find the keywords that are to be extracted.

5.1 Data Preprocessing And Semantic Keyword Extraction

We used the data extracted from Twitter as a real-time dataset**.** Kafka is the tool used to read streaming data from Twitter. It follows the publish-subscribe model where you write messages (publish) and read them (subscribe). Messages are grouped into topics. As messages are consumed, they are removed from Kafka. We carried out the experiments on the real-time Twitter datasets namely Lockdown,Corona, Economy, Technology and Citizenship Amendment Act-19. To extract tweets from Twitter API, we used the python Tweepy library. To use Twitter API, first, we created a Twitter Developer account. Then after creating an app, we got our API Keys and Access Tokens, which helped us retrieve data from Twitter. Each document extracted from Twitter was in JSON format, and it contained over 3000 lines of data. Then we reprocessed and extracted only the tweet text from the data to find the

keyword. The following topics were taken for this work: 1) Corona, 2) Lockdown, 3) Economy, 4) Technology, and 5) Citizenship Amendment Act (CAA). Each dataset contains tweets on a particular topic.

Steps in extracting the data:

- Creation of an App on the Twitter API Website. It gives the keys that are needed to use the Twitter Streaming API.
- Kafka-python and twitter-python are to be installed and used.
- Start Zookeeper and Kafka from the Kafka install directory. Then, the data on a particular topic can be accessed.
- Adaptive Lesk algorithm is used for the semantic relationship between two words, as shown in Table1. One of the most challenging tasks in text processing is identifying ways to improve accuracy.
- One of the most challenging tasks in text processing is identifying ways to improve accuracy.

Then, the noun-phrasing approach is applied to generate keywords to remove redundancy, as shown in Table 2.

Table 1: New keywords after processing through Improvised Lesk Algorithm for Five Datasets

DATASETS	KEYWORDS FROM LESK ALGORITHM
Lockdown	['rule', 'month', 'small', 'jar', 'better', 'left', 'deal', 'seen', 'wave', 'cancel', 'interview', 'depend', 'withdraw', 'impact', 'lose', 'way', 'share', 'inform', 'poster', 'help', 'wound', 'help', 'dog', 'dog', 'evil', 'news', 'hate', 'deaf', 'help', 'take', 'young', 'jumper', 'wait', 'dive', 'camp', 'drink', 'way', 'take', 'age', 'disconnect', 'learn', 'old', 'lower', 'form', 'reveal', 'impact', 'state', 'old', 'case', 'go', 'restrict', 'six', 'employ', 'better', 'feel', 'old', 'paper', 'crew', 'mate', 'holiday', 'lack', 'train', 'accept', 'option', 'follow', 'thousand', 'live', 'power', 'wealth', 'sport', 'night', 'express', 'nail', 'technician', 'spoke', 'stay', 'old', 'feel', 'prison', 'crime', 'divest', 'word', 'shield', 'back', 'list', 'man', 'refrain', 'ill', 'step', 'test', 'contract', 'school', 'front', 'case', 'birth', 'breakdown', 'week', 'case', 'birth', 'amateur', 'birth', 'push', 'ill', 'mean', 'month', 'blur', 'teach', 'mode', 'teach', 'help', 'care', 'say', 'intent', 'want', 'herd', 'hair', 'today', 'tomorrow', 'restrict', 'rise', 'ten', 'rise', 'blow_up', 'rule', 'record', 'old', 'outbreak', 'man', 'guardian', 'common', 'lunch', 'common', 'bit', 'fight', 'test', 'process', 'district', 'see', 'return', 'suppress', 'outbreak', 'region', 'outbreak', 'plant', 'distress', 'drug', 'camp', 'show', 'seek', 'begin', 'sport', 'bouffant', 'want', 'hair', 'news', 'record', 'rise', 'blow_up', 'rule', 'warn', 'month', 'write', 'matter', 'core', 'core', 'aid', 'news', 'lord', 'three', 'month', 'return', 'lose', 'news', 'counsel', 'session', 'daughter', 'keep', 'coat', 'refresh', 'time', 'throw', 'home', 'time', 'outbreak', 'meat', 'author', 'express', 'share', 'hate', 'music', 'prevent', 'tag', 'express', 'sit', 'express', 'outbreak', 'meat', 'time', 'right', 'govern', 'threat', 'refrain', 'mass', 'test', 'give', 'team', 'see']
CORONA	['fund', 'money', 'hold', 'cure', 'case', 'east', 'case', 'hear', 'say', 'trend', 'want', 'perform', 'time', 'question', 'trump', 'visa', 'ban', 'go', 'visa', 'close', 'merchant', 'account', 'speak', 'treatment', 'hear', 'say', 'leader', 'case', 'clinic', 'clinic', 'trial', 'save', 'hold', 'rank', 'share', 'map', 'defend', 'mar', 'case', 'gather', 'break', 'govern', 'succeed', 'protect', 'soldier', 'said', 'cure', 'give', 'despair', 'baba', 'treatment', 'kit', 'cure', 'year', 'prevent', 'see', 'park', 'rent', 'cough', 'strong', 'forget', 'want', 'opinion', 'trump', 'world', 'avoid', 'outbreak', 'baba', 'treatment', 'kit', 'cure', 'conduct', 'close', 'parliament', 'camp', 'mode', 'transport', 'area', 'trend', 'keep', 'aunt', 'symptom', 'problem', 'listen', 'scientist', 'kill', 'utensil', 'baba', 'go', 'cure', 'video', 'cure', 'save', 'world', 'camp', 'treatment', 'inform', 'baba', 'treatment', 'kit', 'cure', 'safe', 'parliament', 'blood', 'plasma', 'donor', 'team', 'ab', 'affirm', 'affirm', 'inform', 'conduct', 'ward', 'par', 'want', 'parent', 'left', 'start', 'listen', 'god', 'saw', 'left', 'thing', 'research', 'scientist', 'camp', 'cure', 'virus', 'baba', 'thing', 'lose', 'despair', 'cure', 'evil', 'mother', 'parent', 'special', 'treat', 'baba', 'treatment', 'kit', 'cure', 'danger', 'survey', 'shoe', 'lost', 'follow', 'session', 'night', 'cite', 'name', 'treat', 'patient', 'evil', 'old', 'baba', 'treatment', 'kit', 'cure', 'hand', 'thing']
Economy	['govern', 'parliament', 'medium', 'scare', 'attempt', 'pitch', 'defend', 'job', 'fact', 'state', 'ebb', 'close', 'daughter', 'predict', 'trump', 'shift', 'home', 'ill', 'prevent', 'evil', 'go', 'plan', 'old', 'clean', 'ill', 'want', 'hear', 'thought', 'help', 'year', 'theft', 'tell', 'tighten', 'restrict', 'close', 'open', 'case', 'group', 'case', 'group', 'gather', 'communist', 'china', 'way', 'world', 'begin', 'way', 'world', 'begin', 'way', 'world', 'begin', 'bang', 'past', 'data', 'expect', 'data', 'expect', 'withdraw', 'drug', 'kill', 'ten', 'thousand', 'kill', 'drug', 'trade', 'trump', 'open', 'matter', 'fraud', 'open', 'hire', 'mouth', 'china', 'see', 'map', 'tell', 'bitt', 'brief', 'camp', 'recess', 'go', 'front', 'flaw', 'china', 'save', 'four', 'year', 'take', 'home', 'freedom', 'parent', 'harm', 'read', 'six', 'east', 'gather', 'communist', 'china', 'north', 'rate', 'swell', 'year', 'correct', 'year', 'opinion', 'bottom', 'food', 'retail', 'base', 'daughter', 'away', 'medium', 'scare', 'diver', 'poor', 'contract', 'medium', 'scare', 'medium', 'scare', 'gather', 'communist', 'china', 'number', 'sell', 'forget', 'minor', 'forget', 'fraud', 'open', 'hire', 'doom', 'succeed', 'sum', 'east', 'go', 'folk', 'mean', 'danger', 'rig', 'trade', 'protect', 'fuel', 'market', 'impact', 'world', 'council', 'advisor', 'bottom', 'trump', 'rise', 'black', 'part', 'gather', 'communist', 'china', 'give', 'home', 'guy', 'help', 'love', 'bit', 'day', 'expect', 'cabin', 'pull', 'say', 'voter', 'forget', 'time', 'trump', 'settlement', 'man', 'graver', 'vote', 'threat', 'warn', 'complex', 'way', 'gather', 'safe', 'shift', 'divest', 'bottom', 'trump', 'rise', 'black', 'part', 'novel', 'project', 'leadership', 'project', 'debt', 'lead', 'review', 'trump', 'support', 'hand', 'trump', 'cage', 'parent', 'protest', 'attack', 'store', 'couch', 'shift', 'sign', 'maker', 'eye', 'invest', 'see', 'dull', 'market', 'benefit', 'camp', 'day', 'fish', 'monger', 'exempt', 'danger']

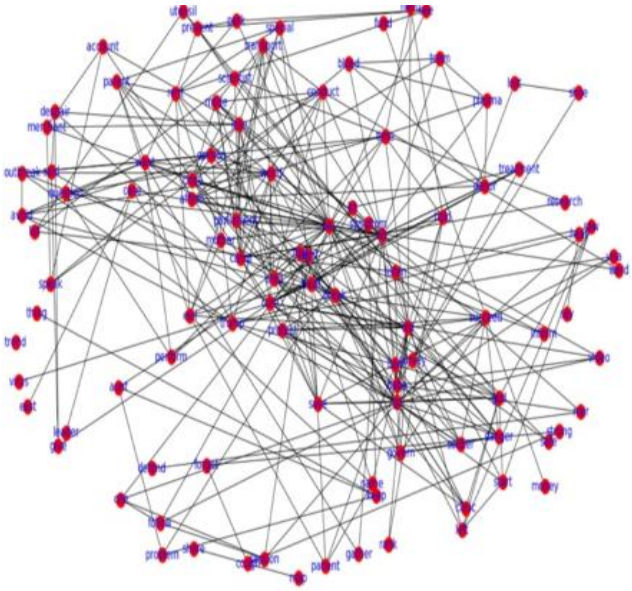
Technology	['help', 'fight', 'gain', 'contain', 'copyright', 'defend', 'glove', 'fight', 'pull', 'regard', 'use', 'thread', 'data', 'buy', 'china', 'thread', 'suffer', 'six', 'exempt', 'evil', 'follow', 'salt', 'man', 'express', 'govern', 'area', 'state', 'farmer', 'tax', 'project', 'retreat', 'end', 'old', 'even', 'ritual', 'help', 'record', 'short', 'record', 'short', 'adopt', 'dot', 'disconnect', 'east', 'east', 'old', 'tool', 'phone', 'set', 'lower', 'user', 'camp', 'user', 'partner', 'player', 'event', 'step', 'step', 'danger', 'lose', 'follow', 'market', 'sad', 'see', 'read', 'thread', 'three', 'time', 'remind', 'trump', 'ban', 'card', 'state', 'begin', 'year', 'take_away', 'partner', 'past', 'danger', 'thing', 'test', 'problem', 'want', 'help', 'custom', 'system', 'close', 'rat', 'protect', 'racket', 'sale', 'bottom', 'secret', 'camp', 'dot', 'disconnect', 'east', 'east', 'scare', 'ill', 'perform', 'camp', 'state', 'back', 'traffic', 'kill', 'three', 'zero', 'night', 'read', 'common', 'follow', 'platform', 'prevent', 'iowa', 'seek', 'inform', 'state', 'stack', 'function', 'help', 'fight', 'gain', 'contain', 'govern', 'build', 'help', 'gather', 'well', 'fun', 'wonder', 'system', 'thread', 'give', 'master', 'attract', 'chairman', 'label', 'trump', 'video', 'medium', 'crack', 'post', 'hear', 'lot', 'inferior', 'way', 'rise', 'crash', 'research', 'team', 'week', 'test', 'old', 'trace', 'tool', 'phone', 'set', 'job', 'eastern', 'east', 'medium', 'size', 'firm', 'unit', 'end', 'old', 'even', 'ritual', 'adopt', 'stupid', 'system', 'use', 'wire', 'tailor', 'water', 'monitor']
Citizenship Amendment Act-19	['voter', 'proof', 'vote', 'prevent', 'forget', 'black', 'call', 'mean', 'drug', 'cartel', 'partner', 'visa', 'visa', 'visa', 'visa', 'visa', 'thread', 'year', 'away', 'dress', 'crush', 'handbook', 'book', 'store', 'handbook', 'train', 'drug', 'cartel', 'partner', 'protest', 'bill', 'vote', 'left', 'camp', 'period', 'wall', 'cancel', 'east', 'back', 'say', 'divest', 'heaven', 'year', 'go', 'visa', 'year', 'visa', 'sought', 'away', 'dress', 'crush', 'use', 'child', 'child', 'document', 'minor', 'award', 'dishonor', 'year', 'post', 'award', 'mean', 'descent', 'canon', 'front', 'thousand', 'age', 'grow', 'enrich', 'camp', 'left', 'vote', 'dollar', 'plan', 'drug', 'cartel', 'partner', 'thread', 'report', 'child', 'child', 'document', 'minor', 'give', 'use', 'keep', 'friendship', 'aid', 'help', 'sharpen', 'say', 'terrorist', 'art', 'counterpoint', 'traitor', 'refresh', 'hour', 'feel', 'freedom', 'refresh', 'say', 'care', 'year', 'dump', 'keep', 'see', 'prevent', 'go', 'visa', 'year', 'visa', 'sought', 'parent', 'id', 'card', 'drug', 'cartel', 'partner', 'mean', 'take', 'feel', 'away', 'dress', 'crush', 'visa', 'visa', 'visa', 'visa', 'visa', 'follow', 'withdraw', 'nation', 'comment', 'nation', 'take', 'keep', 'campaign', 'kite', 'sale', 'good', 'give', 'drug', 'cartel', 'partner', 'child', 'child', 'document', 'minor', 'go', 'visa', 'year', 'visa', 'sought', 'go', 'visa', 'year', 'visa', 'sought', 'go', 'visa', 'year', 'visa', 'sought', 'warm', 'opinion', 'prevent', 'short', 'benefit', 'child', 'child', 'document', 'minor', 'see', 'woman', 'call', 'bill', 'law']

Table 2: Keywords from Noun phrasing approach for five datasets

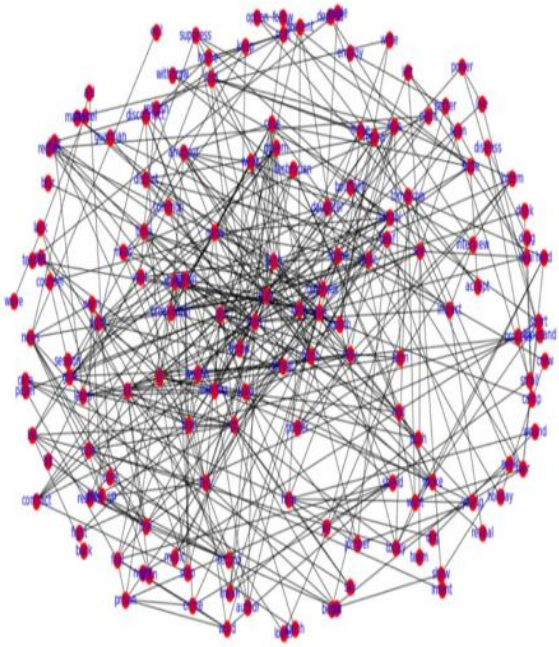
DATASETS	KEYWORDS FROM NOUN PHRASING APPROACH
Lockdown	Month, jar, deal, wave, interview, way, poster, wound, dog, news, deaf, age, paper, crew, technician, man, school, today, tomorrow, guardian, lunch, district, plant, coat, daughter, session, lord, meat, author, tag, team
CORONA	Fund, money, case, time, question, trump, visa, merchant, account, leader, clinic, map, soldier, kit, opinion, treatment, parliament, camp, node, area, aunt, symptom, scientist, utensil, world, blood, plasma, donor, ward, parent, research, virus, thing, mother, survey, shoe, session, nig0ht, baba
Economy	Parliament, medium, pitch, job, daughter, trump, theft, case, group, communist, china, drug, fraud, month, map, flaw, folk, fuel, market, council, advisor, black, home, guy, cabin, voter, settlement, graver, novel, leadership, hand, store
Technology	Copyright, glove, thread, data, china, six, evil, man, area, farmer, state, tax, project, ritual, tool, phone, player, partner, danger, market, trump, racket, system, traffic, platform, stack, function, label, chairman, video, medium, research, team, job, firm, tailor, water, monitor
Citizenship Amendment Act-19	Voter, proof, drug, cartel, partner, visa, dress, store, handbook, bill, heaven, year, child, document, award, age, mean, minor, dollar, plan, friendship, terrorist, freedom, id, card, parent, nation, campaign, opinion, women, traitor, counterpoint

5.2 Semantic Co-occurrence Graph Construction

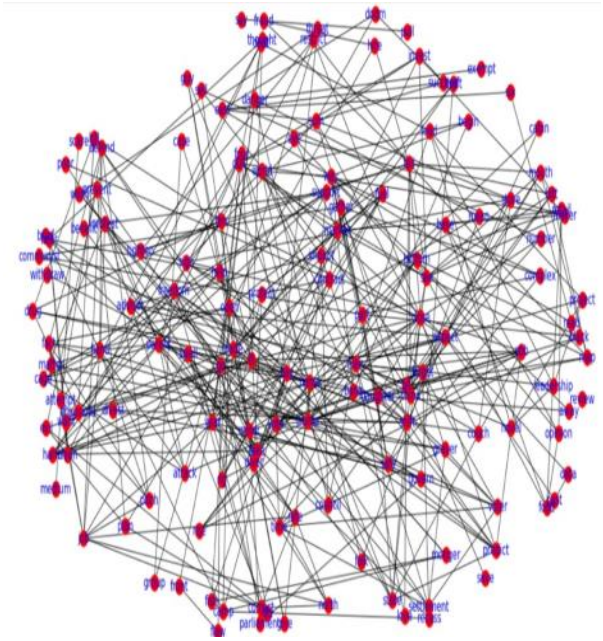
The output from the noun phrasing approach is the critical input to the graph-building phase. Each word in Table 2 is considered as a node and semantic relationship between the nodes as a link in the graph. The elimination of words based on semantic analysis is a significant task in the co-occurrence analysis phase. An example of a co-occurrence graph for five datasets can be seen in Fig 3. These semantic co-occurrence graphs are undirected and well-suited for flat node visualizations between people and concepts. To find the co-occurrence of words in the semantic keywords, we can use bigrams from NLTK. The words in the noun phrasing approach are considered nodes of the graph, as shown in Figure 3



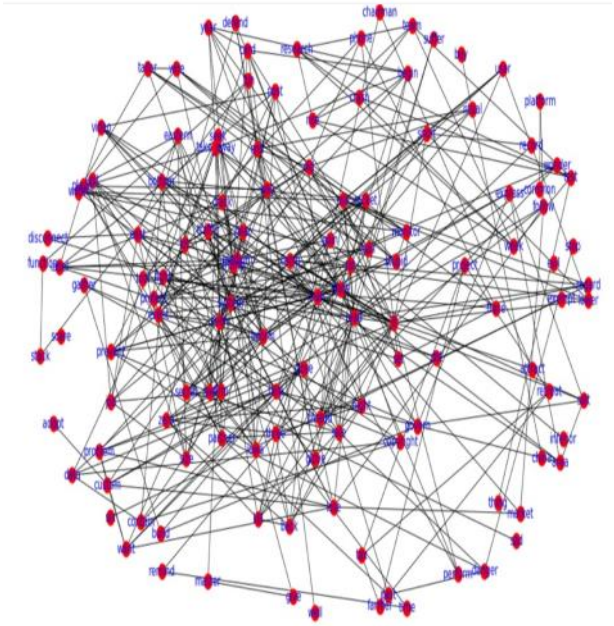
a. Lockdown dataset



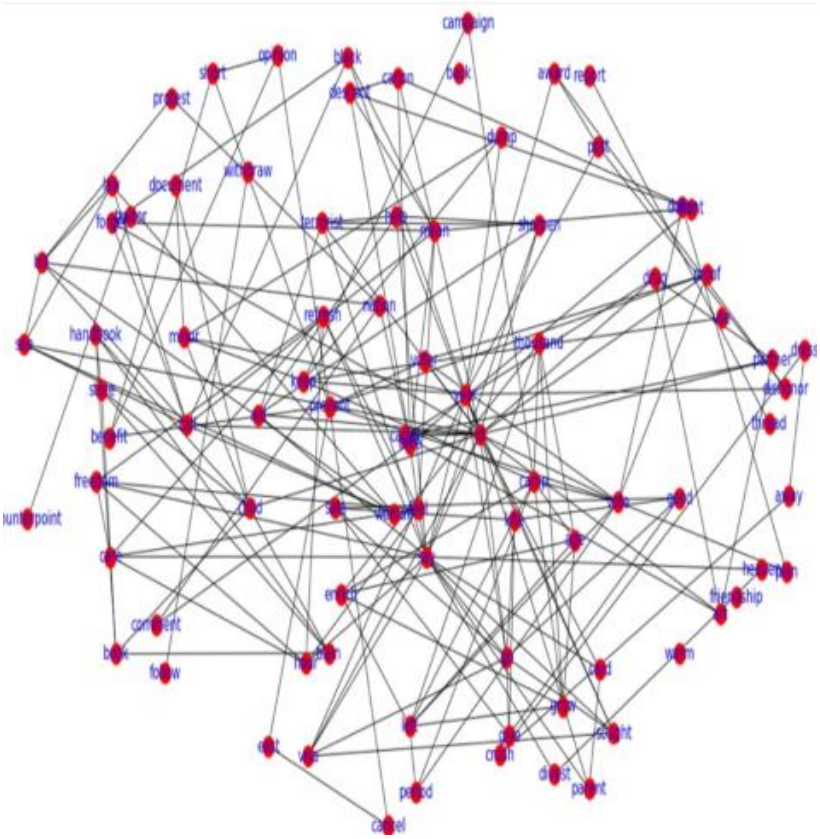
b. Corona dataset



c. Economy dataset



d. Technology dataset



e. Citizenship Amendment Act dataset

Fig 3: Co-occurrence graph representation for the five datasets, namely Lockdown, Corona, Economy, Technology and Citizenship Amendment Act.

5.3 Node Collective Centrality Index (CCI) Calculation

The node weight is calculated [16] as edge betweenness value using the MCI approach, which includes nine centrality measures i.e. betweenness centrality, eigenvector centrality, closeness centrality, and node centrality. The obtained weight is sent into the PageRank algorithm for ranking the keywords, and all these values are normalized using min-max normalization.

Table 3: Node weights for the top 10 keywords of five Datasets

Datasets	Top 10 keywords	Node weight using CCI Approach
Lockdown	"['stay',	1.0
	'men',	0.9712
	'old',	0.9637
	'ill',	0.9515
	'test',	0.9461
	'outbreak',	0.9409
	'time',	0.9279
	'news',	0.9276
	'case',	0.9168
	breakdown']"	0.9019
Corona	"['old',	1.0
	'cure',	0.9812
	'lose',	0.9754
	'treat',	0.9704
	'god',	0.9732
	'world ',	0.9668
	'save ',	0.9657
	'evil',	0.9558
	'transport ',	0.9454
	' ban']"	0.9271
Economy	"['trump',	1.0
	'man',	0.9891
	'home',	0.9731
	'Shift',	0.9697
	' expect',	0.9602
	'china',	0.9597
	'market',	0.9432
	' trade',	0.9335
	'job',	0.9325
	'world ']"	0.9262
Technology	"['man ',	1.0
	'system ',	0.9846
	'state ',	0.9769
	'help ',	0.9748
	'inform ',	0.9664
	'fight ',	0.9581
	'lose ',	0.9538
	'partner ',	0.9504

	‘govern ‘,	0.9444
	‘custom ‘]”	0.9416
CAA	“[‘prevent ‘,	1.0
	‘year ‘,	0.9841
	‘vote ‘,	0.9821
	‘camp ‘,	0.9667
	‘call ‘,	0.9576
	‘child ‘,	0.9569
	‘bill ‘,	0.9481
	‘minor ‘,	0.9333
	‘withdrawal‘,	0.9245
	’traitor ‘]”	0.9174

Table 3 contains the list of top keywords with their normalized weights that are calculated by finding node values for each centrality measure of the keywords. Then by using the CCI approach, a subgroup of centrality measures is created. Next, by using this subgroup, top keywords along with their weights are calculated

5.4. Performance Evaluation:

After finding the top node keywords, performance is analyzed using evaluation metrics such as precision, recall and F1-Score as shown in Table 4. All the individual performance evaluation metrics for each dataset are compared among three different approaches, namely the sCAKE method, Multi Centrality Index (MCI) method without Semantic Analysis, and the proposed SKEM++. These three methods are compared using a bar chart for a better understanding.

From Table 4 and the performance measures, it is evident that Collective Centrality Index (CCI) approach using Semantic Analysis (SKEM++) has an efficiency of nearly 80%. Whereas Multi Centrality method without Semantic Analysis has nearly 70% efficiency and Semantic Analysis without Collective Centrality method has around 60% efficiency. Therefore, the CCI approach using Semantic Analysis is better in manipulating and discovering important document keywords.

Table: 4 Node weights for the top 10 keywords of five Datasets

Datasets	SCAKE Method			Multi Centrality Approach			**Proposed SKEM++**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Lockdown	0.61	0.62	0.61	0.71	0.71	0.7	**0.83**	**0.81**	**0.8**
Corona	0.69	0.63	0.66	0.78	0.74	0.76	**0.88**	**0.84**	**0.86**
Economy	0.71	0.68	0.69	0.73	0.7	0.7	**0.82**	**0.8**	**0.8**
Technology	0.71	0.65	0.68	0.72	0.73	0.72	**0.81**	**0.83**	**0.82**
CAA	0.75	0.62	0.68	0.76	0.71	0.73	**0.86**	**0.81**	**0.83**

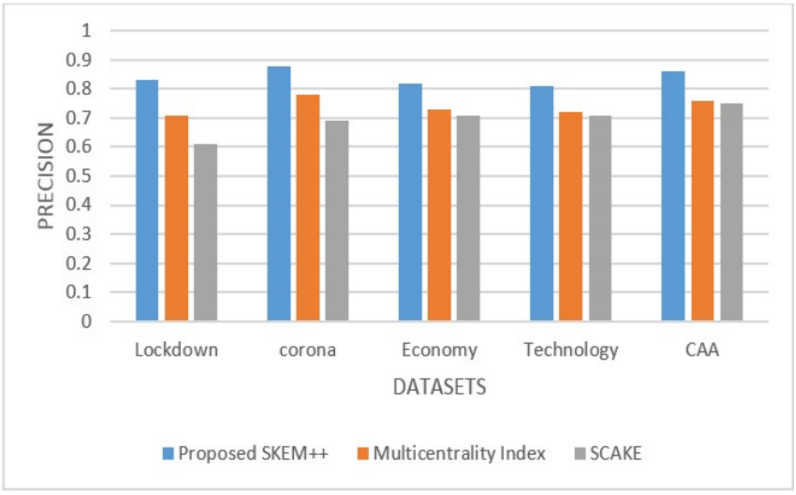


Fig 4: Comparison of different approaches using Precision

From Figure 4 it is seen that SKEM++ achieves the highest precision value of 0.83 for keywords like Lockdown, Technology and 0.89 for other keywords like Corona and CAA, respectively and .80 for Economy and Technology. In the Multi-Centrality Index, the precision value is 0.70 for lockdown for Corona and 0.76 for CAA, whereas 0.70 for Lockdown, Technology, and Economy. The highest recall value for the proposed SKEM++ is on an average of 0.8 for all five datasets, as shown in Figure 5. The other two methods, namely the Multi Centrality Index and sCAKE, are 0.7 and 0.6, respectively. Figure 6. indicates that the proposed SKEM++ has the highest F- measure compared to other existing methods.

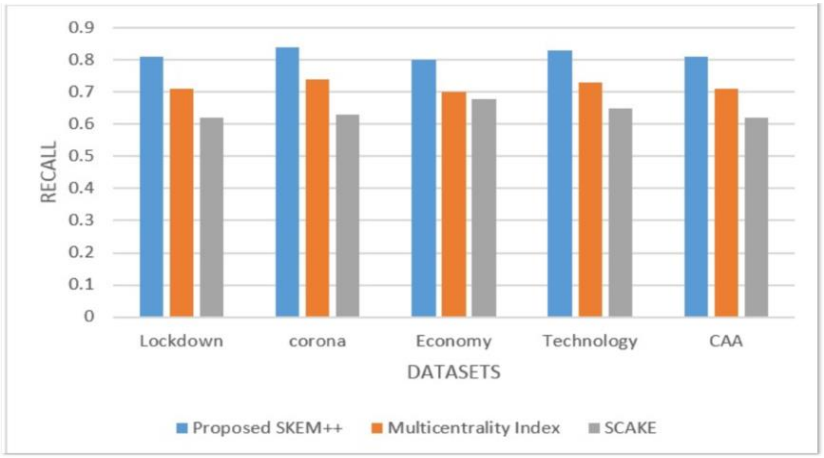


Fig 5: Comparison of different approaches using Recall

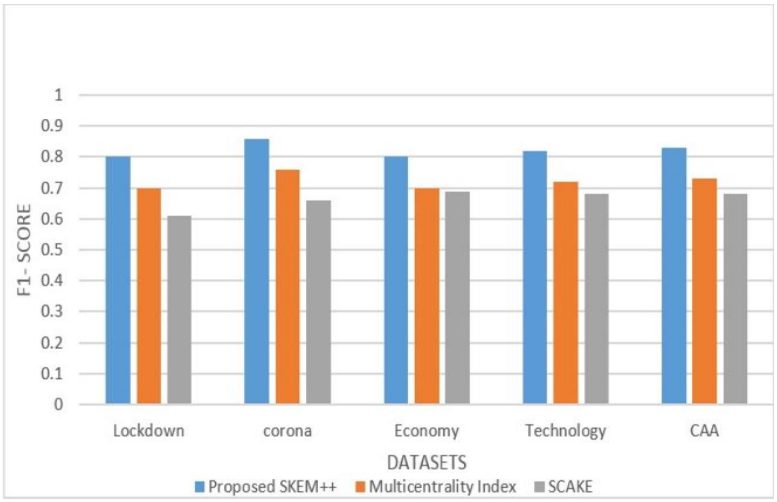


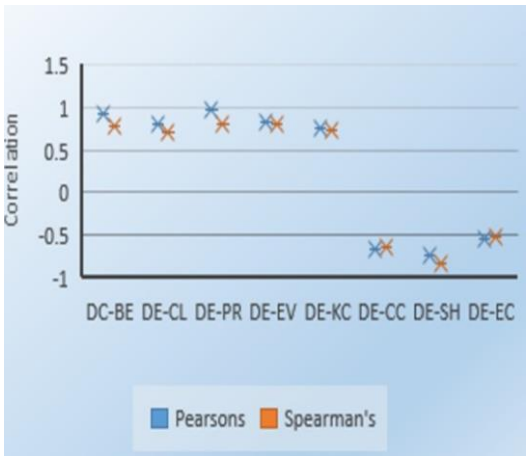
Fig 6: Comparison among different approaches using F1-Score

Table 5: Time complexity for different centrality measures

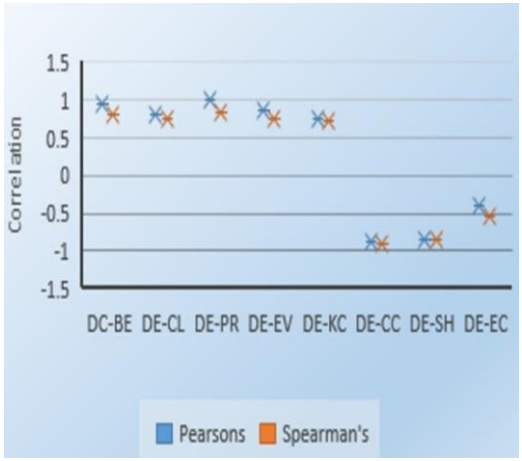
Centrality Measures	**Time complexity**
Degree (DE)	O (n)
Clustering Coefficient (CC)	O (n*<g>^{2})
Structural Holes (SH)	O (n+n*(<g>)^2)
Eigenvector (EV)	O (n^2)
PageRank (PR)	O (n+e)
K-Core (KC)	O (n+e)
Eccentricity (EC)	O (n*e)
Closeness (CL)	O (n*e)
Betweenness (BE)	O (n*e)

Table 5 shows the various centralities measured along with their time complexity. Average path calculation measures have the highest computational cost when compared to other efforts. The measures like clustering coefficient and structural holes are more appropriate for real-time problems were the significant graph size. This is because these measures focus only on particular vertices that is a portion of the graph. These measures also help to interact with the entire graph. Also, while computing centrality measures, a small number of vertices and time complexity descends to the neighborhood of the chosen vertices in a given graph. However, the degree centrality measure gives flat time complexity, and this implies that degree is the most critical measure for various situations.

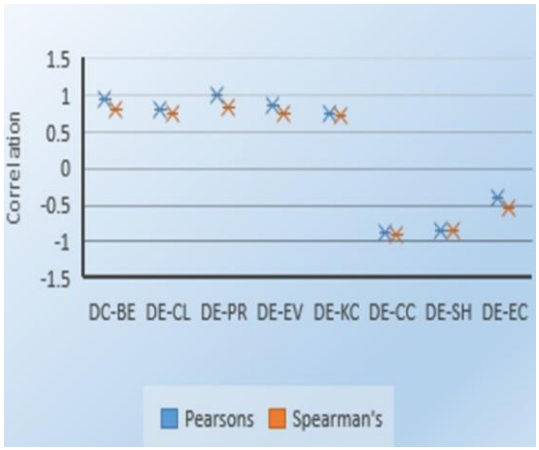
Pearson's and Spearman's correlation for the pairwise gradation with additional centrality measures validating this approach. Pearson correlation is shown in blue color and Spearman's correlation in red color in Fig.7. Pearson's coefficient is higher than Spearman in Lockdown, Corona, Economy, Technology, and CAA datasets. K-core presents a high correlation to a degree since it has more cores, whereas eccentricity gave less correlation with the degree.



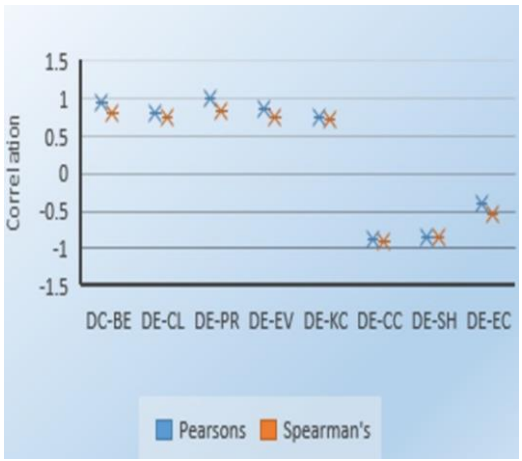
(a) Lockdown Datasets



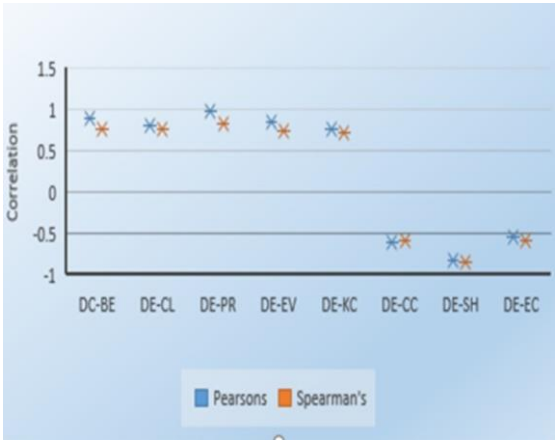
(b) Corona Datasets



(c) Economy Datasets



(d) Technology Datasets



(e) CAA Datasets

Fig. 7: Two kinds of Correlation distribution between Degree and other centrality measures

6.0 CONCLUSION AND FUTURE WORK

In this paper, a novel method called SKEM++ uses CCM measure. Using five different datasets that are collected from Twitter of other domains, the authors of this paper achieved the desired results. It has been demonstrated that a grouping of centrality attributes is more efficient to get enhanced effects by providing semantic analysis using the Improvised Lesk Algorithm. In this algorithm, we used WordNet dictionary to retrieve the meaning, and it is compared with context intention to give the best sense. Words with the best understanding are considered semantic keywords. As a result, this combination of approaches shows a very substantial variance in evaluation to the individual centrality approach. The performance of this method is evaluated using evaluation metrics, i. e. Precision, Recall, and F1-score for each dataset, which is nearly 10% more efficient than using a multi-centrality index[18] without using semantic approach. At the same time, the proposed SKEM++ is 20% more efficient than sCAKE [17] methods. Selection methods and Semantic Analysis are applied to obtain a meaningful context. In the future, it is essential to develop ways that can understand and analyze multimodal big social data to improve the performance of prediction systems.

The main demerits of the adaptive Lesk algorithm are its exponential complexity: This means that the increase in the number of comparisons increases the number of words to disambiguate in the text. Then it is susceptible to the definition of the words. Since words used in the report should not overlap within the context, the other main problem with the improvised Lesk algorithm is overlapping. Deep learning techniques and large sense-annotated datasets are increasingly crucial for supervised training systems to address this problem.

7.0 ACKNOWLEDGMENT

The authors gratefully acknowledge the Science and Engineering Research Board (SERB), Department of Science & Technology, India for financial support through Mathematical Research Impact Centric Support (MATRICS) scheme (MTR/2019/000542). The authors also acknowledge SASTRA Deemed University, Thanjavur for extending infrastructural support to carry out this research work.

ABBREVIATIONS

The following abbreviations are used in this manuscript:

Abbreviation	Description
API	Application Programming Interface
BE	Betweenness
CC	Clustering coefficient
CCM	Collective Centrality MeasuCCre
CL	Closeness
DE	Degree
EC	Eccentricity
EV	Eigenvector
JSON	JavaScript Object Notation
KC	K-Core
MCI	Multi Centrality Index
NLP	Natural Language Processing
OSN	Online Social Network
PR	Page Rank
SNA	Social Network Analysis
SH	Structural Holes

REFERENCES

[1] Tripathy, R. M., Bagchi, A., & Mehta, S. (2010, October). *A study of rumor control strategies on social networks*. In Proceedings of the 19th ACM international conference on Information and knowledge management, (pp. 1817-1820).

[2] Awan, M. N., & Beg, M. O. (2021). *Top-rank: a topical postion rank for extraction and classification of key phrases in text.* Computer Speech & Language, 65, 101116.

[3] Blanco, R., & Lioma, C. (2012). *Graph-based term weighting for information retrieval*. Information retrieval, 15(1), 54-92.

[4] Bookstein, A., & Swanson, D. R. (1974). *Probabilistic models for automatic indexing*. Journal of the American Society for Information science, 25(5), 312-316.

[5] Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual web search engine*. Computer networks and ISDN systems, 30(1-7), (pp 107-117).

[6] Fernandez-Basso, C., Francisco-Agra, A. J., Martin-Bautista, M. J., & Ruiz, M. D. (2019). *Finding tendencies in streaming data using big data frequent itemset mining.* Knowledge-Based Systems, 163, 666-674.

[7] Florescu, C., & Caragea, C. (2017, February). *A position-biased pagerank algorithm for keyphrase extraction*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

[8] Harter, S. P. (1975). *A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing*. Journal of the American Society for Information Science, 26 (5), 280-289.

[9] Kraft, D. H. (1985). *Advances in Information Retrieval: Where Is That/#* &@¢ Record*. In Advances in computers (Vol. 24, pp. 277-318). Elsevier.

[10] Litvak, M., Last, M., Aizenman, H., Gobits, I., & Kandel, A. (2011). *DegExt—A language-independent graph-based keyphrase extractor*. In Advances in intelligent web mastering–3 (pp. 121-130). Springer, Berlin, Heidelberg.

[11] Luhn, H. P. (1957). *A statistical approach to mechanized encoding and searching of literary information.* IBM Journal of research and development, 1(4), 309-317.

[12] Matsuo, Y., Ohsawa, Y., & Ishizuka, M. (2001, November). *Keyworld: Extracting keywords from documents small world.* In International conference on discovery science. (pp. 271-281). Springer, Berlin, Heidelberg.

[13] Mihalcea, R., & Tarau, P. (2004, July). *Textrank: Bringing order into text*. In Proceedings of the 2004 conference on empirical methods in natural language processing, (pp. 404-411).

[14] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). *Kea: Practical automated keyphrase extraction*. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, (pp. 129-152).

[15] Y. Ohsawa, Y., Benson, N. E., & Yachida, M. (1998, April). *KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor.* In Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98, (pp. 12-18). IEEE.

[16] Devika, R., & Subramaniyaswamy, V. (2019). *A semantic graph-based keyword extraction model using ranking method on big social data*. Wireless Networks, 1-13.

[17] Duari,S., & Bhatnagar,V. (2019). *sCAKE: semantic connectivity aware keyword extraction*. Information Sciences, 477, 100-11.

[18] Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., & Berton, L. (2019). *A multi-centrality index for graph-based keyword extraction.* Information Processing & Management, 56 (6), 102063.

[19] Tixier, A. J. P., Malliaros, F. D., & Vazirgiannis, M. (2016). *A graph degeneracy-based approach to keyword extraction*. In J. Su, X. Carreras, & K. Duh (Eds.).

[20] Beliga, S., Mestrovic, A., & Martincic-Ipsic, S. (2014). *Toward selectivity-based keyword extraction for croatian news*. Proceedings of the workshop on surfacing the deep and the social web co-located with the 13th international semantic web conference (ISWC).

[21] Tsatsaronis, G., Varlamis, I., & Nørvåg, K. (2010*). Semanticrank: Ranking keywords and sentences using semantic graph*s. Proceedings of the 23rd international conference on computational linguistics – COLING, 1074–1082.

[22] S.B. Seidman , *Network structure and minimum Degree*, Soc. Netw. 5 (3) (1983) 269–287.

[23] J. Cohen , *Trusses: cohesive subgraphs for social network analysis*, National Security Agency Technical Report, 2008