Text-Based Analysis of COVID-19 Comments Using Natural Language Processing

Kanchan Naithani(✉), Y. P. Raiwani, and Rajeshwari Sissodia

Department of Computer Science and Engineering, HNB Garhwal University, Garhwal, Srinagar, Uttarakhand, India

Abstract. In dialectology, Natural Language Processing is the process of recognizing the various ontologies of words generated in human language. Various techniques are used for analyzing the corpus from naturally generated content by users on various platforms. The analysis of these textual contents collected during the COVID-19 has become a goldmine for marketing experts as well as for researchers, thus making social media comments available on various platforms like Facebook, Twitter, YouTube, etc., a popular area of applied artificial intelligence. Text-Based Analysis is measured as one of the exasperating responsibilities in Natural Language Processing (NLP). The chief objective of this paper is to work on a corpus that generates relevant information from web-based statements during COVID-19. The findings of the work may give useful insights to researchers working on Text analytics, and authorities concerning to current pandemic. To achieve this, NLP is discussed which extracts relevant information and comparatively computes the morphology on publicly available data thus concluding knowledge behind the corpus.

Keywords: NLP: Natural Language Processing · NLU: Natural Language Understanding · TBA: Text-Based Analysis · Knowledge Representation · WSD: Word Sense Disambiguation

1 Introduction

In the text-based analysis, corpus data characterize the "lexical", "syntactical", "semantic", and "positional" relationship with each other. The corpus vector is a relevant term related to Feature extraction emphasizing Ontologies that provide vital information regarding relations between the words [1].

The work done will provide an innovational pathway for the considerate indulgence of "Knowledge Extraction" from the corpus' perception. The conclusions will help in gaining insights into the most important topics discussed during the health and economic crisis based on the most frequent words used on Social Media Platforms. The hot topics deduced after the analysis of comments will help to know how well the people are handling their situations amidst the pandemic.

A. Dev et al. (Eds.): AIST 2021, CCIS 1546, pp. 195–209, 2022.
https://doi.org/10.1007/978-3-030-95711-7_17

2 Related Work

2.1 Recent Research Conducted on COVID19

In literature, quite a few procedures are presented to analyze ontologies that comprehend lexical relationships with the corpus and define how feature vectors are related with the individual words using words association [2–4]. Over the years, Natural Language Processing research has been done with English language orientation. The reason being, according to a survey, the total number of active users on various platforms use English, for expressing views on a global platform [8, 20].

Novel COVID -19 has created a mesh of digital data, where researchers and many help care centers are trying to understand the hot topics discussed by the public via various social media platforms. The most recent researches conducted by researchers for analyzing the web-based content after COVID-19's outburst are mentioned in Table 1. It was observed that while working with NLP, much of the statements present online show personal opinions in the majority and the research activities are performed on subjective datasets [19].

Table 1. Work done on Covid19 using NLP

Ref. no.	Researchers	Year	Work done & tools	Parameters	Outcome	Limitation
[8]	Güngör & Üsküdarli,	2019	"The effect of morphology in named entity recognition with sequence tagging"	Sequence tagging, Character-based embeddings, F1-measure	Augmenting word representations with morphological embeddings improves NER performance, which is further improved when combined with "character-based word representations"	Parameters chosen could have shown more clarity in context with the multiple languages used as the data set
[1]	Pittaras, Giannakopoulos, Papadakis and Karkaletsis	April. 2020	"Text classification with semantically enriched word Embeddings"	Semantic Frequency Vector, Accuracy, macro F1- score, TF-IDF normalized weights	The use of semantic information from resources such as "WordNet" significantly progresses the performance for classification	TF-IDF behavior observed in frequency based semantic vectors was not decent

(*continued*)

Table 1. (*continued*)

Ref. no.	Researchers	Year	Work done & tools	Parameters	Outcome	Limitation
[3]	Kusum Lata, Pardeep Singh, Kamlesh Dutta	Oct., 2020	"A comprehensive review on feature set used for anaphora resolution"	Anaphora Resolution Antecedent	The review presented an understanding of solving AR problems from the perspective of feature selection	Deeper level research work on positional relations of the features was absent
[4]	Klaifer Garcia & L. Berton	Dec., 2020	"Topic Detection and sentiment analysis in Twitter"	Precision and F1 Score	Mostly negative Polarity was observed in the LR and RF and Linear SVM	A more detailed analysis of the feelings was possible
[15]	Matteo Cinelli, Walter Quattrociocchi et al	Oct., 2020	"The COVID-19 social media infodemic"	stochastic gradient descent, word prediction vector, vocabulary size,	Intercept, coefficient, R^2 for various primary datasets were deduced for different datasets	Lexical and Semantic analysis could have been improved w.r.t. the information extracted

Limaye et al. studied some concerns regarding the misrepresentation of situations, information and even statistics on social media especially during a pandemic like COVID-19 and displayed them in a commentary article [12]. The largest social media in China, WeChat, was analyzed by Lu and Zhang to identify the trends referencing COVID-19 [13].

Research in the field of observing COVID -19 situation and its effect on people has created many influential topics, as shown in Fig. 1 that gained mass attention. An online survey conducted on online available social platforms like Facebook, YouTube, Twitter, Instagram, blogging sites and many official web discussing forums has shown that people concerned with the current epidemic have analyzed people's opinions in reference to these topics around the world including India to understand the lenient or harsh situations for them [13–15].

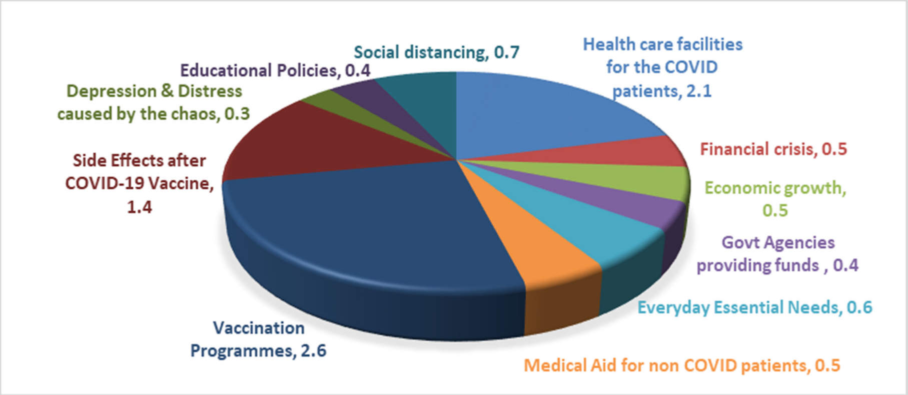


Fig. 1. Influential topics w.r.t. COVID-19 that gained mass attention [3, 4, 9–11, 14, 15]

3 Text-Based Analysis

NLP and Text Analysis techniques are generally practiced to recognize and extract information that is subjective in nature from a piece of text or corpus and this practice is known as Text-Based Analysis (TBA). It refers to the use of "computational linguistics" and "ontology-based analysis" to analytically categorize, extract, quantify, and study varying circumstances along with subjective information [5]. It is an evolving subject, which challenges the analysis and measurement of human language and transforms them into hard facts for enlightening the real, factual or cynical meaning behind the words [15, 21]. The pipeline for TBA can be achieved using Seven Basic steps as shown in Fig. 2:

a) **Identification of Language:** Different idiosyncrasies are present in multiple languages, that's why it is indeed a critical aspect to know what Language and what grammatical features we're dealing with. It involves predicting the natural language of the text by observing the features of the grammar that will be responsible for the other text analytics function. Approaches like Short Word Based Approach. Frequent Word-Based Approach and N-Gram Based Approach are used to Identify the Language of the corpus elements [6].

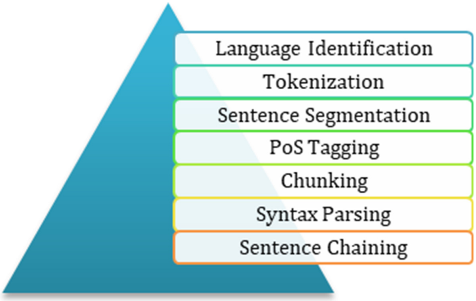


Fig. 2. Steps for TBA

b) **Tokenization:** After the language of the text is known, it can be broken up into Tokens, the basic entities of the connotation that are operated on, and this process of breaking down of the corpus into smaller unique entities is known as Tokenization. Tokens can be words, phonemes, punctuations, hyperlinks or any other smallest component of the grammar. For example, an English sentence made up of 5 words, may contain 5 tokens. Tokenization depends on the characteristics of a language, and each language has varied requirements for tokenization. For instance, in English, practices "white space" and "punctuation" are used for breaks, and could be tokenized without putting much effort. Most languages based on alphabets follow comparatively simple approaches to break the corpus. So, rules-based tokenization is prominently used for alphabetic Languages [6, 7].

c) **Sentence Segmentation:** Also known as Sentence Tokenization is the process of separating a sequence of written language into its component sentences. Once the tokens are identified, places where sentences end can be easily pointed. In order to run more complex text-based analytical functions such as syntax parsing, limits, where grammar ends in a sentence, must be known. In simpler terms, it breaks the paragraph into separate sentences [1, 19].

Example: Contemplate the COVID 19 comment Sample –

If you do not recommend it for 18 and under, how is it remotely safe for above 18-year-olds? Our makeup is not that different??? Thank god you came out and said not safe for 18 and under at least.

Sentence Segment produces the following result:

a) *"If you do not recommend it for 18 and under, how is it remotely safe for above 18-year-olds?"*
b) *"Our makeup is not that different???"*
c) *"Thank god you came out and said not safe for 18 and under at least."*

d) **PoS Tagging:** is the process of tagging every token collected from the corpus with its respective 'Part Of Speech'. PoS tagging helps in finding out how a word is used in a sentence for instance as – "**nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions** or **interjections**." It provides the fundamental step right before chunking to set the path for Word Sense Disambiguation (WSD) by properly identifying the part of speech of each for every token generated from the text-based corpus [8].

Example:

"Staying healthy and "social distancing" are mutually exclusive."

Output: -

[('staying', 'VBG'), ('healthy', 'JJ'), ('and', 'CC'), ('social, 'JJ'), ('distancing', 'NN'), ('are', 'VBP'), ('mutually', 'RB'), ('exclusive', 'JJ')].

e) **Chunking:** calls the PoS-tagged tokens for phrases. Chunking can be defined as the process of mining phrases from unstructured text. It is not responsible for the internal structure of the constituents, nor their usage in the leading sentence. It rather works on top of POS tagging by identifying those constituents in the form of a group of words like Noun Phrase, Verb Phrase, Prepositional phrase, etc. [8, 15].

Example:

The covid patient is lying in the ICU.

Chunking Output:

[The covid patient]_np [is lying]_vp [in the ICU]_pp

(*np* stands for "noun phrase," *vp* stands for "verb phrase," and *pp* stands for "prepositional phrase.")

f) **Syntax Parsing:** determines the structure of a sentence. In simpler terms, *Syntax Parsing* could be called *Sentence Diagramming* that acts as a preliminary step in processing any natural language features. It is considered as one of the most computationally-intensive steps while performing analysis on text-based content to gain insight into grammar and syntax. The Syntax tree for the above given example can be seen in Fig. 3.

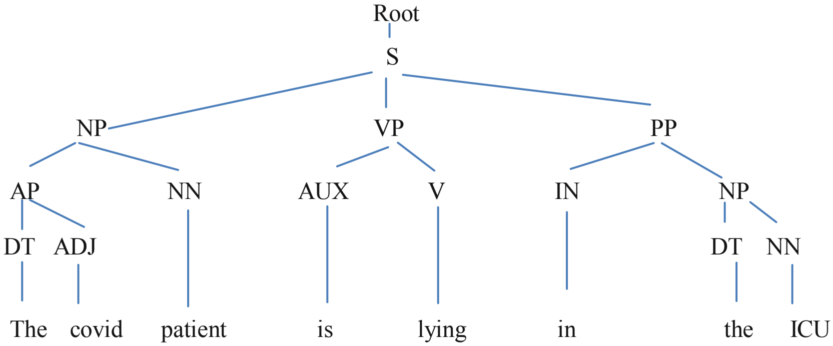


Fig. 3. Syntax parsing

g) **Sentence Chaining:** The concluding step in organizing the raw and amorphous text for further analysis at complex levels is called *sentence chaining*. It is the process to link individual but related sentences by the "strength of association" of the sentence w.r.t. the title of the content. The lexical chain helps in combining sentences, even if they are present apart from each other in a document. It detects the predominant topics for a machine and measures the overall context of the document. It also helps in observing where linkages are shown for ontological meaning to the comment thus providing morphological relations among words.

4 Natural Language Processing

Natural Language Processing (NLP) is categorized as a sub-domain of dialectology, computer science, knowledge engineering and artificial intelligence implicating fundamental relations between computers and humane dialects. Predominantly, it concentrated on organizing systems to process and analyze massive natural language data [18].

NLP makes use of Tokenization, Sentence breaking, Part of Speech tagging, Chunks of tokens and PoS tags. In machine learning (ML) jargon, these series of steps taken are called data pre-processing. The idea is to break down the natural language text into smaller and more manageable chunks. These can then be analyzed by ML algorithms to find relations, dependencies, and context among various chunks. NLP utilizes these fundamental functions in order to achieve its two components while taking ontologies and Knowledge Representation into consideration, i.e. Natural Language Understanding and Natural Language Generation.

4.1 Natural Language Understanding (NLU)

NLU aids the machine in understanding and analyzing the human language with the help of metadata extracted from content such as entities, keywords, relations, semantic and syntactic roles etc. [20]. It involves Mapping the given input into useful representation, Analyzing different aspects of the language, Interpreting Natural Language, Deriving Meaning, Identifying context and Deducing Insights. Word Sense Disambiguation (WSD), a function that is implemented via NLU makes sure that the machine is able to understand the two different senses of a word belonging to a glossary [21].

4.2 Natural Language Generation (NLG)

NLG helps in converting the machine formatted data into a representation that could be read by a human. It is achieved by three common steps i.e. planning of textual content, Planning of Sentence making, and finally Realization of the text that will be represented as a Natural Language [6].

It's important to note that in NLU, the process is to disambiguate the input sentence to produce a language that is known to the machine, whereas in NLG the process is about making decisions regarding the arrangement of representation into words known to humans [6, 21].

5 Data and Methodology

5.1 Data Collection

The data in the current research is collected manually and directly in real-time from three social media i.e., Facebook, Twitter and YouTube's official press conference. The data collection started in mid-July 2020 and continued extraction till mid-May 2021. An unstructured dataset of 60,365 text-based discussions from different posts and various concerning topics, was converted into structured data that included comments, tweets and replies related to the pandemic COVID -19 around the world.

5.2 Pre-processing of Data

Data were preprocessed using various basic steps except stemming or lemmatization in order to analyze the real word association, like applying stop word removal, punctuation removal, emoji removal, hyperlinks removal, numerical removal, eliminating extra white spaces and converting all upper cases to lower cases for achieving feature extraction by selecting most frequently used words.

5.3 Working of Proposed Analysis: Methodology

The stepwise process of the NLP for Novel TBA can be observed as follows:

Step 1. After preprocessing, tokens were generated for words as well as their respective PoS Tags for their respective hypernyms, in order to get maximum content discussed.

Step 2. It can be seen with the help of the following Table 2, that a Glossary is created and a gloss is tagged along the word for better understanding of context to accommodate easier search and lookup for further processing. Thus establishing basic parameters for WSD and understanding the real meaning and context of the comment.

Table 2. Sample Glossary for COVID19 Comments

Sample Words from Corpus	Word	Glossary
	COVID-19	G100 Series: G101Disease
	Bat	G200 Series: G201Mammal G202 An appliance used for hitting the ball in cricket G203 A wedge used for pottery
	Safe	G300 Series: G301 Protected from or not exposed to danger or risk G302 A strong fireproof cabinet with a complex lock, used for the storage valuables
	Passing	G400 Series: G401 Going past G402 The end of something G402 A person's death

Step 3. After that, the tokens are generated for words' distribution of the corpus on the basis of "is-a" relationship thus providing the concepts in the domain and also the relationships that hold between those concepts to observe the ontology behind the words used in some context.

Step 4. A Text bag is created, after extracting word definition corresponding to the Weighted Vector that will lead to a separate bag of features, 10 hot topics are generated on the basis of word usage and Ontology observed.

Step 5. These hot Topics are Labelled from S_1 to S_{10} depending upon the features classified with respect to the Weighted Vector in Lexicon, given by

$$\overline{V_w} = \sum\nolimits_{i=1}^{10} S_i w_i. \tag{1}$$

- where S_i is the feature vector from the text bag and w_i is the average weight of the frequently used word.

Step 6. Evaluation of the maximum weighted overlapping between the context bag of words and the Si bag of Words is observed to chunk the sentences for respective Si's feature from the entire corpus.

Step 7. Corpus Analysis to achieve Word Sense Disambiguation through Proper Knowledge Representation is implemented.

Step 8. Finally results based on the novel TBA approach using NLP steps will deduce the interesting insights regarding COVID 19.

The entire process of how novel Text-Based Analysis was obtained via NLP steps that helped in further classification of the corpus into ten hot topics' categories is shown in the following Fig. 5, thus concluding their insights (Fig. 4).

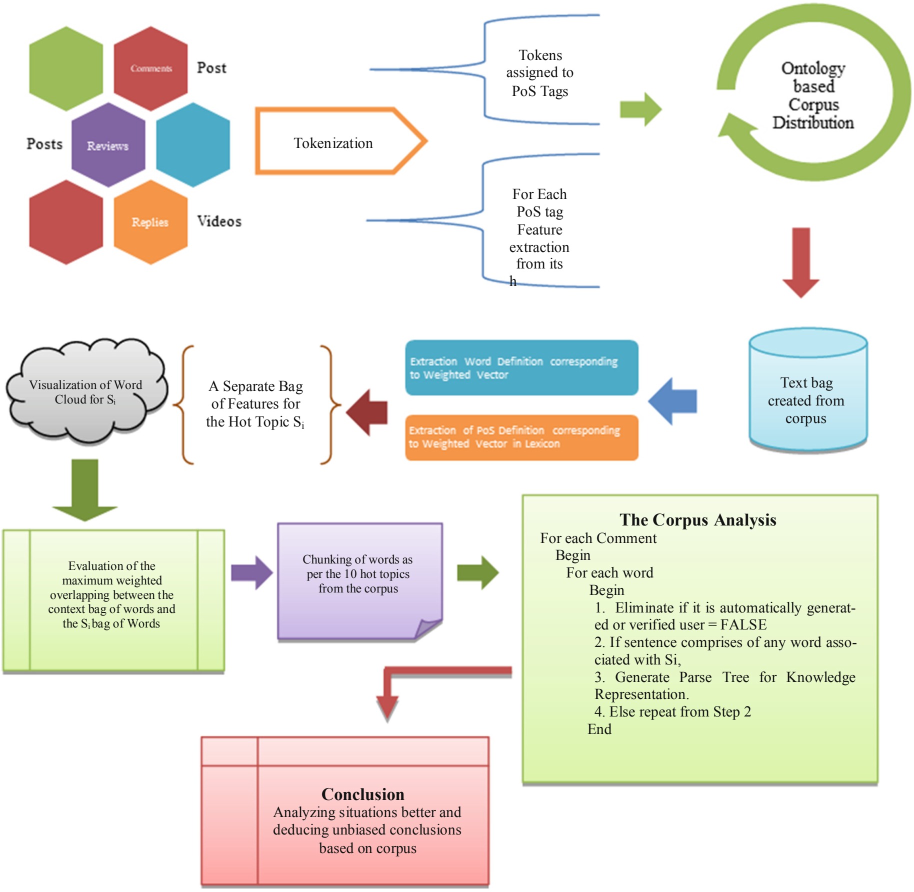
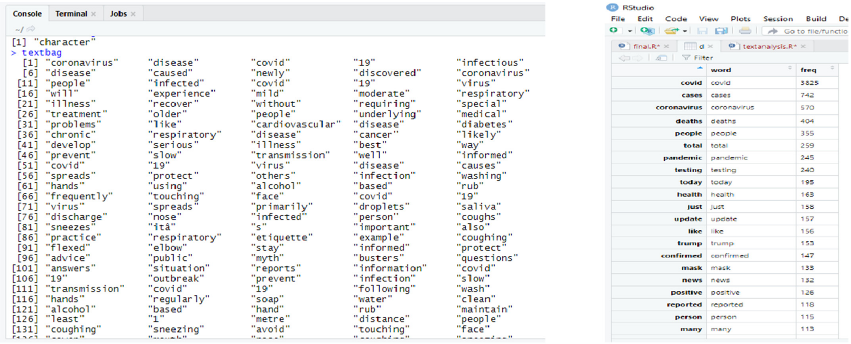

Comments
Post
Posts
Reviews
Replies
Videos
Tokenization
Tokens assigned to PoS Tags
For Each PoS tag Feature extraction from its h
Ontology based Corpus Distribution
Text bag created from corpus
Extraction Word Definition corresponding to Weighted Vector
Extraction of PoS Definition corresponding to Weighted Vector in Lexicon
A Separate Bag of Features for the Hot Topic S_i
Visualization of Word Cloud for S_i
Evaluation of the maximum weighted overlapping between the context bag of words and the S_i bag of Words
Chunking of words as per the 10 hot topics from the corpus
The Corpus Analysis
For each Comment
Begin
For each word
Begin
1. Eliminate if it is automatically generated or verified user = FALSE
2. If sentence comprises of any word associated with Si,
3. Generate Parse Tree for Knowledge Representation.
4. Else repeat from Step 2
End
Conclusion
Analyzing situations better and deducing unbiased conclusions based on corpus

Fig. 4. Proposed analysis steps: methodology

6 Results and Discussions

The corpus analysis achieved in Sect. 5.3 using the novel TBA steps has displayed a decent approach to represent knowledge that is available on the social media platforms and resolving Word Sense Disambiguation. Since many comments, tweets and replies were irrelevant from the topic, a text bag, is created to eliminate those unnecessary comments, then observe the texts, that showed relevance to COVID-19 and are frequently used. The text Bag that was created after the ontology-based distribution as shown in Fig. 5(a) helped in collecting information for the words' usage that in turn helped in collecting the most frequently used words as shown in Fig. 5(b).



```
[1] "character"
> textbag
  [1] "coronavirus"    "disease"        "covid"          "19"             "infectious"
  [6] "disease"        "caused"         "newly"          "discovered"     "coronavirus"
 [11] "people"         "infected"       "covid"          "19"             "virus"
 [16] "will"           "experience"     "mild"           "moderate"       "respiratory"
 [21] "illness"        "recover"        "without"        "requiring"      "special"
 [26] "treatment"      "older"          "people"         "underlying"     "medical"
 [31] "problems"       "like"           "cardiovascular" "disease"        "diabetes"
 [36] "chronic"        "respiratory"    "disease"        "cancer"         "likely"
 [41] "develop"        "serious"        "illness"        "best"           "way"
 [46] "prevent"        "slow"           "transmission"   "well"           "informed"
 [51] "covid"          "19"             "virus"          "disease"        "causes"
 [56] "spreads"        "protect"        "others"         "infection"      "washing"
 [61] "hands"          "using"          "alcohol"        "based"          "rub"
 [66] "frequently"     "touching"       "face"           "covid"          "19"
 [71] "virus"          "spreads"        "primarily"      "droplets"       "saliva"
 [76] "discharge"      "nose"           "itâ"            "person"         "coughs"
 [81] "sneezes"        "itâ"            "s"              "important"      "also"
 [86] "practice"       "respiratory"    "etiquette"      "example"        "coughing"
 [91] "flexed"         "elbow"          "stay"           "informed"       "protect"
 [96] "advice"         "public"         "myth"           "busters"        "questions"
[101] "answers"        "situation"      "reports"        "information"    "covid"
[106] "19"             "outbreak"       "prevent"        "infection"      "slow"
[111] "transmission"   "covid"          "19"             "following"      "wash"
[116] "hands"          "regularly"      "soap"           "water"          "clean"
[121] "alcohol"        "based"          "hand"           "rub"            "maintain"
[126] "least"          "1"              "metre"          "distance"       "people"
[131] "coughing"       "sneezing"       "avoid"          "touching"       "face"
```

	word	freq
covid	covid	3825
cases	cases	742
coronavirus	coronavirus	570
deaths	deaths	404
people	people	355
total	total	259
pandemic	pandemic	245
testing	testing	240
today	today	195
health	health	163
just	just	158
update	update	157
like	like	156
trump	trump	153
confirmed	confirmed	147
mask	mask	133
news	news	132
positive	positive	126
reported	reported	118
person	person	115
many	many	113

(a) (b)

Fig. 5. **(a)** Screenshot for COVID 19 Text Bag. **(b)** Screenshot for Word Frequencies

For each word belonging to the corpus and its corresponding frequency, Extraction Word Definition corresponding to Weighted Vector and Extraction of PoS Definition corresponding to Weighted Vector as per English Lexicon was achieved. A Separate Text Bag of Features for the Hot Topic S_i is created. The word-clouds for this separate featured set of text bag is shown in Fig. 6.

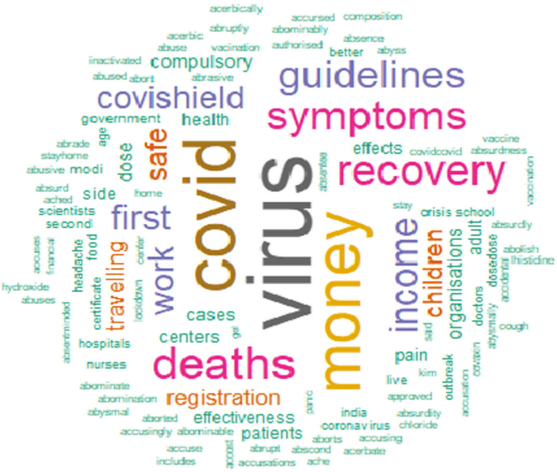


Fig. 6. Visualization of Word Cloud for Si

After that, the selected Si text bag is explored as per the word definition corresponding to Weighted Vector V_w. A sample of COVID 19 with their meaning representation is present in Table 3. The correct form of the word is replaced with the tokens.

The metadata about the knowledge base is provided that is used for classifying and organizing the content into sub-categories. It is achieved using three basic entities: users, tags, and resources to focus on Sentence chaining that categorized the entire corpus into 10 main HOT Topics:

1. Following Social distancing
2. Distribution of Masks, Hand Sanitizer and food entities
3. Medical Conditions and Aid
4. Political Agendas during the crisis
5. Bed Availability
6. Oxygen Cylinders availability
7. Online Education Policies
8. Work from Home Routine
9. Financial Crisis
10. COVID Vaccination

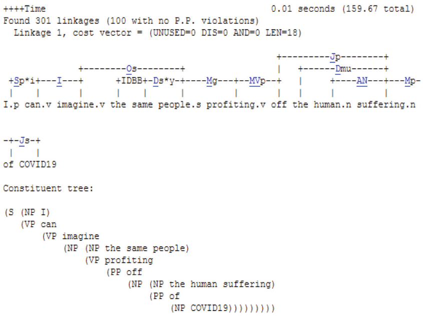
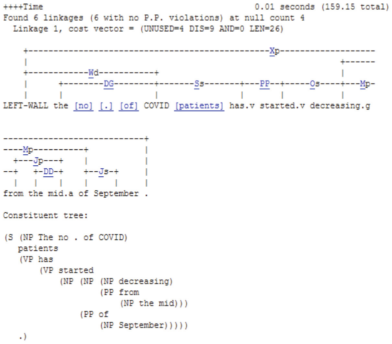
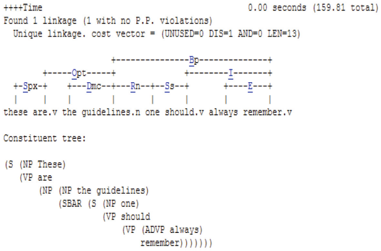
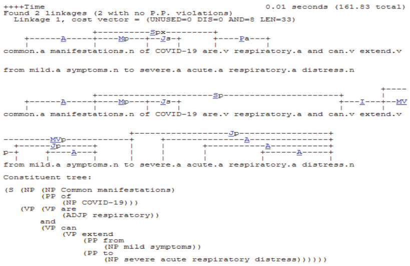
In a similar way the entire corpus was classified into the previously mentioned ten Hot Topics with the observation being made that people are concerned about certain important factors affecting life and sciences during COVID19, that must be taken into consideration by authorities and researchers of the concerned field.

It was analyzed that the major problems faced during this crisis as shown in Fig. 7 were queries regarding Vaccine, for handling Finances during critical times and Medical Conditions not only for COVID patients but for patients suffering from other critical conditions as well, respectively. These are considered the most discussed topics over the social media platforms among the entire corpus from three platforms.

For the morphological understanding of the corpus and to observe the relation between topics for the Weighted Vectors, the top five most used words from the Si i.e. #deaths, #recovery, #Money, #work, #symptoms that have and were taken compared with all topics that were frequently used among these. The analysis of the features extracted from Si with respect to these weighted vectors is shown in Fig. 6.

The data relating to COVID-19 is mostly about sufferings, losses, the crisis faced by the public, guidelines levied by governing authorities and impact on the educational and working sector. The weighted vector score observed from Fig. 8, showed that for some families, it became very hard to even manage food two times a day and were completely dependent on the Distribution of food entities. It was extremely difficult to follow social distancing guidelines with family members staying under one roof even if one of them was diagnosed with the disease. Correspondingly, by observing the meaning representation as shown in Table 3 and weighted vector score w.r.t. work from home and online education policies people working on startup businesses and science researches had to suffer great loss, as their businesses, resources and researches that were ongoing for a long period of time faced serious consequences due to lack of regular monitoring and inconsistent interactions.

Table 3. The Sample meaning representation of comments from the corpus with their classification

Word	Sample Corpus	Meaning representation	Hot Topic Classification
profit	I can imagine the same people profiting off the human suffering of #COVID19		Financial Crisis
patients	The no. of COVID patients has started decreasing from the mid of September.		COVID Vaccination
guidelines	These are the guidelines one should always? èmember *(?èmember is replaced with remember using Weighted Vector of the respective word in the Lexicon)*		Following Social Distancing
symptoms	Common manifestations of COVID-19 are respiratory and can extend from mild symptoms to severe acute respiratory distress.		Medical condition and Aid

```
++++Time                                          0.01 seconds (159.67 total)
Found 301 linkages (100 with no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=18)

                                                         +---------Jp--------+
              +--------Os--------+                       |   +------Dmu------+
 +Sp*i+---I---+      +IDBB+-Ds*y-+----Mg---+--MVp--+     |      +----AN---+---Mp-
 |    |       |      |    |      |         |       |     |      |         |
I.p can.v imagine.v the same people.s profiting.v off the human.n suffering.n


-+-Js-+
 |    |
of COVID19

Constituent tree:

(S (NP I)
   (VP can
       (VP imagine
           (NP (NP the same people)
               (VP profiting
                   (PP off
                       (NP (NP the human suffering)
                           (PP of
                               (NP COVID19))))))))
```

```
++++Time                                        0.01 seconds (159.15 total)
Found 6 linkages (6 with no P.P. violations) at null count 4
  Linkage 1, cost vector = (UNUSED=4 DIS=9 AND=0 LEN=26)

    +------------------------------------------------Xp-------------------
    |                                                              +------
    +------------Wd-----------+                                    |
    |      +--------DG--------+-------Ss-------+---PP--+-----Os----+---Mp-
    |      |                  |                |       |           |
LEFT-WALL the [no] [.] [of] COVID [patients] has.v started.v decreasing.g


----------------------------+
----Mp----------+           |
  +---Jp---+    |           |
--+   +-DD-+    +--Js-+     |
  |   |    |    |     |     |
from the mid.a of September .

Constituent tree:

(S (NP The no . of COVID)
   patients
   (VP has
       (VP started
           (NP (NP (NP decreasing)
                   (PP from
                       (NP the mid)))
               (PP of
                   (NP September)))))
   .)
```

```
++++Time                                          0.00 seconds (159.81 total)
Found 1 linkage (1 with no P.P. violations)
  Unique linkage. cost vector = (UNUSED=0 DIS=1 AND=0 LEN=13)

                  +---------------Bp--------------+
        +-----Opt-----+               +--------I-------+
  +-Spx-+    +---Dmc--+---Rn--+--Ss--+       +----E---+
  |     |    |        |       |      |       |        |
these are.v the guidelines.n one should.v always remember.v

Constituent tree:

(S (NP These)
   (VP are
       (NP (NP the guidelines)
           (SBAR (S (NP one)
                    (VP should
                        (VP (ADVP always)
                            remember)))))))
```

```
++++Time                                              0.01 seconds (161.83 total)
Found 2 linkages (2 with no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=8 LEN=33)

            +----------Spx---------+
  +------A-----+----Mp---+--Js-+   +----Pa---+
  |            |         |     |   |         |
common.a manifestations.n of COVID-19 are.v respiratory.a and can.v extend.v

from mild.a symptoms.n to severe.a acute.a respiratory.a distress.n

                                                                        +----
            +----------------------Sp----------------------+            |
  +------A-----+----Mp---+--Js-+                           +---I---+--MV
  |            |         |     |                           |       |
common.a manifestations.n of COVID-19 are.v respiratory.a and can.v extend.v

                  +-----------------Jp-----------------+
---------MVp-----------+     +---------------A---------------+
 +------Jp------+      |     |       +-----------A-----------+
p-+     +----A---+     |     |       |          +------A-----+
  |     |        |     |     |       |          |            |
from mild.a symptoms.n to severe.a acute.a respiratory.a distress.n

Constituent tree:

(S (NP (NP Common manifestations)
       (PP of
           (NP COVID-19)))
   (VP (VP are
           (ADJP respiratory))
       and
       (VP can
           (VP extend
               (PP from
                   (NP mild symptoms))
               (PP to
                   (NP severe acute respiratory distress))))))
```

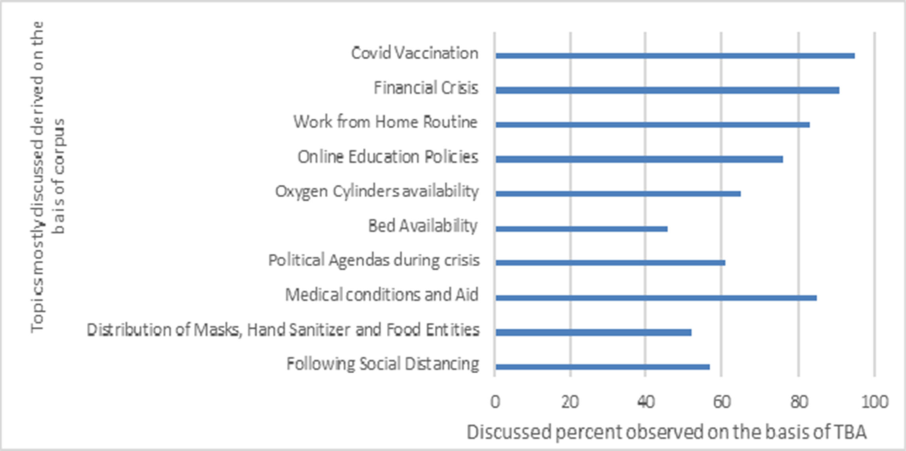


Fig. 7. Percentage analysis of hot topics discussed over the entire corpus

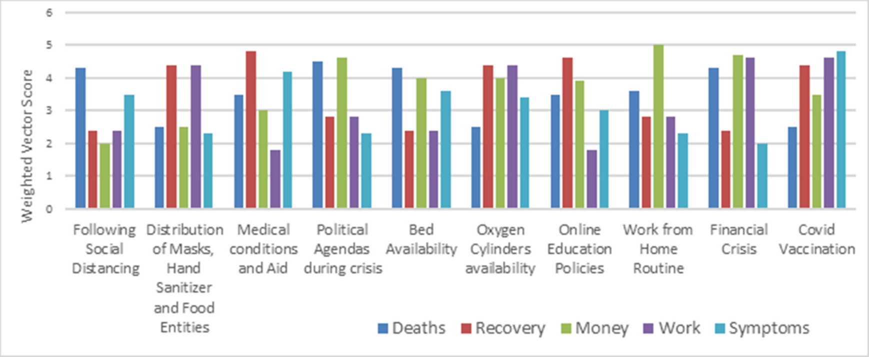


Fig. 8. Hot Topics w.r.t. Weighted Vectors of S_i

7 Conclusion and Future Work

The Novel TBA approach has provided crucial analysis steps to better understand the basic concepts of Natural Language Processing that were encountered during the initial level of data analysis. The implementation of Knowledge Representation as shown in Table 3 has revealed that the taxonomy of Constituent Tree could be parsed with the ontologies to get a better understanding of text analytics while trying to achieve text classification.

The corpus is constructed and categorized on the basis of most discussed topics derived from frequently used words of comments, tweets and replies on Facebook, Twitter and YouTube respectively with the help of R and Python packages. The collected data was converted to corpus and the ten hot topics that need attention were discussed.

The proposed work could make use of a more complex version for the morphological analysis of the corpus. Future research could use an investigative approach to find the

Lexical relations of varying features and their enactment on the sub-contexts and domains at deeper levels. Also, sentiments based on the knowledge extracted can be obtained to get even the cynical insights into the situation.

References

1. Pittaras, N., Giannakopoulos, G., Papadakis, G., Karkaletsis, V.: Text classification with semantically enriched word embeddings. Nat. Lang. Eng. **27**(4), 391–425 (2020). https://doi.org/10.1017/S1351324920000170
2. Nemes, L., Kiss, A.: Social media sentiment analysis based on COVID-19. J. Inf. Telecommun. **5**(1), 1–15 (2020). https://doi.org/10.1080/24751839.2020.1790793
3. Lata, K., Singh, P., Dutta, K.: A comprehensive review on feature set used for anaphora resolution. Artif. Intell. Rev. **54**(4), 2917–3006 (2020). https://doi.org/10.1007/s10462-020-09917-3
4. Garcia, K., Berton, L.: Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Appl. Soft Comput. **101**, 107057 (2021). https://doi.org/10.1016/j.asoc.2020.107057
5. Zhuhadar, L., Nasraoui, O., Wyatt, R., Yang, R.: Visual knowledge representation of conceptual semantic networks. Soc. Netw. Anal. Min. **1**(3), 219–229 (2011). https://doi.org/10.1007/s13278-010-0008-2
6. Van Harmelen, F., Lifschitz, V., Porter, B. (eds.): Handbook of Knowledge Representation, vol. 1. Elsevier (2008)
7. Martin, M.K., Pfeffer, J., Carley, K.M.: Network text analysis of conceptual overlap in interviews, newspaper articles and keywords. Soc. Netw. Anal. Min. **3**(4), 1165–1177 (2013). https://doi.org/10.1007/s13278-013-0129-5
8. Güngör, O., Güngör, T., Üsküdarli, S.: The effect of morphology in named entity recognition with sequence tagging. Nat. Lang. Eng. **25**(1), 147–169 (2019). https://doi.org/10.1017/S1351324918000281
9. Park, J., Chung, E.: Learning from past pandemic governance: early response and public-private partnerships in testing of COVID-19 in South Korea. World Dev. **137**, 105198 (2021)
10. de las Heras-Pedrosa, C., Sánchez-Núñez, P., Peláez, J.I.: Sentiment Analysis and Emotion Understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. Int. J. Environ. Res. Pub. Health **17**(15), 5542 (2020).https://doi.org/10.3390/ijerph17155542
11. Chen, Q., Min, C., Zhang, W., Wang, G., Ma, X., Evans, R.: Unpacking the black box: how to promote citizen engagement through government social media during the COVID-19 crisis. Comput. Hum. Behav. **110**, 106380 (2020). https://doi.org/10.1016/j.chb.2020.106380
12. Limaye, R.J., et al.: Building trust while influencing online COVID-19 content in the social media world. Lancet Digit. Health **2**(6), e277–e278 (2020). https://doi.org/10.1016/S2589-7500(20)30084-4
13. Yue, L., Zhang, L.: Social media WeChat infers the development trend of COVID-19. J. Infect. **81**(1), e82–e83 (2020). https://doi.org/10.1016/j.jinf.2020.03.050
14. Rajkumar, R.P.: COVID-19 and mental health: a review of the existing literature. Asian J. Psychiatry **52**, 102066 (2020). https://doi.org/10.1016/j.ajp.2020.102066
15. Cinelli, M., et al.: The COVID-19 social media infodemic. Sci. Rep. **10**(1), 16598 (2020). https://doi.org/10.1038/s41598-020-73510-5
16. Dias, G., Moraliyski, R., Cordeiro, J., et al.: Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. Nat. Lang. Eng. **16**(4), 439–467 (2010). https://doi.org/10.1017/S135132491000015X

17. Dornescu, I., Orăsan, C.: Densification: semantic document analysis using Wikipedia. Nat. Lang. Eng. **20**(4), 469–500 (2014). https://doi.org/10.1017/S1351324913000296
18. Akhtar, M.S., Ghosal, D., et al.: A multi-task ensemble framework for emotion, sentiment and intensity prediction, computation and language (2018). https://arxiv.org/abs/1808.01216
19. Malla, S.J., Alphonse, P.J.A.: COVID- 19 outbreak: an ensemble pre-trained deep learning model for detecting informative tweets. Appl. Soft Comput. **107**, 107495 (2021)
20. Macherey, K., Och, F.J., Ney, H.: Natural language understanding using statistical machine translation. In: 7th European Conference on Speech Communication and Technology (2001)
21. Russell, S.J.; Norvig, P.: Artificial Intelligence: A Modern Approach, p. 19. Prentice Hall (2003). ISBN 0-13-790395-2. http://aima.cs.berkeley.edu/