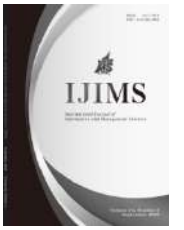
http://ijims.ms.tku.edu.tw/main.php

International Journal of
Information
and
Management
Sciences

International Journal of
Information and Management Sciences

34 (2023), 327-342. DOI:10.6186/IJIMS.202309_34(4).0005



PHWordSet: the Chinese Sentiment Lexicon for Emotional Analysis of Public Health Emergencies

Wei Shi[1], *Guang-cong Xue*[2], *Yue Fu*[1*]
[1] ZheJiang Ocean University and [2] Huzhou University.

Keywords

Sentiment lexicon; network public opinion; Word2Vec; label propagation

Abstract

In this paper, we propose a method to automatically construct a target-specific sentiment lexicon. The core of our method is a unified framework that incorporates three kinds of methods for sentiment lexicon construction, i.e., seed emotional words are extracted from the existing general sentiment lexicon, candidate emotional words are extracted by calculating cosine similarity based on the Word2vec model, and candidate emotional words are extracted from large-scale public opinion corpus based on syntactic relationship. To avoid random propagation of label information, the LeaderRank algorithm is introduced to calculate the emphasis degree between nodes and update the emotional information of words in a certain order. Finally, the PHWordSet (sentiment lexicon of public health public opinion) is obtained. A test set containing 2000 public health data is randomly constructed, and the experimental results on the constructed public health dataset prove the effectiveness of this method.

1. Introduction

As an important emotional resource, sentiment lexicon plays an important role in sentiment analysis tasks with different granularity, such as words, phrases, attributes, sentences and text levels (Chauhan & Meena, 2020). The sentiment lexicon can be divided into general sentiment lexicon and domain sentiment lexicon according to its applicability (Fan et al., 2018). General sentiment lexicons, such as General Inquirer (GI) and SentiWordNet, are mainly constructed manually or semi-automatically, which are difficult to cover different fields of emotion words, and their domain adaptability and reliability are limited. The domain sentiment lexicon is a sentiment lexicon tailored to the content of a specific domain, with clear pointing. It adds domain-specific words to the general sentiment lexicon and rearranges the categories or polarities of words (Gupta et al., 2020; Chua & Banerjee, 2016). Therefore, many researches focus on the automatic construction of domain sentiment lexicons. One kind of methods is to use semantic

*Corresponding author. Email: 286824081@qq.com

knowledge base to expand the sentiment lexicon. This kind of method mainly uses a group of words with known polarity as the seed set, and judges the emotion tendency of unknown words through the semantic relationship in the knowledge base. It can easily and quickly expand the sentiment lexicon, but there are also problems such as relying on semantic knowledge base and limited lexicon coverage.

In order to solve the above problems, this paper proposes a sentiment lexicon construction method based on label propagation. This method obtains candidate emotion words by using Word2Vec model to train word vectors on the corpus and analyzing the relationship between words in sentences, and constructs a graph network based on the similarity between candidate emotion words and seed words, and uses label propagation algorithm to mark the emotional polarity of candidate words, and finally expands to obtain a sentiment lexicon suitable for public health research.

2. Literature Review

The construction of the sentiment lexicon is an important basis for emotional analysis. At present, foreign emotional classification research has achieved good results, thanks to the convenience of English words in emotional analysis tasks and a large number of English data sets, such as English general sentiment lexicons General Inquirer (Zeng et al., 2019) and SentiWordNet (Soumya & Pramod, 2020). Due to the variability of Chinese sentences, the multiplicity of word meanings, and the lack of data sets, domestic emotional classification research started late. In recent years, the Chinese general sentiment lexicon has also made good progress, such as the HowNet sentiment lexicon (Qi et al., 2021), which includes public domain emotional words, negative words and degree words, DLUT-Emotionontology (Wang et al., 2022) with the attributes of emotional category, emotional polarity, and emotional intensity, and the NTUSD (Jiawa et al., 2021) simplified Chinese polarity sentiment lexicon based on the binary division of text emotion. The word polarity of the general sentiment lexicon will not change with the change in the field of sentiment analysis. Although it has the advantages of large scale and high accuracy, there may be polysemy in different fields (Wang et al., 2022). Therefore, the construction of a sentiment lexicon for specific fields has become the focus of academic attention. At present, the construction methods of domain sentiment lexicon mainly include seed word-based method and corpus-based method.

The method based on seed words is to carry out communication construction based on seed words. (Shunli & Xiangxian, 2016) combined multiple general sentiment lexicons, selected various emotional seed words according to the calculated emotional word frequency, expanded the candidate words through the synonym forest, and calculated the emotional polarity of the candidate words by using the improved SO-PMI algorithm to build a sentiment lexicon for the Chinese book review. This method improves the problem of data sparsity caused by too low word frequency. In terms of word vector training, the biggest feature of a word vector is to represent the semantic information in the form of a vector distributed. When constructing the sentiment lexicon, the cosine similarity between word vectors is usually calculated to extract candidate emotional words (Li & Fan, 2019). To better analyze the emotions of investors in the financial field, Oliveira et al. (2016) takes the stock market platform StockTwits comment data as the corpus, selects the representative words in the stock field as the seed words, and constructs a sentiment lexicon for the stock market field. Chen & Chen (2020) introduced the Word2vec

model to extract high-frequency candidate emotional words, combined with the basic sentiment lexicon, degree adverbs, and negative words constructed earlier to generate the stock market sentiment lexicon and used the improved simulated annealing algorithm to optimize the emotional index of words and improve the performance of the stock market sentiment lexicon. These sentiment lexicons are closely related to their application fields and have good domain specificity.

The corpus-based method is based on a large-scale comment corpus, obtains candidate emotional words according to the co-occurrence relationship and context of words in the corpus, and continuously reduces the candidate set for lexicon construction by calculating the emotional value and emotional intensity of words. For example, because of the limited coverage of health-related terms in the general sentiment lexicon, Asghar et al. (2016) proposes a hybrid method to build a sentiment lexicon for the medical field based on the guiding concept, SWN technology, and large-scale corpus and uses TF-IDF to update the affective intensity of emotional words. Cui et al. (2018) used PMI-iIR and SO-PMI methods to extract emotional words from the large-scale fire public opinion event comment corpus to construct the original word set and combined with the general sentiment lexicon, the original word set, and the network language sentiment lexicon to construct the fire emergency network public opinion sentiment lexicon. (Zhong et al. (2016) used the association rule algorithm to extract and identify the emotional words and evaluation objects in the music commodity comment information corpus and obtained a strong association relationship between the candidate emotional words and the evaluation objects by constructing the association rule transaction set and semantic relationship tree, introduced the random walk model and mixed correlation to judge the polarity of unknown emotional words and effectively constructed the sentiment lexicon in the music field based on the quantitative model of the emotional tendency of emotional words. Based on the ultra short comment data set.

Based on the above findings, there are still some problems in the construction of sentiment lexicons: 1) there is little research on sentiment lexicons in specific fields, especially in the field of public health public opinion; 2) In practical application, the features of the social network platform, such as colloquialism and non-standard expression of words, cause many difficulties in the extraction of emotional words; 3) Emotional words have different emotional polarity in different fields, and some words even express the opposite meaning.

To solve these problems, this paper combines the Word2vec model, dependency parsing and label propagation algorithm to build a public health sentiment lexicon. Firstly, seed words are extracted from the existing sentiment lexicon. Secondly, the Word2vec model and dependency parsing are used to expand candidate emotional words from the large-scale corpus. Then label propagation algorithm and the LeaderRank algorithm are introduced to label the emotions of unknown emotional words. Compared with the existing general sentiment lexicon, the domain sentiment lexicon constructed in this paper has better emotional classification performance.

Research design and process

The construction process of sentiment lexicon in this field is shown in Figure 1, which mainly includes five parts.

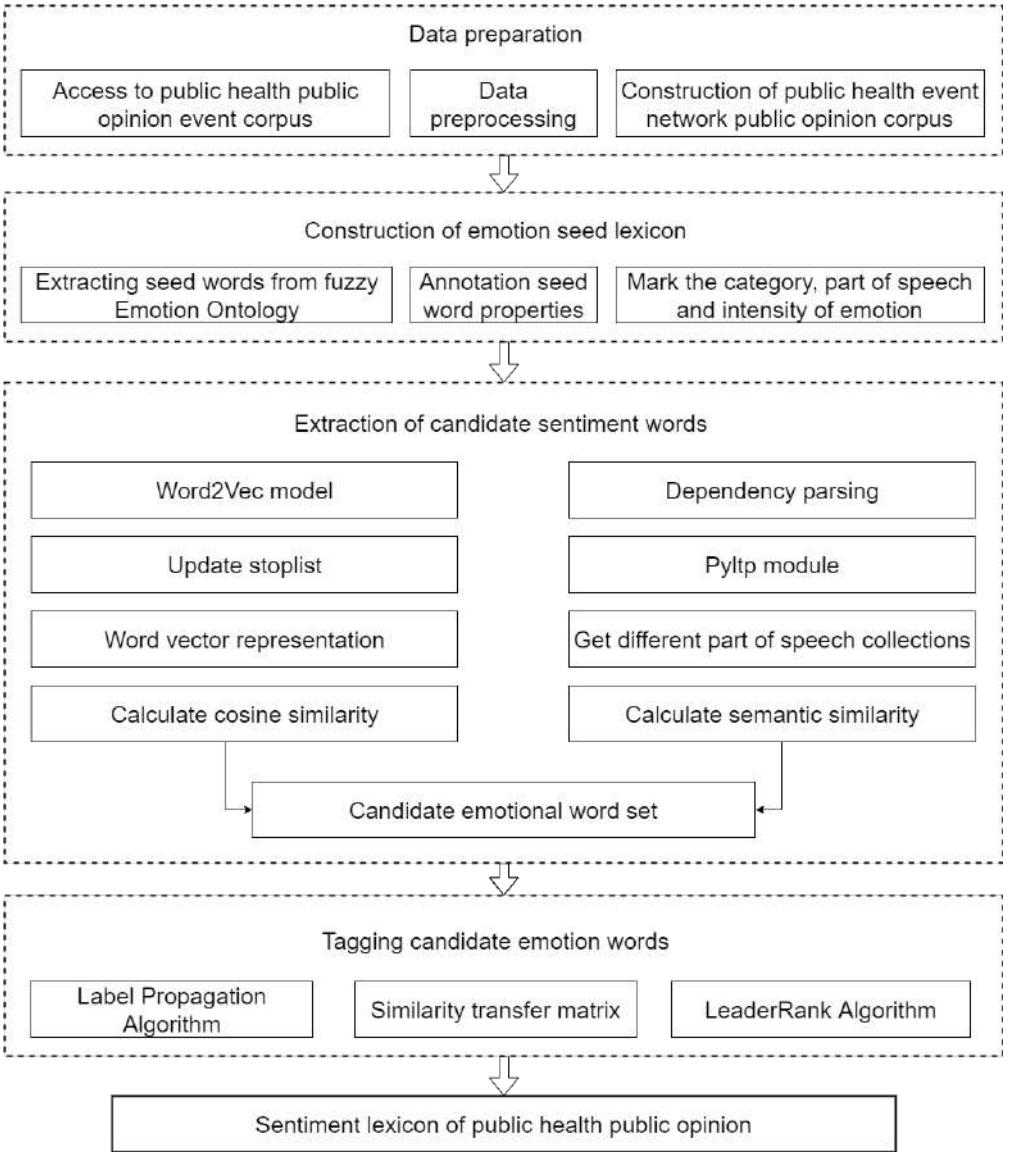


Figure 1 *The flow chart of domain sentiment lexicon construction*

3.1 Construction of corpus

This paper mainly collects the microblog and its comment data related to public health events as the online public opinion corpus. Firstly, public opinion events are screened according to the social hot spot aggregation platform. The selected types were "society" and "disaster", and the selected time was "2015-2021". Some public health events after retrieval are shown in Table 1.

Table 1 *Some public health public opinion events from 2015 to 2021*

Time (Year)	Public Health Events
2015	北京 pm2.5(Beijing PM 2.5)、杭州小学营养午餐被曝脏乱差(Hangzhou primary school nutritious lunch had ever considered)、千吨走私冻肉流入中国(One thousand tons of smuggled frozen meat into China)
2016	毒跑道(Poison runway)、上海假奶粉事件(Shanghai fake milk powder)、重庆疫苗"调包"事件(Chongqing vaccine transfer package)、医疗垃圾被做成餐具(Medical waste is made into tableware)
2017	湖南肺结核疫情(Tuberculosis epidemic situation in Hunan)、香港爆发 H3 甲型流感(H3 influenza A outbreak in Hong Kong)、Petya 勒索病毒爆发(Petya blackmail virus outbreak)
2018	吉林春芽幼儿园中毒事件(Poisoning in kindergartens)、福建泉州碳九泄漏事故(Fujian C9 leakage accident)、天津蓟州确诊非洲猪瘟(African swine fever confirmed in Jizhou)
2019	新型冠状病毒(Novel coronavirus)、天津学生感染诺如病毒(Students in Tianjin infected with norovirus)、甘肃省岷县发生非洲猪瘟疫情(African swine fever occurred in Gansu province)
2020	台湾出现首例新冠病毒变种病例(First case of novel coronavirus variant in Taiwan)、上海发现变异新冠病毒感染病例(Cases of variant novel coronavirus infection found in Shanghai)、新冠肺炎死亡数已超 SARS(COVID-19 has already surpassed SARS in the number of deaths)
2021	山东发现首例新冠变异毒株(The first new coronavirus variant found in Shandong)、河北出现瘦肉精羊肉(Clenbuterol mutton appeared in Hebei)

Then, the filtered public health events were retrieved based on the Sina Weibo platform, and the Python crawler was used to crawl the comment data of public health events. Each comment data crawling field includes user ID, comment time, primary comments, secondary comments, and other information, and a total of 620128 comment data were obtained. Secondly, as a kind of User Generated Contentdata, microblog comments contain more noise interference information, so it is necessary to clean the microblog data. This process mainly includes removing the forwarding information (forwarding microblog), reply information (reply @), mention of users (@), topics (#... #), emoticons, URL links, picture comments, and other information, and finally 317838 microblog comment data are obtained.

3.2 Construction of emotional seed lexicon

Because the selection of seed words affects the accuracy of the sentiment lexicon and the particularity of the field of public health events, we should choose the words with significant emotional tendency and unique when selecting seed words. Fuzzy emotion ontology (Shi & Fu, 2021) is a sentiment lexicon based on HowNet for online comments according to the fuzzy attributes of natural language and emotion. It divides emotion into eight categories: expectation,

happiness, love, surprise, anxiety, sadness, anger, and hate. Fuzzy emotion ontology includes fuzzy emotional word ontology and fuzzy evaluation word ontology. Among them, fuzzy emotional word ontology contains 2090 words expressing emotion, while fuzzy evaluation word ontology contains 6862 words expressing their views or positions.

First, count the frequency of emotion words of each emotion category in the public health event network public opinion corpus, and extract the top 10 words as the seed words of each emotion category; Secondly, we label the selected emotional seed words with emotional attributes, which mainly include emotional category, part of speech and emotional intensity. The traditional sentiment lexicon divides the polarity of emotion into positive, neutral and negative. This paper follows the classification of sentiment in the fuzzy emotion ontology. The positive emotion category includes expectation, happiness, love and surprise, and the negative emotion category includes anxiety, sadness, anger and hate; Part of speech includes noun, verb, adjective, adverb and idiom; The intensity range of positive emotion words is [1,4], and the intensity range of negative emotion is [- 4, - 1]. The larger the number is, the stronger the emotion intensity is.

Because the emotional attribute has a certain human-centered view, the method of multi-person labeling is adopted, that is, three different labels label the same unit. When the label results of two people are consistent, the scheme is adopted for labeling; When the labeling results are inconsistent, the most consistent scheme will be adopted in combination with the labeling results of the third labeling person. The final emotion seed words are shown in Table 2, and each emotion word is saved in the form of triple.

Table 2 *Some seed emotional words and their attributes*

Emotional Words	Attributes	Emotional Words	Attributes	Emotional Words	Attributes
期望 (expect)	[expectation, verb, 4]	牵挂 (care)	[love, verb, 2]	悲痛 (grieved)	[sadness, noun, -2]
欢迎 (welcome)	[expectation, verb, 4]	新奇 (novel)	[surprise, noun, 1]	指责 (accuse)	[anger, verb, -3]
欢呼 (cheer)	[happiness, verb, 3]	怀疑 (doubt)	[surprise, verb, 1]	严肃 (serious)	[hate, adj, -4]
爱惜 (cherish)	[love, verb, 2]	苦恼 (distress)	[anxiety, verb, -1]	忽视 (ignore)	[hate, verb, -4]

3.3 Extracting candidate emotional words based on the Word2vec model

The process of representing text data as word vectors is called word embedding. In 2013, Google released the Word2vec model(Mikolov et al., 2013), an open-source word embedding tool based on deep learning. The model includes a two-layer neural network to represent words in high-dimensional vector space in the form of word vectors, to predict the similarity between words. This paper uses the skip-gram model in word2vec to construct the word vector. The word vector dimension is set to 128, the word nearest neighbor window is set to 5, and the minimum word frequency of the word vector is 5.

Combined with Baidu stop word list, Harbin Institute of technology stop word list, Sichuan University stop word list, and Jieba stop word list, the user-defined stop word list is obtained. Remove more meaningless conjunctions and modal particles, such as “的”, “啊”, “吧”, “吗” and other words, because removing these words has no practical impact on the textual sentiment analysis. 3137 stop words are obtained after the above operation. Finally, Chinese word segmentation is carried out on the corpus data, and the word vector representation of each word is obtained through model training. The cosine similarity between the segmented word and the seed word is calculated, and the words greater than the set similarity threshold are extracted as candidate emotional words. The cosine similarity calculation formula is shown in formula 3.1. Word1 and word2 represent two words respectively, which are mapped into an n-dimensional vector after word2vec model training, n represents dimension, and word1i and word2i represent the value of its dimension respectively.

$$\mathrm{Sim}_{w2v}(\mathrm{word1}, \mathrm{word2}) = \frac{\sum_{i=1}^{n} \mathrm{word1}_i \mathrm{word2}_i}{\sqrt{\sum_{i=1}^{n} \mathrm{word1}_i^2}\sqrt{\sum_{i=1}^{n} \mathrm{word2}_i^2}} \quad (3.1)$$

3.4 Extracting candidate emotional words based on dependency parsing

Dependency parsing is an important task in natural language processing. Its goal is to describe the relationship between each unit afterword segmentation, so as to find the corresponding relationship between comment object and emotional word. These relationships include subject-predicate relationships, verb object relationships, inter-object relationships, and so on, but there is only one core relationship in the sentence. There are two main ways to realize dependency parsing: graph-based dependency parsing and transfer-based dependency parsing. The former uses the algorithm to obtain the maximum spanning tree (MST) as the dependency tree, while the latter uses the machine learning model to predict and construct actions according to some features of the sentence, and the computer assembles the correct dependency tree according to these actions(Pan et al., 2019).

Common syntactic analysis tools include Stanford parser, FoolNLTK, and the language technology platform (LTP) of Harbin Institute of technology. Among them, the language technology platform (LTP)(Che et al., 2010) developed by the social computing and information retrieval research center of Harbin Institute of technology provides richer and more accurate Chinese natural language processing tasks such as Chinese word segmentation, part of speech tagging, named entity recognition, dependency syntactic analysis, semantic role tagging and so on. Based on the open-source Pyltp module provided by LTP, this paper analyzes the dependency syntax of the corpus of public health public opinion events. Firstly, part of speech tagging is carried out for the results of corpus word segmentation (see Table 3 for the relationship of part of speech tagging), and then the tagged words are summarized into the corresponding set according to part of speech. After obtaining different parts of speech sets, with the help of the concept of “义原(sememe)” proposed by HowNet, the sememe similarity between words and seed words in verb set, object set and attribute set and the sememe similarity of two words in parallel set are calculated respectively, and words greater than the set threshold are extracted as candidate emotional words. The calculation process of sememe similarity is shown in formula 3.2, where Sim (S1, S2) represents the similarity between two sememes, and the value range is

0 ~ 1(The larger the value, the higher the similarity between the two sememes), dis (S1, S2) is the distance between the two sememes, and Min (depth (S1), depth (S2)) is the smaller value of the depth of the two sememes.

Table 3 *Part of speech relations*

Label	Relationship Type	Example
SBV	主谓关系 （predicate relations）	我害怕这次疫情 [我<-害怕] (I'm afraid of the epidemic [I <- afraid])
VOB	动宾关系 （verb-object combination）	我害怕这次疫情 [害怕<-疫情] (I'm afraid of the epidemic [afraid <- epidemic])
ATT	定中关系 （attribute and center）	日常防疫 [日常<-防疫] (Daily epidemic prevention [Daily <- prevention])
COO	并列关系 （parallel relations）	不要慌张和松懈 [慌张<-松懈] (Don't panic and relax [panic <- relax])

$$Sim_{LTP}(S_1, S_2) = \frac{\text{Min}(\text{depth}(S_1), \text{depth}(S_2))}{\text{Min}(\text{depth}(S_1), \text{depth}(S_2)) + \text{dis}(S_1, S_2)} \quad (3.2)$$

3.5 Optimized LPA algorithm

Label propagation algorithm (LPA) is a graph-based semi-supervised learning algorithm. (Raghavan et al., 2007) first applied the label propagation algorithm to the field of complex community detection, and then LPA was also widely used in the fields of feature discovery, sentiment lexicon construction and so on. The core idea of the LPA algorithm is that similar data should have the same label. The graph model is established by using the relationship between nodes. The connection of each node in the graph represents the "relationship strength" between nodes. Label propagation is completed through continuous iteration. Each node selects the label with the largest number of labels in adjacent nodes to update its label. When more than one label appears in the neighbor nodes of the node, the node will randomly select one label as the label of the node. Such division results are random, which is easy to leads to different results in each iteration, as shown in Figure 1. Assuming that the labels of nodes A, B, C, and D in Figure 2 are different, point C randomly selects the label of a neighbor node (such as A) in the first iteration, and node D also determines its label according to the number of labels of adjacent nodes in the second iteration. A has the largest number of labels in adjacent nodes, so node D selects the same label as node A, and so on. Finally, all nodes in the network are labeled with the label node A.

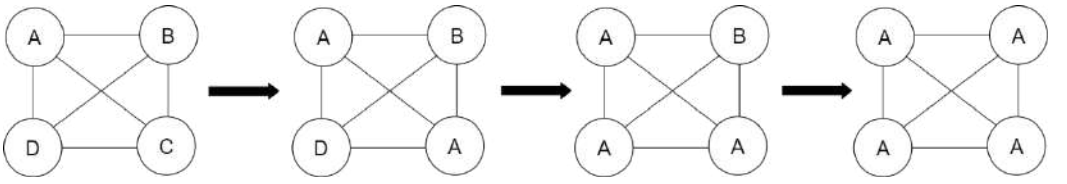


Figure 2 *The propagation randomness of LPA*

This paper takes the seed emotional words and candidate emotional words obtained in sections 3.2, 3.3, and 3.4 as nodes, takes the similarity between the seed words and candidate words as the weight of the edge, and takes the word emotional intensity as the label of the node to form a similarity relationship diagram, that is, in the process of propagation, the greater the similarity between the node and adjacent nodes, the greater the influence weight of the node marked by adjacent nodes, and the more consistent the label of the node. Since the similarity between seed words and candidate words obtained from the Word2vec model and dependency parsing may not be unique, the average value is taken as the similarity between words. Therefore, its average value is taken as the similarity between words. Assuming that there are n nodes in the graph, an N×N similarity probability matrix P is defined. The calculation method of transition probability is shown in Formula 3.3, where P_{ij} represents the transition probability between node i and node j, and $Sim(S_i, S_j)$ represents the similarity between node i and node j.

$$P_{ij} = \begin{cases} \dfrac{Sim(S_i, S_j)}{\sum_{k=1}^{n} Sim(S_i, S_k)} & (When\ the\ similarity\ is\ unique) \\ \dfrac{Sim_{w2v}(S_i, S_j) + Sim_{LTP}(S_i, S_j)}{2\sum_{k=1}^{n} Sim(S_i, S_k)} & (When\ the\ similarity\ is\ not\ unique) \end{cases} \tag{3.3}$$

In order to reduce the randomness in the process of label propagation, this paper improves the random selection part of the label propagation algorithm, introduces the LeaderRank algorithm to quantify the importance of nodes in the network, and then sorts the propagation order of nodes according to the quantization results. LeaderRank algorithm is an improved node importance ranking algorithm based on the PageRank algorithm. The algorithm does not need additional parameters, reduces the complexity of the algorithm and the impact of parameters on the accuracy, and makes the LeaderRank algorithm more suitable for exploring important nodes in complex networks. By adding a common node G to the original network, the node becomes a background node connected with other nodes, and the original network becomes a strong connection network with stronger liquidity. First, 1 unit LR value is assigned to all nodes in the figure except the common node G, and 0 unit LR value is assigned to the common node G. Then, formula 3.4 is used to calculate the LR value of each node when convergence occurs.

$$LR_i(t+1) = \sum_{j=0}^{N} P_{ij} LR_j(t) \quad t = 0,1,2,\cdots \tag{3.4}$$

Where t is the convergence times, N is the total number of nodes in the network except for the common node G, and i and j are different nodes. In the initial state, the common node and the remaining node represent the transition probability from node i to node j. Then, the LR value of the common node G in convergence is evenly distributed to each node in the network to obtain the LR value of all nodes. The calculation process is shown in Formula 3.5.

$$LR_i = LR_i(t+1) + \frac{LR_g(t+1)}{N} \tag{3.5}$$

Where $LR_g(t+1)$ represents the LR value of common node g at convergence. Sort all nodes from high to low according to LR value, and update the label of each node according to the order of LR value.

3.6 Construction process of sentiment lexicon

Based on the above methods, the construction process of the sentiment lexicon in the public health field is as follows:

Step1. Build an emotional seed lexiconWordSet based on the Fuzzy Emotion Ontology;

Step2. Use Word2vec model and dependency parsing to calculate word similarity Sim and extract words with similarity greater than 0.7 as candidate emotional words;

Step3. Based on WordSet and the candidate emotional word set, a graph network G= (N, E) is constructed, N={N1, N2,... Nn} represents the set of nodes, and E={E1, E2,... En} represents the set of edges between nodes. The transfer probability P of labels is calculated according to the similarity between nodes, and the similarity probability matrix T is constructed;

Step4. Introduce common node G, calculate LR values of all nodes and sort them from high to low;

Step5. Select the word emotional intensity S as the node labelin the process of traversing all nodes. When updating the label of node y, obtain the adjacent nodes and label information corresponding to node y, and find the label with the largest number of occurrences as the label of node y; If there is more than one label at most, the label of node y is updated according to the order of LR values of adjacent nodes;

Step6. keep the label information of the seed word unchanged in each iteration. After continuous iteration, until the labels of all nodes in the graph are no longer changed, otherwise, repeat step 5.

Step7. After the iteration process, put all nodes and their label information into the candidate emotional word set U, further filter and screen the obtained emotional words, eliminate the corpus of obvious non-emotional words, improve the accuracy of the lexicon, and manually label the part of speech of emotional words to obtain the sentiment lexicon PHWordSet in the field of public health public opinion.

The final domain sentiment lexicon contains 1263 emotional words, which are divided into eight categories: expectation, happy, love, surprise, anxiety, sadness, anger, and hate. The emotional intensity is S={-4, -3, -2, -1,1,2,3,4}. The number of words contained in each emotional category and some extended emotional words is shown in Table 4.

Table 4 *Examples of some emotional words in the sentiment lexicon of public health public opinion*

Categories	Number	Examples of Sentiment Lexicon	
期待(expectation)	102	[许愿(hope)，verb，4]	[渴望(thirst)，verb，4]
高兴(happy)	164	[胜利(victory)，noun，3]	[信仰(belief)，noun，3]
喜爱(love)	165	[守护(guard)，verb，2]	[拥抱(embrace)，verb，2]
惊讶(surprise)	43	[离谱(Outrageous)，adj，1]	[震撼(shock)，adj，1]
焦虑(anxiety)	198	[惨痛(miserable)，adj，-1]	[焦躁(impatience)，verb，-1]
悲伤(sadness)	220	[心痛(distressed)，noun，-2]	[无情(ruthless)，adj，-2]
生气(anger)	218	[造假(falsification)，verb，-3]	[浪费(waste)，verb，-3]
讨厌(hate)	153	[贬低(disparage)，verb，-4]	[致命(deadly)，adv，-4]

Experiment and results

The experiment of this paper includes two parts: construction of test set and evaluation of experimental effect. In the construction part of the test set, the user comment data of the Sina Weibo platform is used as the test sample of the sentiment lexicon; The precision, recall, and F1 value were used as the measurement indicators for the evaluation of the experimental effect.

4.1 Construction of test set

To verify the effectiveness of the sentiment lexicon in the field of public health public opinion constructed in this paper, the "pothole pickled cabbage" event is selected as a case study. On March 15, 2022, China's 315 evening party exposed several cases of violations of laws and regulations, especially the "pit pickled vegetables" incident. In a short time, the terms "Laotan pickled vegetables" and "pit pickled vegetables" quickly rushed to the hot search. After the video of its production process was exposed, it aroused public indignation. For the Chinese pothole pickled cabbage incident, statistics on the propagation trends of the topic "315 potholes pickled cabbage" incident on the microblog, Wechat and online media (As shown in Figure 3

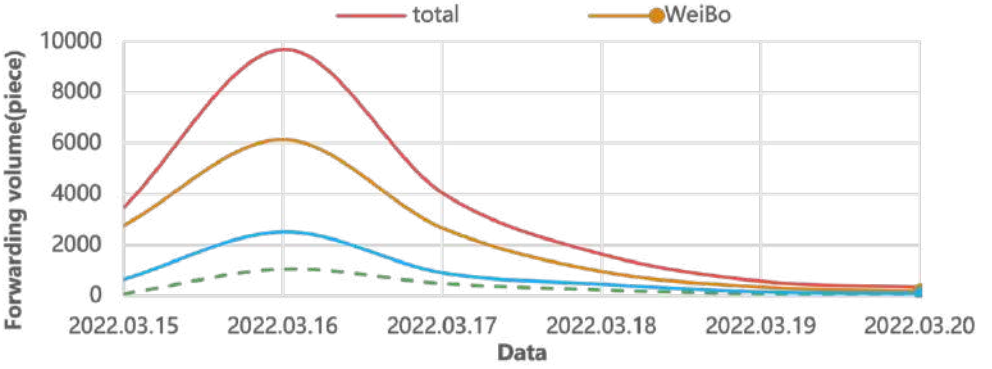


Figure 3 *The spread trend of "315 pit pickled cabbage" event by platform*

As can be seen from Figure 3, the event spread the most on Sina Weibo. The crawling time

is set from March 15 to March 18, 2022. During this time period, the topic was highly discussed and active on the Sina Weibo platform. A total of 45364 popular microblogs and comments related to the event were collected in the above time period. Since this experiment only needs the "comment" field in the comment data, the irrelevant field data such as "user ID", "comment time" and "comment connection" in the comments are excluded. Then data cleaning was carried out on the corpus. After removing the invalid comment data only including URL links, pictures, forwarded microblogs, and @ users, a total of 41284 valid comments were obtained, from which 2000 comments were randomly selected. Three taggers were selected to mark the emotional words in each comment respectively. Only when the tagging results of the three were consistent, the tagging results were output.

4.2 Experimental evaluation index

This paper uses precision, recall, and F1 measure as the measurement indicators. Precision (P) refers to the proportion of correctly labeled words in the total number of dictionaries. The calculation formula is as follows 4.1; Recall rate (R) refers to the proportion of the number of correctly labeled words in the total number of tested words. The calculation formula is as follows 4.2; the F1 value represents the harmonic average of precision and recall, and the calculation formula is shown in formula 4.3.

$$\mathrm{P} = \frac{n1}{n3} \times 100\% \tag{4.1}$$

$$\mathrm{R} = \frac{n2}{n3} \times 100\% \tag{4.2}$$

$$\mathrm{F1} = 2 \times \frac{P \times R}{P + R} \times 100\% \tag{4.3}$$

Where n1 is the number of words consistent with the manual annotation in the lexicon, n2 is the number of emotional words recognized in the lexicon, and n3 is the number of emotional words recognized in the test corpus.

4.3 Experimental result

In order to illustrate the classification effect of the sentiment lexicon constructed in this paper, it is verified from the following two aspects: emotional word recognition and emotional classification accuracy. In the aspect of emotional word recognition: firstly, the emotional words in the test corpus are manually labeled, and the positive emotional words and negative emotional words are divided to obtain the number of emotional words manually judged; Secondly, the HowNet, the DLUT-Emotionontology, the NTUSD and the PHWordSet constructed in this paper are used to identify emotional words in the test corpus, and the number of emotional words judged by the lexicon is obtained. These dictionaries are widely used and validated universal sentiment dictionaries in the field of Chinese sentiment analysis. Finally, the manual judgment results are compared with the sentiment lexicon judgment results, as shown in Table 5.

Table 5 *Comparison of the number of emotional words recognized*

	Manual Marking	HowNet	DLUT-Emotionontology	NTUSD	PHWordSet
Number of positive emotion words	374	231	282	225	296
Number of negative emotion words	586	463	501	438	510

From Table 6, it can be seen that the recognition effect of the domain sentiment lexicon constructed in this paper is better than that of the traditional general sentiment lexicon, which indicates that the lexicon can better recognize the domain emotional words. The HowNet sentiment lexicon and NTUSD sentiment lexicon identify fewer affective words, which are determined by the characteristics of new words on the Internet and the domain of public opinion, such as "拉两个背锅的(pulling two backs)", "这就很离谱(this is outrageous)" and "监管部门处罚力度太小了(the punishment of regulatory authorities is too small) ". In the network environment with the continuous emergence of new words, the traditional general sentiment lexicon can no longer meet the needs of emotional analysis in specific fields. The DLUT-Emotionontology subdivides emotions into 7 categories and 21 subcategories, and the recognition effect is better than the traditional general sentiment lexicon, but its defect is that it can not recognize the special words in specific fields.

In terms of emotional classification performance, the HowNet sentiment lexicon, the DLUT-Emotionontology, NTUSDsentiment lexicon, and the PHWordSet sentiment lexicon constructed in this paper are used to calculate the precision, recall, and F1 values on the test corpus. The comparison results of emotional classification are shown in Table 6.

Table 6 *Comparison of emotional classification performance of each sentiment lexicon*

Sentiment Lexicon	Precision（%）	Recall（%）	F1（%）
NTUSD	61.46	69.06	65.04
HowNet	67.92	72.29	70.04
DLUT-Emotionontology	71.14	81.46	75.95
PHWordSet	**78.96**	**83.96**	**81.38**

As can be seen from table 6, the precision, recall, and F1 value of the PHWordSet sentiment lexicon constructed in this paper are 78.96%, 83.96%, and 81.38% respectively. Compared with other sentiment lexicons, the precision has increased by 7% - 17% and the recall has increased by 2% - 14%. Compared with the HowNet sentiment lexicon, DLUT-Emotionontology and NTUSD sentiment lexicon of Taiwan University, the PHWordSet constructed in this paper has higher precision and domain applicability in the emotional analysis of public health emergencies. By analyzing the emotional words contained in the PHWordSet sentiment lexicon but not in the other three general sentiment lexicons, the results show that: the PHWordSet sentiment lexicon constructed in this paper better contains the unique words of public health events, such as "白衣天使(angel in white)", "支援(support)", "触目惊心(shocking)" and so on. In addition, the PHWordSet sentiment lexicon can also better identify new words on the Internet, such as "杠精(internet troll)", "吐槽(roast)", etc. Chinese has a complex sentence structure. The sentiment lexicon can simply identify the words of positive emotion and negative emotion, but

it will increase the difficulty of emotional analysis for negative sentences, interrogative sentences, double negative sentences and so on.

Conclusion

As one of the important tools in the field of emotional analysis, the part of speech distribution of the sentiment lexicon directly affects the performance of the sentiment lexicon in the task of emotional analysis. This article proposes a sentiment lexicon for the field of public health public opinion using large-scale public health public opinion event review texts, combined with a universal sentiment lexicon and deep learning methods. Compared with a universal sentiment lexicon, it can effectively improve the accuracy of sentiment analysis in the field of public health public opinion, providing a good foundation for subsequent sentiment analysis research and the construction of sentiment lexicon in other fields. The sentiment analysis based on the PHWordSet also helps to explore the public's attitudes and emotions towards public opinion events, helps relevant government departments accurately grasp the evolution process of public opinion events, and carries out correct public opinion guidance work to prevent the situation of online public opinion losing control.

Although this study has made some achievements, there are still some deficiencies. Future research will introduce a deep learning model to further expand and optimize the public opinion subject sentiment lexicon, increase the size of training corpus samples, and improve the universality of the public opinion subject sentiment lexicon in different fields. At the same time, the method of calculating emotional words is introduced to realize the research of emotional evolution and emotional prediction of network public opinion.

Acknowledgements

This study is supported by National Social Science Fund of China (No. 20BXW013); the Fundamental Research Funds for Zhejiang Provincial Universities and Research Institutes(No.2023k002); Social Science Planning Project in Zhejiang Province(No.24NDJC272YBM).

References

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016). SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, *1139,* 1-23. https://doi.org/ 10.1186/s40064-016-2809-x

Chauhan, G. S. & Meena, Y. K. (2020). Domsent: Domain-specific aspect term extraction in aspect-based sentiment analysis, In: Somani, A.K., Shekhawat, R.S., Mundra, A., Srivastava, S., Verma, V.K. (Eds.), *Smart Systems and IoT: Innovations in Computing, Smart Innovation, Systems and Technologies* (pp. 103-109). Springer Singapore. https://doi.org/10.1007/978-981-13-8406-6_11

Chen, K. & Chen, Y. (2020). Automatic construction and optimization of stock market sentiment lexicon. *Science Technology and Engineering*, *20*(5), 8683-8689. https://www.jsjkx.com/EN/Y2017/V44/I1/42

Chua, A. Y. & Banerjee, S. (2016). Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior*, *54*(c), 547-554. https://doi.org/10.1016/j.chb.2015.08.057

Che, W., Li, Z., & Liu, T. (2010). LTP: a Chinese language technology platform. *Proceedings of the 23rd International Conference on Computational Linguistics*, *Demonstrations, 15*(2), 13-16. https://aclanthology.org/C10-3004.pdf

Cui, Y., Zhang, P., & Lan, Y. (2018). On the construction of sentiment lexicon for emergency network Public Opinion of Fire Services. *Journal of Intelligence*, *37*(8), 154-160. https://doi.org/10.3389/fpsyg.2022.857769

Fan, Z., Guo, Y., Zhang, Z., & Han, M. (2018). Sentiment analysis of movie reviews based on dictionary and weak tagging information. *Journal of Computer Applications*, *38*(11), 3084-3090. https://doi.org/10.1145/3319921.3319966

Shunli, G. & Xiangxian, Z. (2016). Building sentiment Analysis dictionary for Chinese book reviews. *Data Analysis and Knowledge Discovery*, *32*(2), 67-74.

Gupta, S., Singh, R., & Singla, V. (2020). Emoticon and text sarcasm detection in sentiment analysis. In: Luhach, A., Kosa, J., Poonia, R., Gao, XZ., Singh, D. (Eds.), *First International Conference on Sustainable Technologies for Computational Intelligence: Vol. 1045. Advances in Intelligent Systems and Computing* (pp.125-150). Springer. https://doi.org/10.1007/978-981-15-0029-9_1

Jiawa, Z., Wei, L., Sili, W., & Heng, Y. (2021). Review of methods and applications of text sentiment analysis. *Data Analysis and Knowledge Discovery*, *12*(5), 1-13. https://doi.org/ 10.54097/fbem.v10i1.10171

Li, F. & Fan,Y. (2019). Research on construction method of domain sentiment lexicon. *Library Theory and Practice*, *12*(5), 60-65.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, *24*(4), 1301-1310. https://doi.org/10.48550/arXiv.1301.3781

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*(2), 62-73. https://doi.org/10.1016/j.dss.2016.02.013

Pan, H., Wei, Y., & Pan, E. (2019). Emotion analysis based on automatic extraction of syntactic patterns. *Journal of Chinese Information Processing*, *33*(9), 129-140.

Qi, F., Xie, R., Zang, Y., Liu, Z., & Sun, M. (2021). Sememe knowledge computation: a review of recent advances in application and expansion of sememe knowledge bases. *Frontiers of Computer Science*, *15*(2), 155322-155327. https://doi.org/10.1007/s11704-020-0002-4

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, *76*(7), 36106-36112. https://doi.org/10.48550/arXiv.0709.2938

Shi, W. & Fu, Y. (2021). Microblog short text mining considering context: A method of sentiment analysis. *Computer Science*, *48*(6A), 158-164.

Soumya, S. & Pramod, K. V. (2020). Sentiment analysis of Malayalam tweets using machine learning techniques. *ICT Express*, *6*(4), 300-305. https://doi.org/10.1016/j.icte.2020.04.003

Wang, C., Zhang, H., & Mo, X. (2022). Overview on sentiment analysis of microblog. *Computer Engineering and Science*, *44*(1), 165-175.

Wang, X., Liu, Y., & Li, Y. (2022). A study on the emotional evolutionary mapping of public opinion on citizens' privacy leakage in public health emergencies. *Information studies: Theory & Application*, *45*(3), 19-27. https://doi.org/10.16353/j.cnki.1000-7490.2022.03.004

Zeng, J., Duan, J., & Wu, C. (2019). Empirical study on lexical sentiment in passwords from Chinese websites. *Computers & Security*, *80*(1), 200-210. https://doi.org/10.1016/j.cose.2018.10.004

Zhong, M., Changxuan, W., & Dexi, L. (2016). Opinion lexicon construction based on association rule and orientation analysis for production review. *Journal of China Social Science Technology Information*, *35*(2), 501-509.

Wei Shi
School of Economics and Management, ZheJiang Ocean University, Zhoushan, P.R. China
E-mail: Mikeshi108@163.com
Major area(s): Sentiment analysis, text mining, Business intelligence

Guang-cong Xue
School of Information Engineering, Huzhou University, Hu Zhou 313000, P.R. China
E-mail: xgc735570365@163.com
Major area(s): Data mining, Electronic Commerce, Business intelligence

Yue Fu
School of Economics and Management, ZheJiang Ocean University, Zhoushan, P.R. China
E-mail: 286824081@qq.com
Major area(s): Sentiment analysis, information management, Business intelligence.

(Received February 2023; accepted September 2023)

Guide to Authors

1. Submission of Manuscripts

   - All manuscripts should be submitted (as either PDF, LaTeX or Word files) through the online submission system at https://www.ipress.tw/J0177.
   - Three files are necessary. The first is a regular manuscript. The second keeps the same as the first except all informations related to author(s) are removed, including acknowledgment. The third is Authors' Information (see item 12).
   - Papers submitted to this journal must NOT been previously published nor be under review by another journal. Any form of duplicate publication or plagiarism is absolutely prohibited. Violation of it will be penalized. Author(s) alone should take all responsibilities of all possible accusations due to violation of those previously mentioned.
   - Any addition, deletion, or rearrangement of the author's name in the author list should be done before the manuscript is accepted and approved by the journal editor. Please consider the list and order of authors carefully before submitting.
   - Once the files are deemed to have complied with the **APA Formatting and Citation (7th Ed.)** requirements, an acknowledgement email will be send together with a paper ID to the corresponding author.

2. Form of Manuscripts

   - The length limit for each manuscript is 20 pages including figures, tables and others.
   - Manuscript should use 12-point Times New Roman font and 1.5 spacing throughout. Space should be fully utilized.
   - All illustrations, photographs, tables, etc., should be on separate sheets, and should be included in each copy. Each page of the manuscript should be numbered.
   - Manuscript should be written in impeccable English (either US or UK spelling is accepted, but not a mixture of both).

3. Organization of the Paper

   - HEADING: The title of the paper should be concise and informative. Successive lines should give the author's name, academic or professional affiliation, and address.
   - ABSTRACT: Every manuscript must start with a concise abstract, no more than 150 words, followed by no more than 5 keywords.
   - MAIN TEXT: Structured clearly with sections such as (but not limited to) Introduction, Literature Reviews, Notations and Assumptions, Data and Methodology, Model Formulation, Research Design, Data Analysis and Results, Discussion, Conclusions, etc.
   - ACKNOWLEDGEMENTS: This section includes acknowledgements of assistances and financial support, etc. Please note that only accepted manuscripts may include this section, else place this section in your authors' file.
   - REFERENCES: See detailed instructions in item 11.

4. Illustrations
    - All Illustrations should be submitted in a form suitable for reproduction. Number the illustrations according to the sequence of their appearance in the text, where they are to be referred to as “Fig. 1”, “Fig. 2”, etc. Each illustration may have a legend, if required.
    - Photographs should be glossy prints. The authors name and figure number should be indicated on the back of each illustration.
5. In-text Citations
    - In the text, follow the rules implied in these examples. For one authr use: (Foxall, 2018) Or According to Foxall (2018).
    - For two authors use: (Mason & Missingham, 2019) Or According to Mason and Missingham (2019).
    - For more than three authors use: (Ewert et al., 2014) or According to Ewert et al. (2014).
6. Tables
    - Tables should be typed on separate pages. Number the illustrations according to the sequence of their appearance in the text, where they are to be labeled as “Table 1”, “Table 2”, etc.
    - Table titles should be short and self-explanatory.
    - A brief title should be given above each table, and any footnotes below (see Section 8).
7. Headings
    - The following sequence of headings should be used: 1, 1.1 (then 1.2, 1.3, ...), 1.1.1 (then 1.2, 1.1.3, ···).
    - Please refrain from using fourth level sections/headers.
8. Footnotes
    - In the text, footnotes are not permitted. They should be properly included in the context.
    - Use smaller font size for table footnotes.
9. Symbols and Abbreviations
    - Please use widely accepted symbols and forms of abbreviation.
    - If there is any doubt in your mind about a particular symbol or abbreviation, give the full expression followed by the abbreviation, when it appears in the text for the first time.
10. Mathematics
    - Mathematical expressions and equations should be properly typewritten, with all symbols aligned as they are to appear in print.
    - All Greek letters and other special symbols must be identified.
    - Vectors will be set in bold face and should be indicated in the manuscript by underlining with a wavy line.

- Equations or formulae should be numbered serially on the right-hand side by Arabic numberals in parentheses. For example, the first formula in Section 3 is numbered by (3.1). Only equations explicitly referred to in the text should be numbered.

11. References

- All literature citations should be collected in a list at the end of the paper and numbered alphabetically according to the last name of the first author. Retain the original title for publications in languages using the Roman alphabet. However, those employing Cyrillic and other non-Roman alphabets should be transliterated. Please note the original language at the end, e.g., "(in Russian)."
- Include DOI of the references whenever available.
- Please follow the style below:

(1) References to Book & eBook: Ewert, E. W., Mitten, D. S., & Overholt, J. R. (2014). *Natural environments and human health.* CABInternational. https://doi.org/10.1079/978184593 9199.0000 (Note: Include the DOI using the format https://doi.org/10.xxxx/xxxx)

(2) References to Chapter in an edited book or eBook: Aron, L., Botella, M., & Lubart, T. (2019). Culinary arts: Talent and their development. In R. F. Subotnik, P. Olszewski-Kubilius, & F. C. Worrell (Eds.), *The psychology of high performance: Developing human potential into domain-specific talent* (pp. 345-359). American Psychological Association. https://doi.org/10.1037/0000120-016 (Note: Use this format for both print and eBook edited book chapters, including edited book chapters from academic research databases. Include DOI if online. Please also take note how the names are given for the Editors, the initials come first.)

(3) References to Article from research databases: Washington, E. T. (2014). An overview of cyberbully in higher education. *Adult Learning*, 26(1), 21-27. https://doi.org/10.1177/1045159514558412 (Note: For journals, we italicise the Journal Name and Volume, instead of the article title. Always find articles from prominent research databases instead of some random websites, as they may be predatory or fake.)

(4) References to Conference proceedings published in a journal: Duckworth, A. L., Quirk, A., Gallop, R., Hoyle, R. H., Kelly, D. R., & Matthews, M. D. (2019). Cognitive and noncognitive predictors of success. *Proceedings of the National Academy of Sciences*, USA, 116(47), 23499-23504. https://doi.org/10.1073/pnas.1910510116 (Note: Normally Journals will indicate the Proceedings title, if not given, then please cite the source like a Journal with DOI as shown above instead of conference proceedings.)

(5) References to Report by a Government Agency: National Cancer Institute. (2019). Taking time: *Support for people with cancer* (NIH Publication No. 18-2059). U.S. Department of Health and Human Services, National Institutes of Health. https://www.cancer.gov/publications/patient-education/takingtime.pdf(Note: The specific agency responsible for the report appears as the author. The link to the report should be given. If publication number is given, then include it as well.)

(6) References to Unpublished Dissertation or Thesis: Harris, L. (2014). *Instructional leadership perceptions and practices of elementary school leaders* [Unpublished doctoral dissertation]. University of Virginia. (Notes: 1. A dissertation or thesis is considered as unpublished when you gain access through direct reading at specific library/repository but not accessibly publicly. If you find it online following published dissertation or thesis reference shown above. 2. When a dissertation or thesis is unpublished, include the description "[Unpublished doctoral dissertation]" or "[Unpublished master's thesis]" in square brackets after the dissertation or thesis title.)

(7) References to Webpage: Bologna, C. (2019, October 31). *Why some people with anxiety love watching horror movies.* HuffPost. https://www.huffpost.com/entry/anxiety-love-watching-horrormovies l 5d277587e4b02a5a5d57b59e (Notes: 1. You are advised to only use webpages with clear indication of authors and date of publication. 2. The name of the website (e.g. HuffPost) should be included before the URL. 3. No retrieval date is needed when date of publication is given.)

(8) You may refer to the full manual or follow the key information from https://apastyle.apa.org/style-grammar-guidelines/references/examples#textual-works.

12. Authors' Information

- Authors should provide the following informations in a separate file:

(1) Grant-supported research like funding, fellowship or any financial backing from person, university, organization or government.

(2) Authors' full affiliations, (department, university, country) email address and major area (at most 3 items) are to be given in order of authors. For example, Department of xxx, xxx University, ..., Canada.
E-mail: xxx
Major area(s): xxx
Department of xxx, xxx University, ..., USA.
E-mail: xxx Major area(s): xxx etc..