Soft Computing (2021) 25:3291–3315
https://doi.org/10.1007/s00500-020-05377-7

METHODOLOGIES AND APPLICATION

A systematic review to identify the effects of tea by integrating an intelligence-based hybrid text mining and topic model

You-Shyang Chen[1] · Ching-Hsue Cheng[2] · Wei-Lun Hung[2]

Published online: 16 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

An emerging problem regarding tea and health has been pyramidally addressed as research concerns from limited literature reviews to identify an important and interesting challenge. Although past studies activated for various topics and diverse purposes of tea and health, a gap for using hybrid intelligence-based techniques to discover useful information from literature analysis exists in the exploration of curative effects on tea against fatal diseases other than Western medicine therapies. This study is motivated to bridge this gap by solving this research issue for healthcare applications between tea and health. Thus, this study proposes a hybrid method of an intelligent/objective text mining technique and topic modeling principally by latent Dirichlet allocation with VEM method and Gibbs sampling as along with measurements for three evaluation metrics for model performance from published articles. In the experiment materials, this study sets conditions to collect 2109 journal articles from the Web of Science from 2010 to 2017. We divided this into three datasets that each corresponded to the three periods to differentiate discrepancies in future trends. This study contributes eight beneficial directions as follows: (1) From a technical view, the figure of VEM's perplexity has a screen plot, but Gibbs sampling is smooth and good; and, interestingly, the greater the number of topics, the lower the perplexity is; (2) in empirical results, the terms for primary topics are tea and tea compounds, and secondary topics are associated with terms for issues regarding tea and health; (3) this study yields five research findings with key empirical evidence that tea has a natural and important preventive impact on treating diseases, especially cancer; (4) as to any managerial implications, early preventions and treatments by greater tea consumption as a valuable healthcare activity with medicinal purposes; (5) regarding the novelty of this research, this study fills the gap in a hybrid knowledge-based objective text mining and topic modeling technique than past researchers have as regards tea and health issues that were only based on traditional content analysis methods; (6) for this study's strengths, it achieved manpower cost reductions and relative objectivity because the objective LDA method is rarely used for topic modeling when compared to past studies; (7) for the research significance, the proposed method benefits efficient analysis from massive amounts of extant data for exploring latent information, accelerates research processes quickly, improves understanding for new hypotheses, and identifies key questions for further research; and (8) for conclusive research importance, this study offers new rationales for medical application and discovers differentiations and gaps for studying research trends.

Keywords Tea effects for health · Literature analysis · Text mining technique · Jaccard index · Variational expectation–maximization (VEM) method · Gibbs sampling

Communicated by V. Loia.

✉ Ching-Hsue Cheng
chcheng@yuntech.edu.tw

[1] Department of Information Management, Hwa Hsia University of Technology, 111, Gongzhuan Rd., Zhonghe Dist. 235, New Taipei City, Taiwan, ROC

[2] Department of Information Management, National Yunlin University of Science and Technology, 123, University Rd., Section 3, Douliou 640, Yunlin, Taiwan, ROC

1 Introduction

This section introduces the research background and research problem, research importance and research motivation, and research gap and research objective in detail.

1.1 Research background and research problem

From a practical and academic view, tea plays an important role with a variety of significant purposes from time to time, particularly for healthcare. In tradition, tea is originated and always taken as medicinal plants in Asia areas. Tea bush is also called camellia sinensins and is originated from the lines of tropical wet and hot regions crossed in the mountainous Eastern Himalayas, such as China, India, and Myanmar. The varieties of tea are highly dependent on different types of main manufacturing processes and oxidation degrees; interestingly, the preventive effects of different teas for proper health are diverse. Initially, tea is used as a common herbal medicine (Munday 2016) useful for healthcare in there. Afterward, Europe was imported to tea, but it became rapidly popular as a common consumed beverage from the early sixteenth century. More importantly, tea has produced classical predominance forms, including black tea, green tea, and oolong tea, accounting for about 78%, 20%, and 2% of tea production worldwide due to the method differences on production processes and oxidation levels. Black tea is made for full oxidation, green tea has used for non-fermented or un-oxidized components, and the other types of teas have various ratios of semi-oxidized forms (Jain et al. 2013). Black tea has the most consumed production, but green tea attracts numerous research interests due to its major components "catechins" (flavan-3-ols) (da Silva Pinto 2013). In particular, flavan-3-ols are an extended subclass of flavonoids; evidence-based cases show that flavan-3-ols have functional effects for significantly endothelial activity, blood pressure, and mechanisms of insulin resistance. In addition, tea has functions of other compounds, as each element has a different benefit for affecting healthcare. For example, Yao et al. (2004) indicated that tea polyphenols extracted from a positive and important compound of tea can provide a great taste as well as substantial prevention effects of beneficial health against serious illnesses, such as cancer, obesity, and cardiovascular diseases. Conversely, flavonoid compounds may have potentially adverse impacts to health due to the excessive ingestion, as it has harmful to the lipid membranes and DNA (Jain et al. 2013).

Given the above descriptions, it is clear that the emerged problems of tea and its compounds have extensively been addressed by past studies for the issues of tea and health on various topics. Thus, the related issues of tea and health are activated for the key research concern, and tea becomes the core research target in this study. This research has a major interest for further and deeply studying a research problem to explore the curative effects of tea against fatal diseases other than Western therapies from the literature.

1.2 Research importance and research motivation

Although tea and its compounds benefit from an important application topic for treatment exploration of diseases, there is a lack of tea and health-related articles for the literature review analysis based on emerging text mining techniques with hybrid models. By the way, the prominence of tea as a healthcare topic will win major priority of researchers from hybrid intelligence-based angles to mining useful information. This study just has the trigger to focus on such a meaningful issue and intelligence-based text mining technique based on the following four importance of significant research application topics, as follows: (1) text mining techniques are used to progress un-structured or semi-structured data in the information age, in which enormous amounts of information or data are widely available for best use through Internet platforms, particularly for text data, and thus they have become a popular concern in research fields. Techniques and tools for text mining have functions and applications, such as text classification, text representation, text summarization, text clustering, text retrieval, document classification, opinion mining, text analysis, sentiment analysis, and entity-relation model (ER model), for numerous application fields. Interestingly, researcher has well-established the literature analysis of key text mining techniques (Amado et al. 2018; Delen and Crossland 2008; Feldman et al. 2003) to review collections from large documents to find out the potentials of semantics automatically from the articles. (2) Review what has been learned can learn something new; thus, analysis of the literature reviews can result in supporting comprehensible evidence-based and effective output. More importantly, the literature review has a conspicuous success in identifying the issues of topics addressed from the articles, in which the topic model is exploited to analyze quickly the articles published and identify effectively the topics studied by a specific goal within them. Topic model uses a generative process to offer an automated approach for a dimension reduction technique in identifying the most common subjects of documents analysis (Blei 2012). Thus, this study uses it as a key tool for research purposes and strategic foresight. (3) According to Mietzner and Reger (2005), this study finds two core shortcomings on traditional literature analysis. First, the traditional methods for analyzing the literature have easily a great waste for running times and biased sampling problems, leading to

unrepresentative of practice. Second, it has the difficulty that massive amounts of the literature reviews are handled and coded under no common standard form with a difference between research purposes and research objects. (4) Although studies have proven that more tea-drinking could lower the cause risk associated with diseases (van Dieren et al. 2009; Koga and Meydani 2001; Shen et al. 2012), related impacts connected with tea and health for papers published are relatively scarce from limited literature analysis by using a notable method of integrating hybrid intelligence-based text mining and topic model. In particular, it is found that hybrid models yield a better performance than single (or stand-alone) models from past works (Sangaiah et al. 2019a, b).

Based on the above reasons, the more diseases by practicing more aggressive therapeutic medicine triggers with the more healthcare requirements and necessity in advance to search and identify potential preventable materials for lowing medical resource-wasting from perspectives on tea of natural ecology. The challenge on tea and health thus attracts much attention from practitioners and academicians. This study is motivated by exploring the potential treatments or botanical herbs with tea and thereby studying such an interesting and important issue.

1.3 Research gap and research objective

In the conclusive remarks, identify the early preventable effects against disease (particularly for cancer) on tea is a valuable application topic. Although past researchers study on addressing tea and health issues based on traditional content analysis methods, it has a gap and is scarcely seen in hybrid knowledge-based text mining and topic modeling techniques. Thus, the above gap needs to be bridged from a broader view of relatively existing methods in a good under-used technique for applications of tea and health for natural medical researches. This study thus focuses on proposing such an integration model (Sangaiah et al. 2019a) of hybrid intelligent and objective text mining technique and topic modeling method for discovering published literature to examining and mining tea with its mixtures for healthcare application domains as well as measuring model performance by three evaluation metrics.

Four core objectives of this study are emphasized as follows. (1) Construct a hybrid intelligent technique to mine hidden and helpful information or knowledge from views of medicinal application fields. (2) Measure the relationship of tea and health by using an approach of topic modeling with useful tools to define the potential semantics. (3) Examine the reasonable meaning for the topic modeling addressed after text clustering approach. (4) Rank the importance of different topics by learning different weighting distance approaches and find out trending topics and weighting terms (keywords).

The paper is set as the arrangement organized as follows. Section 1 highlights the importance of tea and health for protruding the issues of the related text mining techniques, and Sect. 2 describes a literature review for tea and health together with text mining. The proposed method to use hybrid literature analysis and topic model approaches is presented in Sect. 3. Section 4 explores and discusses the empirical results for experiment analyses and analytical findings. Section 5 presents our conclusions and subsequent research.

2 Literature work

The section mainly describes the key conceptions of related literatures for the issue of tea and health, text mining application, analysis of literature review, and topic model approach.

2.1 Issue for tea and health

Camellia sinensins has some evergreen shrubs, in which leaf buds and leaves can be used to make tea; thus, it is also called tea. Tea actives as a plant for a healthy drink and native to Asian countries, including Taiwan, India, Japan, China, and Thailand (Cao 2013). Regarding original place with time for tea, it may be initialed in Yunnan, China in 2700 B.C. Based on a Chinese ancient book: A matter of chance, tea was found by Emperor Shen Nong who had a most well-known work that contributed to his perception as an herbal medicine figure, as he discovered that tea has common detoxifying benefits; thus, it can be taken as a continuous advantage of medicinal plants (Mahmood et al. 2010). During the Tang Dynasty in the ancient China (618–907), tea used as a consumed beverage began to flourish and famous to both the aristocrats and royals as well as also popular for common people. Until the early sixteenth century Europe, tea was imported by businessman for the purpose to commodity transaction (Chrystal 2014). Until now, tea is a most popular drink except water around the world due to its multiple helpful functions or other beneficial effects.

Interestingly, all the multiple helpful or other beneficial effects of tea highly depend on its processing skills; diverse forms of compounds make varieties of tea (Mair and Hoh 2009). Three major catalogues, black tea, green tea, and oolong tea, are grouped for different purposes and benefits. (1) Regarding black tea in the West, the brewed and fully oxidized time is about 4 min, but it is about 15 min or even longer for the boiled time in the East. In usual, the main types for black tea have traditional teas of Assam, Ceylon,

Darjeeling, Keemun, Nepal, Nilgiri, and Rize (Mair and Hoh 2009). (2) Green tea processes a more special un-oxidized and un-wilted means than that of others (both kinds of tea); thus, it can retain more certain elements than others like flavonoids (Cabrera et al. 2006). (3) The semi-oxidized oolong tea (Benn 2015) is made from partially oxidize before pan-frying also with some specific advantages. Recently, tea has extensively been focused and accepted in medical fields for healthcare purpose, and evidence indicates that tea can be used as a stimulant for healthcare benefits to prevent certain conditions or diseases, including different types of cancers, cardiovascular disease, diabetes, obesity, vomiting, or stroke (Martin et al. 2017; Pastoriza et al. 2017), particularly for beneficial effects on diabetes for green tea (Hosoda et al. 2003; Neyestani et al. 2010). Additionally, black tea and green tea are positively associated with a powerful antioxidant efficacy that has health benefits (Munday 2016). Thus, several core terms, such as tea leave, catechins, polyphenols, theaflavins, and flavanols of tea compounds are accordingly and valuably introduced, as follows. (1) The fresh tea leaves contain useful compounds of caffeine, chlorophyll with other free pigments, lignin, organic acids, theobromine, theophylline with other methylxanthines, and massed flavor elements (Graham 1992). (2) The catechin is recorded about the 30–42% dry matter escaped from the solids material within the brewed processes for green tea associated with the most plentiful polyphenols. (3) Both the polyphenols and theaflavins are strongly associated with both the helpful moieties of pyrogallol and catechol, and it was shown that they scavenge common oxidizing kinds like hydrogen peroxide, hydroxyl radicals, hypochlorous acid, peroxyl radicals, peroxynitrous acid, and superoxide radical. (4) For the flavanols, it also has a polyphenol element extracted from tea components, such as glycosides, kaempferol, and quercetin (Martin et al. 2017; Balentine et al. 1997).

The effects of tea in beneficial healthcare are examined from studies. For example, it is found that common consumption of tea-drinking lower the factor of suffering a typical type 2 diabetes that recommended at least a total of consumption for a "three cups of tea" as a daily activity (van Dieren et al. 2009). Next, a research on prospective cohort was studied and indicated that tea lessens the hazard of developing the diabetes; however, this fact can actually apply only to nonelderly patients who had previously losing weight (Greenberg et al. 2005). A research on Shanghai woman's healthcare (Nechuta et al. 2012) had assessed 69,310 women within 11 years to validate the relationship between tea-drinking consumption and cancer risk, and it was shown that Chinese women associated with regularly drinking green tea had a significant impact against cancer progression for a digestive tract. Afterward, a research on meta-analysis method (Arab et al. 2009) was studied and indicated that to consume a "three cups of black tea or green tea" habit daily prevents a serious illness with ischemic stroke. The research of Shen et al. (2012) also used a meta-analysis approach to review 14 papers published and showed that an inverse association between tea consumption and risk factor of stroke has reached statistical significance, particularly in association with an ischemic stroke. It has validated that the relations among catechins from tea, diabetes, and obesity are examined and showed that epigallocatechin-3-gallate (known as an abbreviation of EGCG) has effects of anti-obesity and anti-diabetic to the two above diseases from Kao et al. (2006). The metabolites of flavonoids effectively restrain the reaction of monocyte adhesion to endothelial cells for simulating human aortas to decrease the risk factor of suffering cardiovascular disease (Koga and Meydani 2001). It is reported that to consume over 5–6 cups of drinking green tea on a daily basis is an improvement for the prevalence of a major metabolic and cardiovascular healthcare due to the beneficial effects of EGCG (Wolfram 2007). Kajima et al. (2017) proposed a quantitative approach of text mining, by using term-frequency, Jaccard index, and visualization, to analyze two main substances; one is the minutes of two local assemblies, both tea and sake, are local agricultural products in Japan, and the other is using a sake brewery for the purpose of promoting sales pamphlet. Their experimental results implied that the place-based identification for character assessment keeps level certification at sustainable products and rather finite at landscape-level perspective; both tea and sake have the health effects for nutritional medicine goal from the literature reviews. Given the above review, it is clear that tea with various compounds have a wide diversity of useful functions associated with real-life applications.

2.2 Text mining application

Prior to the introduction of text mining, this study clear its definitions to better see data mining. For its definition, it is a specific type of uncovering trends or patterns for mining data from huge datasets, and a key step for processing the practical knowledge discovery in databases (KDD) is addressed and modeled (Fayyad et al. 1996). It uses hybrid techniques of learning advanced assemble machine learning (ML) and artificial intelligence (AI) applications, or methods of statistical analysis to extract understandable patterns and useful information from data given. However, the term of "text mining" was first appeared in 1980s (Fan et al. 2006), and mainly applied to the issues of real-life science researches and governmental intelligence development. Main function of text mining is to deal with semi-structured or un-structured data highlighted as a text-based

document application (Hobbs et al. 1982). In particular, the un-structured data has a sustained growth because of the fast increase of multiple social media usages like Facebook, Instagram, and Twitter via Internet as well as a rapid increase in numerous publications, news, and web pages. Thus, to extract helpful features from the un-structured data is an interesting and important work trend, and this emerged requirement gains much focuses of researchers for searching an automated system to process it over the last years. Basic functions for text mining have the following main activities: Text categorization, document clustering, document summarization, and sentiment analysis.

In addition, a key term for text mining also needs to be accordingly identified. A well-trusted essential statistic weight approach referred for the term-frequency—inverse document frequency (TF-IDF) (Salton and McGill 1986) is good for information retrieval (IR) and text mining of data given, and it has a weighted factor used to measure the relevant importance from a word in a collection (or corpus) of documents addressed. Now, TF-IDF has a most prevalent scheme for weighting term, and it is extensively studied in beneficial purposes of reducing dimensions, selecting features, and extracting features, to raise effectively the mining accuracy (Jing et al. 2002). Although TF-IDF approach has realizable advantages of learning past research outcomes, there are several existing main drawbacks. (1) It is not necessarily required to have occasionally a high frequency from documents for an important term. (2) TF-IDF approach considers only a term-frequency to decide its importance. (3) The range of calculation is restricted only by the documents or works collected, and thus it is incomprehensive sufficiency to comprehend all the potential semantics. Based on coping with these disadvantages, this study uses an integration (hybrid) model to avoid it.

2.3 Analysis of literature review

The emerging techniques of text mining (Marcos-Pablos and García-Peñalvo 2020; Hao et al. 2018) are unfolded widely and quickly, and they are contingent with flourishing the literature analysis (Hao et al. 2018) successfully from researchers.

A LitMiner^{TM} system developed in Feldman et al. (2003) and used some text mining techniques to analyze the biomedical literatures to find out the relationships among diseases, drugs, genes, and proteins to accelerate a complex prediction for handling biological processes. A case study conducted in Delen and Crossland (2008) to use text mining to determine research tendency from the three well-known major managerial journals, including Journal of Management Information Systems, Information Systems Research, and MIS Quarterly, to help do literature analysis from a research subject given. Abbas et al. (2014) proposed a hybrid approach to combine text mining and data mining techniques helpful for providing a visualization-based technique and a taxonomy to classify these approaches via doing literature analysis from patent reviews. Amado et al. (2018) studied the literature reviews also to use text mining approaches to identify the key research climates from big data perspective in marketing domain and emphasized on five measurements: Big data, location of authors' affiliation, marketing, sectors, and products, to analyze their relationship to discover a hidden useful knowledge. The research result shows that applications of big data to marketing industry are still in an embryonic period. Thus, it has rooms for new initiatives with combining big data with marketing. Moro et al. (2015) employed text mining approach to organize relevance terms from intelligent fields for the business and banking industries and used latent Dirichlet allocation (LDA) to make up the terms as topics addressed to find out trends in the domains, and the analytical results show that bankruptcy, credit, and fraud risk are key application trends from viewpoint of the banking industry.

Based on the above retrospect and prospect, intelligent/objective techniques for topic modeling, such as effective LDA approach with using various types, for text mining to mine application purpose are investigated and emphasized.

2.4 Approach of topic model

Topic model approach uses a generative modeling for research application purpose covered under the published documents, and it also uses a probability procedure and application to unearth the important topics addressed from a collected document (Deerwester et al. 1990; Blei et al. 2003; Hofmann 1999). Thus, topic modeling approach can define large amount of un-structured data from texts to help identify the potential semantic variables for measurable and complex articles via effective dimension reduction approaches. In addition, to resolve weak points of TF-IDF mentioned in Sect. 2.2, three technical methods for reducing dimensions are raised through IR investigators for topic modeling, which are presented and summarized concisely, as follows.

(1) By exploring definition of latent semantic analysis (LSA) (Deerwester et al. 1990), a commonly used singular value decomposition (SVD) is needed to be introduced first. SVD is a generalization algorithm in linear algebra to take a factorization of real or complex rectangular matrixes from data given and expressed for a machine learning context. LSA applies a X matrix by SVD approach to determine a vector subspace from learning the context of TF-IDF feature spaces to capture and explain mostly the sum of variances from a given collection (Blei et al. 2003).

More importantly, LSA has driven varying success depended on application domains to contribute dimensionality reduction works from past text analytics methods and progresses the research development of the topic model approach. Unfortunately, SVD is to process a tedious job with time consuming; thus, LSA is incapable to overcome the problematic issue of polysemy on terms. (2) Probabilistic latent semantic analysis (PLSA), a probability-based topic model, proposed by Hofmann (Hofmann 1999), is a type of statistically generative model to use EM—expectation maximization algorithm to learn parameter estimation. Thus, PLSA has the real conception for a topic modeling approach to permit documents with multiple topics and solves the term polysemy problem for LSA approach. However, PLSA also has a main drawback. It will lead to an overfitting case when the corpus size increases and the parameters number also linearly increases with. (3) Regarding an extensive topic modeling approach, a LDA technique must be generally explored. Fundamentally, LDA inherits the concept of a generative model in PLSA (Xie et al. 2019; Blei et al. 2003), and it uses a hierarchical Bayesian model with three levels and adds a Dirichlet condition prior on distributing the topics for per-documents in a collection. It is needed for every document to first identify the possibility belonged to each topic and allocate the document to a most topic matched; it obviously heightens the defects of previous models like LSA and PLSA and has become the most commonly applied topic modeling approach nowadays. Figure 1 shows its graphical expression for LDA structure.

In Fig. 1, operators for mathematical symbols are interpreted, as follows. The M is a mth document from the sum of M documents; the N is a nth word from the sum of N words; α is a Dirichlet parameter prior to distribute the topic of per document; β is a Dirichlet parameter prior to distribute the word of per topic and accordingly generates the word assignation for this topic; θ is the topic assignation for this document M; z is this topic for a nth word in this document M; and w is this specific word.

Regarding LDA approach, two different types are used for estimating parameters, including variational expectation–maximization (VEM) method and Gibbs sampling method. LDA approach is followed by Blei et al. (2003) that proposed the VEM method. The Gibbs sampling method makes a common use by means of Bayesian inference in statistics, and it uses a randomization algorithm for statistical inference like the expectation–maximization (EM) algorithm. Thus, the Gibbs sampling method has a prevalent way and well use for estimating parameters from past literature review (Hao et al. 2018).

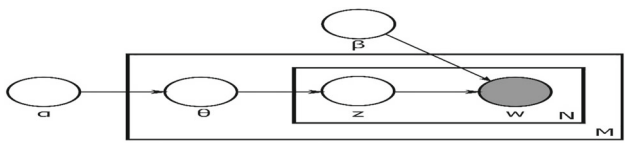

Fig. 1 Representation of LDA structure in a graphical model

3 Research methodology

The section mainly and clearly presents the basic concepts and the procedures for the proposed method for an integration model to examine the literature review for digging the related problems of tea and health.

3.1 Concepts of the proposed method

Traditional analysis of the literature review method has the following problems to be resolved. (1) It will take a long time for data preprocessing, (2) it will contain numerous papers or documents, and (3) the analysis processes for literatures are perplexed and amazed. To conquer these predicaments, techniques for text mining are emerged and helps analyzing plenty of documents mechanically, which thus attracts experts and researchers to use a variety of text mining methods, such as a notable topic model, to analyze the literature. For example, Choi et al. (2017) used it to dig out the relevancy between research tendencies and popular keywords from articles. Topic model manipulates a large of collections for data of un-structured text to find out the latent semantics from documents targeted; thus, it is suitable for the purposes of this study. On the one hand, LDA is a popular algorithm for common topic modeling (Alibasic et al. 2020; Wei et al. 2019); when it is contrasted to previous TF-IDF approach for considering term-frequency, LDA has more ability for quarrying and learning possible semantic analysis and the relevance within terms. On the other hand, the relation for the tea with purpose of health is not represented adequately in terms of term-frequency; thus, it requires a better objective modeling approach. This study offers a bridge over such a valuable issue and benefits the rationale for using LDA. LDA holds advantages of a most representative with a probabilistic algorithm for interpretable topics; in particular, it is the most widely used topic modeling approach to overcome the associated problems of tea and health. Conclusively, this study proposes a hybrid intelligent/objective method to integrate techniques of text mining, literature analysis, and topic modeling with different types of LDA approach. Thus, the proposed method has some broader benefits: It achieves efficient analysis from massive amounts of extant data for exploring latent information to identify entities and extract associations and trends to develop helpful knowledge objectively, increases and accelerates research process quickly, improves understanding for new hypotheses

effectively, and identifies key questions that need to be further researched.

3.2 Procedures of the proposed method

Figure 2 illustrates the procedures for the proposed method in this study; it is included four core/typical expressions (steps): data collection, preprocessing for data, topic modeling, and evaluation metrics. The first step for collecting data uses a release of Python 3.4 interpreter to scrambler and thus is saved as an Excel file in CSV format, and the other remaining steps are executed in R-package programming language. The four steps are introduced in detail with some examples step by step, as follows.

3.2.1 Step 1: Data collection

Initially, to select appropriate target journals indexed in a popular subject-specific database and to use the dataset adequately relevant to the computer science and social science fields is firstly and absolutely necessary. Therefore, the main data source of documents from a noted Web of Science (WoS; or called ISI Web of Knowledge) database is used for original experiment materials and datasets. Accordingly, this study writes and performs a web spider with the Python environment to select and collect data. The web spider is implemented to set the condition for "article" of document types from English-target journal papers during 2010–2017, but excluding papers for publication in conference proceedings and book citation indexes. In addition, the articles collected are mainly focused and set on the conditions of inputting keywords (or attributes) with the related effect issues of tea and health or disease from WoS, and then it uses the Query defined for selecting and extracting the existing literature in the following formation: "TI = (tea) AND TS = (memory OR fatigue OR metaboli* OR cardiovascular OR stomach OR intestin* OR digest* OR teeth OR cancer OR fat OR obesity OR slim* OR "lose weight" OR cataract OR diabetes OR stroke)," for ensuring this data collection.

Where the first term TI denotes title, and the second term TS denotes topics from WoS. By the way, tea can boost effects of body-related metabolism; thus, add "metaboli*" to TS for all related similar terms: metabolism, metabolic, metabolites, and metabolize. The rest * is done for the same manner.

Subsequently, only the Abstract part of documents from total of 2109 articles are collected and used as a complete dataset in 2010–2017 (8 years) for the experiments, and the title and keywords of articles are not involved to avoid bias for experimental results, because the Abstract part already includes the title and keywords in most cases of articles. Productively, the 2109 articles are employed for the target dataset for the following examinations. It is valuably noted

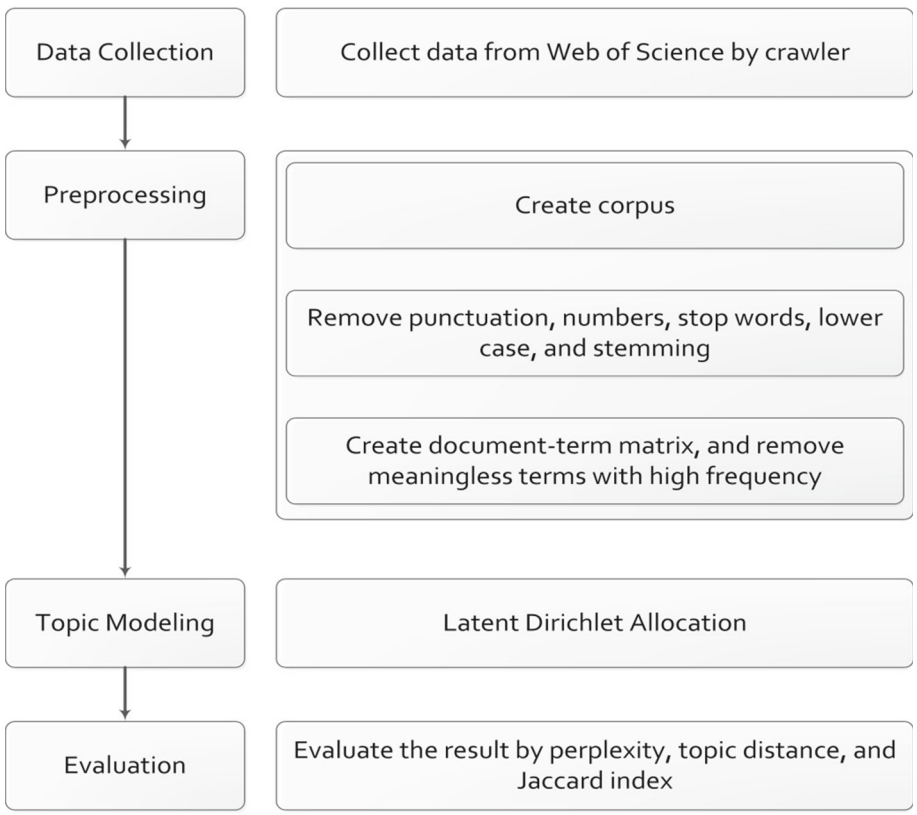

Fig. 2 Procedure flowchart of the proposed method

that the attributes for the input variables are the mentioned-above related keywords selected for documents, and all data types of the experiment model are belonged to characters of text line.

3.2.2 Step 2: Preprocessing for data

Completing the data collection, the articles are divided into three datasets to examine differences for issue of tea and health based on the three time periods: Whole dataset (2010–2017) of 2109 articles, the first half part dataset (2010–2013) with 955 articles, and the last half part dataset (2014–2017) with 1154 articles, for the application issue of tea and health. This step is implemented on "tm" package in R environment, and its pseudocode for the data preprocessing is listed and descripted briefly in Algorithm 1 as supplementary materials helpful for comprehensive understanding below.

Algorithm 1: Pseudocode of data preprocessing

Input: D = collected data, SL = stopwords list
Create a corpus from D;
Make all words in corpus lowercase;
Remove all punctuations in corpus;
Remove all numbers in corpus;
Remove all whitespace in corpus;
Stem all words in corpus;
Remove stopwords that accord to SL in corpus;
Create document-term matrix from corpus;
Examine terms with high frequency in document-term matrix, and remove meaningless terms from corpus;
Create document-term matrix from corpus again;
Remove sparse terms in document-term matrix.
Output: Document-term matrix

The procedure of data preprocessing is described in detail in the following four descriptions. (1) Datasets are employed to create corpus. (2) Do data preprocessing of texts, including removal processes for having the cases of punctuation, stop words, lower case, numbers, or stemming. In particular, the case of stop words represents a list of some most commonly used words, such as "a," "or," "at," "and," "the," and "about." They may be appeared heavily (over 200) within the articles, but they have not really meaningful for the application issue. Therefore, it will get a better result to remove the above stop words from the experiments when the frequency of the term is over 200 with a meaningless or insignificant term. Consequently, there are a total of 385 stop words removed in this step. Similarly, the stemming refers to reduce the developed full-term to stem based on an R function with a parameter 0.990. For example, all the "develop," "developed," "developing," and "development" are stemmed to the "develop." (3) Produce a document-term matrix (DTM) when the data preprocessing of texts is ended. Regarding DTM, the insignificance terms or making no sense terms, such as study, research, and paper, with high frequency will be found and removed preferentially. Although the above unimportant terms do not affect the effectiveness of the topic modeling, they make the modeling results harder to interpret. Furthermore, it is also needed to remove the term "tea" to preclude from influencing the modeling outcomes because this term appears in almost all articles collected. (4) Create DTM again and also remove sparse terms from it based on the above conditions. It is also noted that after the procedure of data preprocessing, the features extraction by the way of similar dimension reduction method mentioned-above is achieved. Consequently, the DTM result can be used to guild and model topics in the following steps.

3.2.3 Step 3: Topic modeling

This step applies two types of topic models to the experiments: LDA based on VEM method (LDA-VEM) and LDA using Gibbs sampling (LDA-Gibbs) (Blei and Lafferty 2005; Xu et al. 2020; Zahedi and Saraee 2018).

The following experiments are implemented by using R for the package "topicmodels" that has a high consistency to the package "tm." The output for the DTM provided in previous step is taken directly as an input variable for using the "topicmodels." Table 1 lists its related setting and detailed descriptions. Initially, to determine the number (K) of modeling topics is a more critical task for function-modeling. For the comparable effects of K (topics), this study computes the perplexity for learning various numbers of Ks (i.e., testing 25, 50, 75, 100, 150, 200, 250, 300, and 400, respectively) topics, and then chooses the result for the best numbers of topics from them. Accordingly, three evaluation metrics (i.e., perplexity, topic distance, and Jaccard index) are utilized to assess these topics obtained and will be used to the following step in detail.

Furthermore, it is helpful for readers to correlate the related information for Steps 1–3 of the algorithm mentioned-above for completing the main parts of the proposed method; thus, the core relation between usages of the hybrid algorithms with key functionality provided is presented with examples graphically (Fig. 3).

3.2.4 Step 4: Evaluation metrics

In this step, three metrics of evaluation standards with parameters are introduced for measuring the LDA results.

Table 1 Information of parameters for two types of topic modeling

	Characterization
LDA-VEM	
var	It controls the variational inference for a single document (default: iter.max = 500 and tol = 10^{-6})
em	It controls the variational EM algorithm (default: iter.max = 1000 and tol = 10^{-4})
initialization	For using one of "model," "random," and "seeded," here set "random"
estimate.alpha	Indicate if the parameter alpha is fixed for a priori or estimated value, and then set TRUE
LDA-Gibbs	
iter and thin	For both the respective number of Gibbs iterations and in-between Gibbs iterations, here set the value of 500
burnin	Set a beginning 0 at the number for the omitted Gibbs iterations
initialization	Set "random"

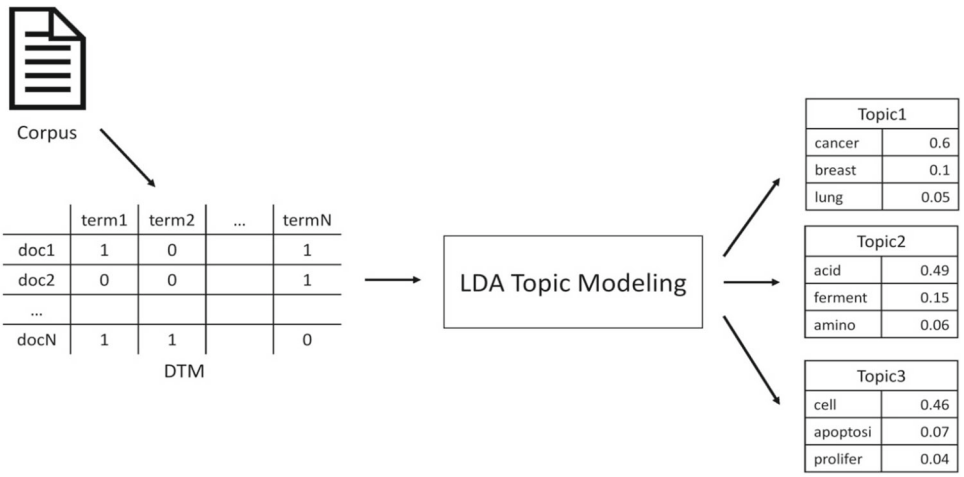

Fig. 3 The related process information of the proposed algorithm with key functionality

(1) *The perplexity*: The perplexity makes a purpose for modeling prediction by measuring probability distribution, and it has a monotonic decrease within the probability for the testing data and has algebraic equivalence to the opposite relation on the geometric mean for per word-probability (Blei et al. 2003). More importantly, the lower the perplexity is, the better generalization performance the number of topics increases with. Perplexity for forming a testing set on M documents is formatted in the Eq. (1) below:

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\}, \tag{1}$$

where D is a whole corpus, M means the document numbers within this testing corpus, N_d denotes the size d of the document (i.e., the words number), and $p(w_d)$ refers to the likelihood for these documents.

(2) *The topic distance*: It is used to select classifiable topics and is defined as Eq. (2) below:

$$\text{dist}\left(\text{topic}_j\right) = \text{dist}(\text{word})w + \text{dist}(\text{doc})(1-w), \tag{2}$$

where $\text{dist}(\text{topic}_j)$ refers to topic distance for the topic_j, dist(word) denotes word distance for the topic_j, dist(doc) means document distance for the topic_j, and w denotes a weight and is set up to 0.5 in this step. Both the equations of dist(word) and dist(doc) are formatted in the Eqs. (3) and (4) below, respectively:

$$\text{dist}(\text{word}) = \sqrt{\sum_{i}^{N_w}\left(p_{wi} - \frac{1}{N_w}\right)^2}, \tag{3}$$

where p_w denotes a word-probability, and N_w denotes terms number for the topic addressed.

$$\text{dist}(\text{doc}) = \sqrt{\sum_{i}^{N_d}\left(\frac{p_{di}}{\sum_{i=1}^{N_d} p_{di}} - \frac{1}{N_d}\right)^2}, \tag{4}$$

where p_d denotes a document probability, and N_d denotes documents number for the topic addressed.

(3) *The Jaccard index*: It is referred from the literature (Kajima et al. 2017; Kohsaka and Matsuoka 2015), and it is

utilized to analyze co-occurring relationships between terms and has a rational number for the discretization interval [0, 1]. Jaccard index, $J(X, Y)$, for the terms X and Y, is set as in the following Eq. (5):

$$J(X, Y) = \frac{A}{A + B + C}, \tag{5}$$

where A is the number of co-occurrences for the terms X and Y within a same text, B denotes the occurrences number for the term X, and C denotes the occurrences number for the term Y. Importantly, the larger Jaccard index indicates that the two terms have the higher correlation.

4 Empirical analysis with experimental result presentation

The section experiences the experiments with the proposed method for the collected tea and health data to implement data analysis and achieve experimental results and study findings.

4.1 Data analysis

The experimental data used and collected from WoS are focused on 2010–2017 having 2109 articles, which are divided into the three time periods of datasets, including 2010–2013, 2014–2017, and 2010–2017. In view of data analysis, Fig. 4 shows diagrammatically these authors' geographical location for knowing the classification of data in collected articles addressed for the issues of tea and health, which is based on nation or area helpful for readers to correlate the related information. Figure 4 shows that the top five numbers of articles published for this issue in famous journals are China (672 articles), USA (422), Japan (209), India (147), and South Korea (116). In them, it is interesting to find out four countries (1144) of the top five highest publications (1566) in Asia, and the four Asia countries exceeds half of all the 2109 articles. This information seems to bridge just the gap for providing a reasonable cause for tea originated from Asia. Figure 5 shows the numbers of publications in collected data every year. It is found that the publications for tea and health are increasing year after year. Accordingly, the experiments are performed for achieving its analysis results and study finding.

4.2 Experiment results

Based on the procedures of Steps 2–4 for the proposed method, all experiments are implemented in R, and the same preprocessing steps and parameter settings are used to the three datasets.

Subsequently, Table 2 and Figs. 6, 7 and 8 list the perplexities with different numbers (Ks) of topics after modeling topics. Table 2 and Figs. 6, 7 and 8 show that the variation for the perplexities of VEM method and Gibbs sampling method. Three best Ks are thus identified and selected. (1) The perplexities based on Gibbs sampling method are significantly lower than VEM method before K is equal to 150 (the low value for K represents the better) in the dataset of 2010–2017. Thus, this study chooses 50 to K with Gibbs sampling method to run experiments. (2) In the dataset of 2010–2013, the perplexities based on VEM method decrease steeply, and its values almost have a better outcome than those of Gibbs sampling. Therefore, when $K = 50$ in the VEM method is chosen, it has both low perplexities and low topics. (3) Regarding the 2014–2017 dataset, the line for the perplexity based on Gibbs sampling method is relatively flat; however, the VEM method is steeper. The more numbers the topics are, the lower perplexity it has; thus, $K = 50$ with a better result for the Gibbs sampling method is chosen. In summary, the LDA with Gibbs sampling method is successful as a powerful sampler when direct sampling is difficult, because the Gibbs sampling for estimating parameters benefits from some advantages. It can directly resolve the sample space for samples for each unknown parameter from the complicated conditional distributions, which are intractable, by means of importance-based resampling (Koch 2007), and it thus benefits a fast convergence.

The analytical results will only list the most distinctive of top 10 topics with first 20 terms as the more topics it has, the more complexities the models have. The top descriptive 10 topics selected are identified by the conditions of calculating the topic distance over the word/topic probability and the document/topic probability. The more descriptive or interesting topics are emphasized on the first 20 terms, which have the highest word-topic probabilities (beta) for each topic. Tables 3, 4 and 5 show the analytical results of topic modeling for differentiating the three datasets. The boldface letter in Tables 3, 4 and 5 shows something interesting and important, and it is worth noting attention to the potential impact for some interesting facts here. The topic distance is represented in parenthesis. The implications and findings will be explored in the following subsection.

Accordingly, Table 6 shows the top 20 term-frequencies and normalized term-frequencies in the three datasets. Table 6 also shows that three informational results is discovered and identified. (1) The first four terms of the three datasets are exactly same: Green, cell, egcg (i.e., EGCG), and cancer, with higher frequency of occurrence than others. (2) Five terms, such as catechin, treatment,

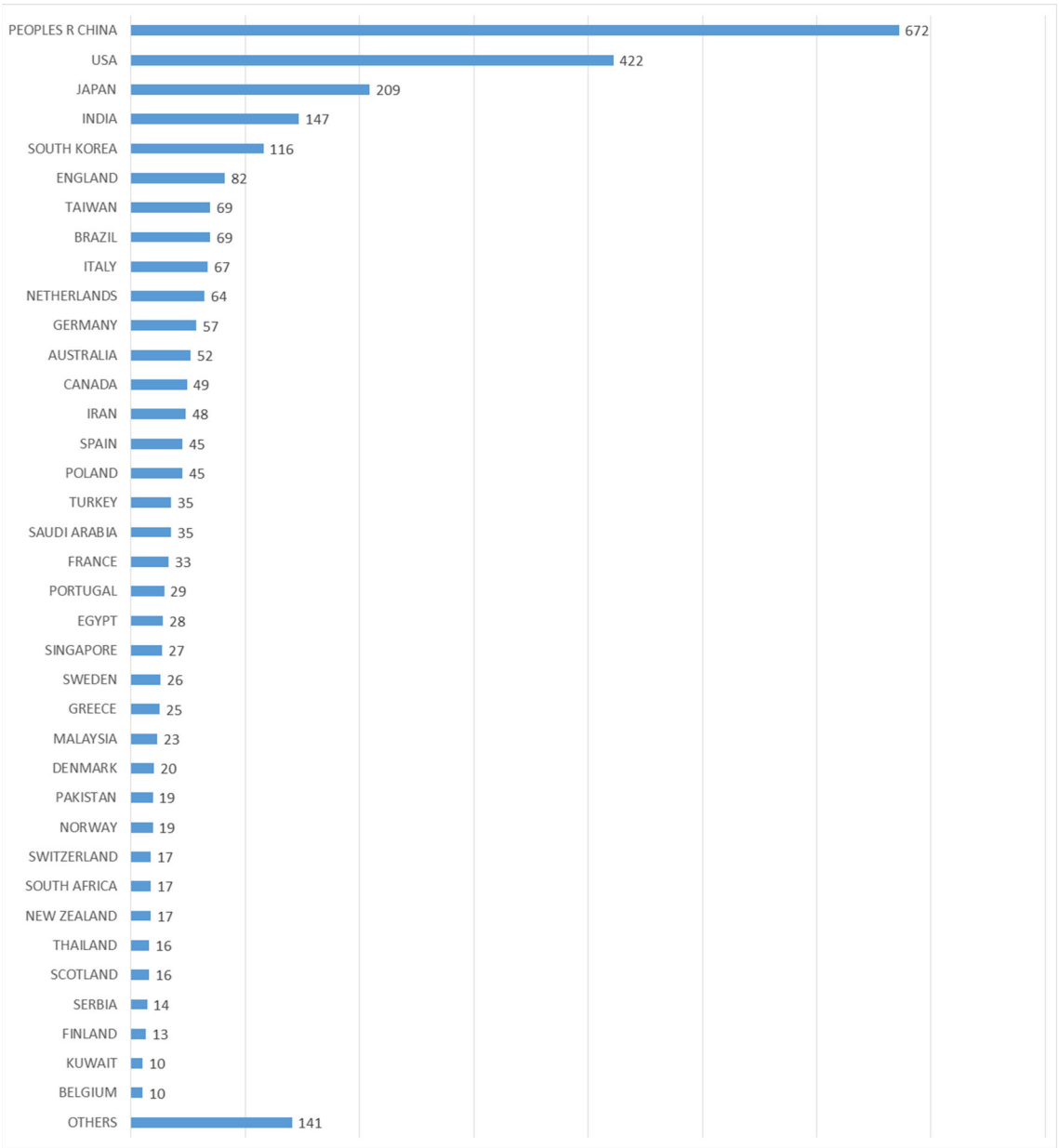

Fig. 4 Authors' geographical location for selected articles in 2010–2017

metabolism, protein, and gene, have a significant increase in term-frequency between the last two datasets: 2010–2013 and 2014–2017. (3) This phenomenon concern on the top 20 terms in 2010–2017 implies that these terms are still a research trend and the interests for researchers in coming recent years.

Nevertheless, a word cloud (or called tag cloud) is also a text mining approach to find out the most frequently used words from different sources like an article, and its procedure can be processed in function of R programming language. Subsequently, Figs. 9, 10 and 11 are word clouds created from the three datasets. These figures represent the hot terms with higher frequency in the three datasets. Figure 9 shows that the top four terms are also same (i.e., green, cell, egcg, and cancer), and the hot terms are found out the differentiation between the two figures: Figs. 10 and 11.

In this study, Jaccard index is used to measure the consistency that represents coefficients between tea term and individual terms. Table 7 lists the top 20 terms ranked with the highest Jaccard index that is calculated from the three datasets. In other words, this table shows the top 20 terms that co-occur more often with the tea than the others.

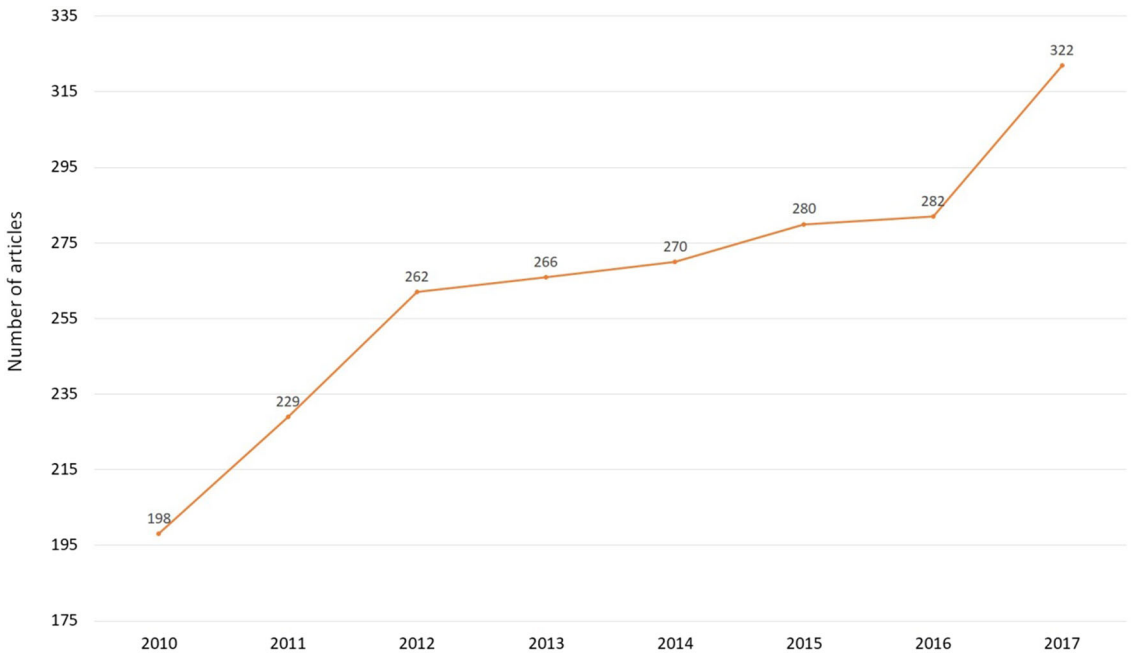

Fig. 5 Numbers of publications in collected data in 2010–2017

Table 2 Perplexities with different *K*s

K	2010–2017		2010–2013		2014–2017	
	VEM	Gibbs	VEM	Gibbs	VEM	Gibbs
25	507.8238	247.1948	214.1319	237.1799	236.3259	256.8378
50	244.5802	**202.1424**	**188.7429**	199.5894	218.7022	**212.6592**
75	242.2451	179.5979	192.9328	182.1514	214.6568	195.5684
100	240.3562	167.5289	153.9937	176.0092	179.9733	185.1748
150	151.9287	155.8329	106.7036	174.3366	123.4803	179.1017
200	136.1021	151.0141	88.7623	177.7913	104.5129	181.1312
250	122.0383	152.568	81.1798	185.8115	93.8285	188.7782
300	115.8954	154.0889	76.2804	189.8326	83.8043	193.9084
400	107.3306	160.9533	63.8197	198.9478	76.7376	201.5617

4.3 Experiment finding

After the complicated experiments were performed, some experimental findings with managerial issues are found and explicated in a list of enumeration descriptions.

(1) For the experiments, test different numbers (*K*s) on topics to calculate better perplexity. In particular, the perplexities of Gibbs sampling are flatter and more stable than others in most of cases in the three datasets. In contrast, perplexities of using VEM method are steep; that is, the higher *K* has the lower perplexities of it from the empirical results. It is found that although Gibbs sampling is more widely used, its perplexities do not necessarily better performance than using VEM method. In the managerial and theoretical basis, the above two methods should be further measured to different datasets for the comparable studies in practice.

(2) After the procedures of modeling topic, Tables 3, 4 and 5 show the empirical results for the listed topic number, its topic distance (in parenthesis), terms, and word-probability (beta). The first term in the distinctive topics has far higher word-probability than other terms in same topic, such as "green," "egcg," "catechin," "treatment," "cancer," "acid," and "cell." Thus, the key terms are found. Interestingly, these key terms have also high Jaccard index, and it is very probably for this result due to a high term-frequency of these terms. In summary, it is also found that terms with high frequency will have a high word-probability in this topic to make this topic distinctive and always have a high Jaccard index.

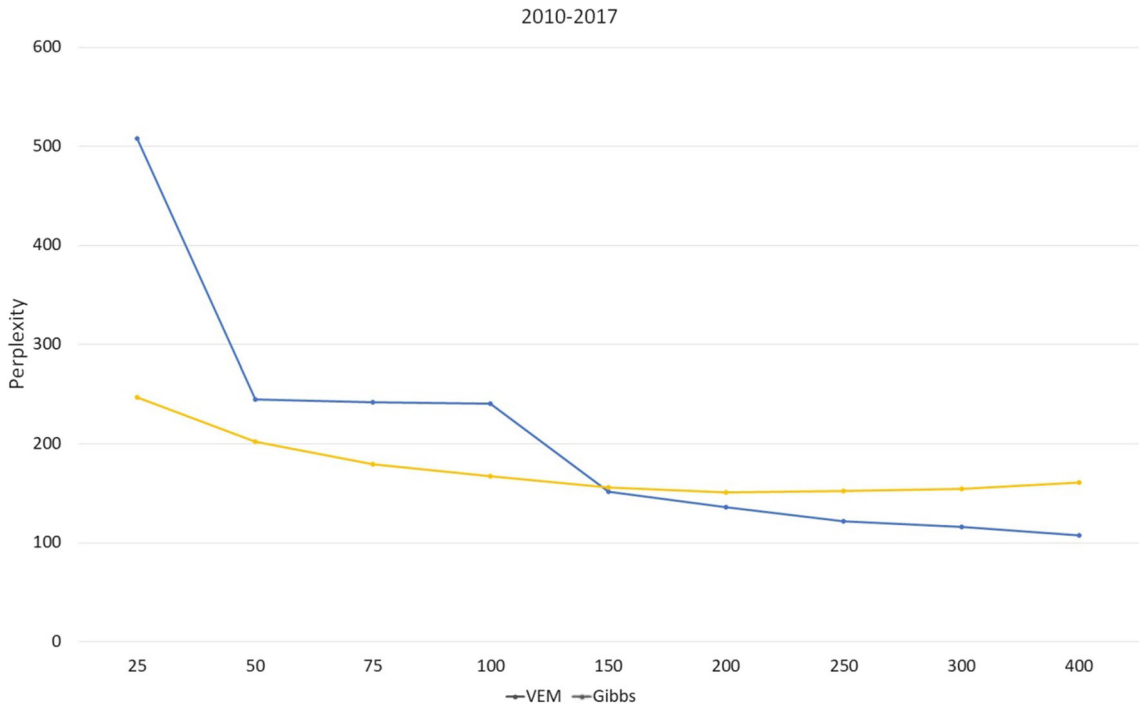

Fig. 6 Presentation on perplexity by using different Ks in the 2010–2017

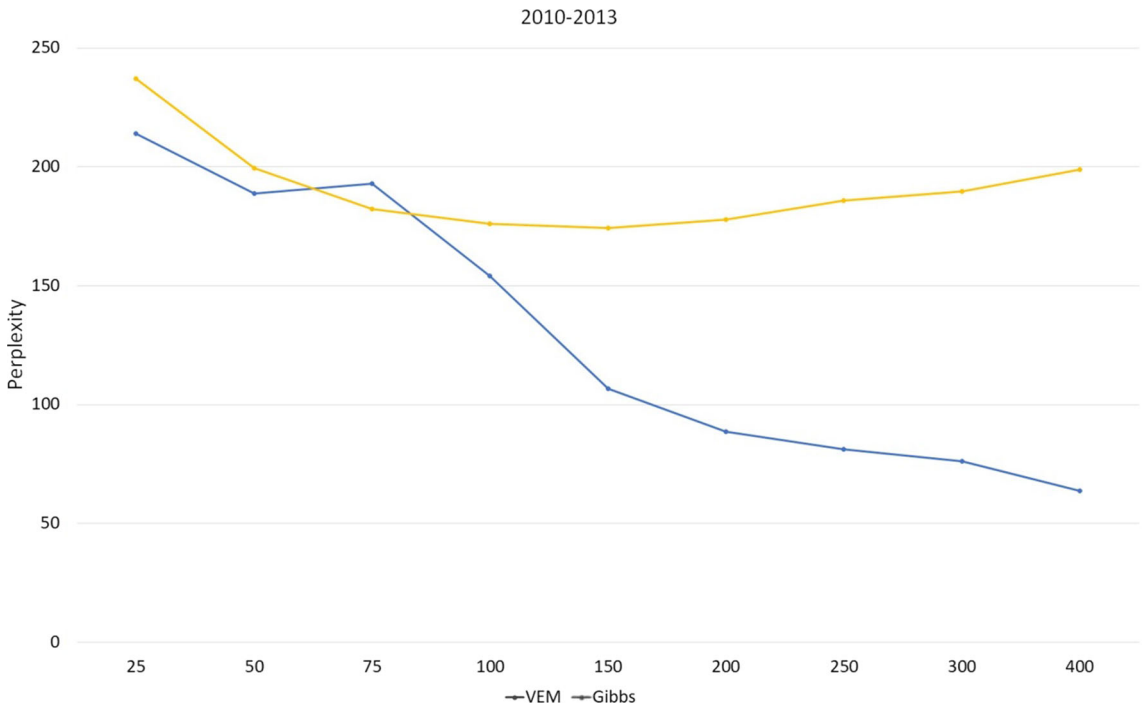

Fig. 7 Presentation on perplexity by using different Ks in the 2010–2013

(3) Tables 3, 4 and 5 show that it is found that terms related to tea compounds, green tea, and health as well as disease are identified as most topics. It is also found that some specific topics with great results highlight the following three study benefits and contributions. (a) Regarding Gibbs sampling in the 2010–2017 dataset, topic 10 has a main term "green" for an abbreviated green tea; it is found that the number of occurrence on researching green tea has much higher frequencies than those of

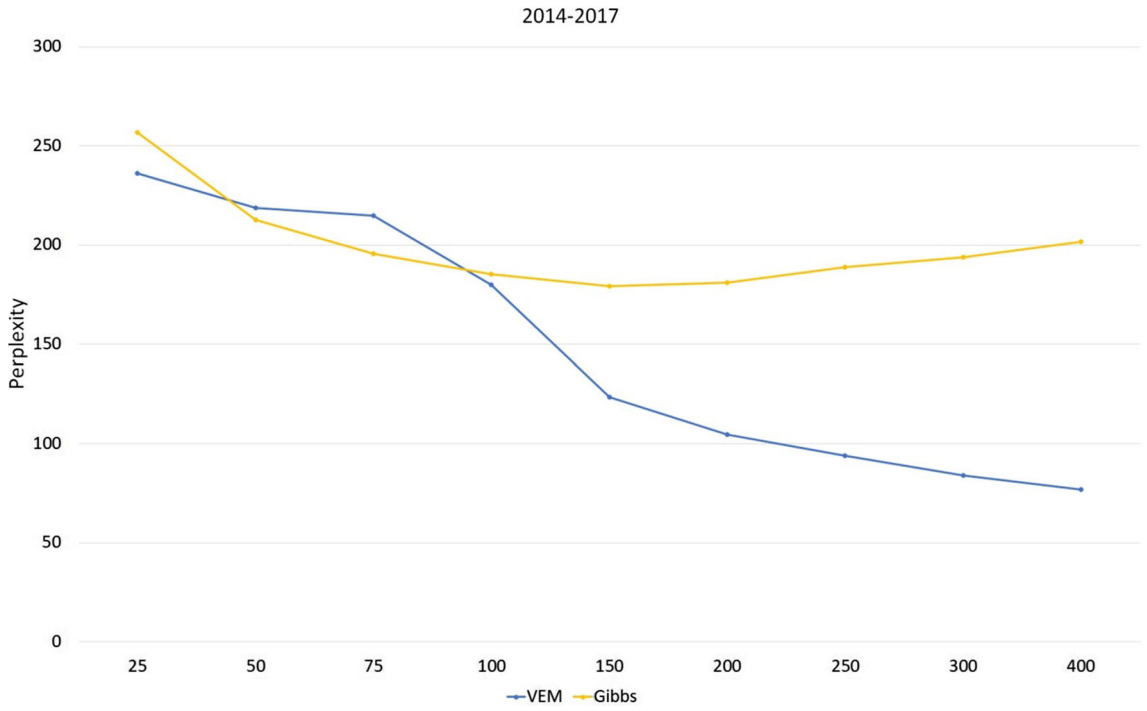

Fig. 8 Presentation on perplexity by using different Ks in the 2014–2017

studying other teas. This topic shows that many studies analyze pharmacological and therapeutic effects of green tea. Next, it is evidence that topic 41 shows a focus on the important compounds, EGCG, and EGCG is most studied in the field of pharmacology. Topic 29 has a good presentation in cancer concerns, and its core terms include breast cancer (Shelton and Badejo 2018; Cheng et al. 2019), lung adenocarcinoma (Cheng et al. 2019), colorectal cancer (Cheng et al. 2019; Alam et al. 2018), and ovarian cancer (Kiselev et al. 2018). From the above linkage, it is clear that tea is regarded as a critical objective for cancer prevention. Topic 13 shows that tea polyphenols has rich benefits on effect of inhibition. Topic 9 clearly indicates the relationship between fermentation and acid in tea and health. Topic 48 implies tea has an effect of treatment for alleviating disease. Topic 7 shows that catechin is an important compound in green tea, especially for EGCG. Topic 47 illustrates main effect of apoptosis and cytotoxicity that has anticancer benefits in tea. In topic 15, other teas (e.g., black tea and oolong tea) are addressed and introduced, and the main term "theaflavin" is mainly found from them. Topic 35 is focused and associated on the terms of green tea extract (GTE), EGCG, and tea polyphenol. (b) Using VEM method in the dataset 2010–2013, some main terms, such as green tea, EGCG, cancer, GTE, and acid, in each topic are found and are similar to those of the dataset 2010–2017. For example, for green tea, topic 36 mentions that it is regarded as a benefit of beverage. Topic 38 shows that EGCG has effects of inhibition and treatment. For cancer term, topic 50 mentions cancers with colon and adenocarcinoma, as well as chemoprevent and anticarcinogenic (Jiao et al. 2018). Regarding GTE, topic 31 shows it can help lose weight. Finally, for acid, topic 24 focuses on it and other compounds of tea, such as amino acid, phenol, quercetin, and flavonoid. Additionally, other topics also provide different descriptions. Topic 22 shows polyphenols as the main term, and other terms in this topic are mainly for the field of molecular biology. Black tea and oolong tea are found in topic 20, and theaflavin is mainly found in black tea. Topic 30 shows the effect of tea on diabetes using rat experiments. Topic 44 focuses on the antioxidant properties of tea (Munday 2016; Tian and Huang 2019). Accordingly, although most of terms in topic 7 have not a low word-probability, to find out the relationship between these terms is more difficult than other topics. (c) For using Gibbs sampling in the 2014–2017 dataset, the following main terms in each topic are similar to the previous two datasets, such as green (topic 20), EGCG (topic 25), catechin (topic 17), treatment (topic 4), cancer (topic 24), acid (topic 5), polyphenols (topic 35), cell (topic 40), and antioxidant (topic 14). Additionally, topic 41 is mainly focused on protein. Topic 25 has

Table 3 Analytical results of top 10 topics using Gibbs ($K = 50$) in the 2010–2017

Key topic	Core term	Beta value	Key topic	Core term	Beta value	Key topic	Core term	Beta value
10 (0.45)	**green**	0.9004	29 (0.31)	**cancer**	0.6083	9 (0.26)	**acid**	0.4981
	reduct	0.0054		**breast**	0.1027		**ferment**	0.1506
	inhibit	0.0025		**lung**	0.0526		**amino**	0.0668
	pharmacolog	0.0025		**colorect**	0.0375		gallic	0.0448
	comparison	0.0021		**ovarian**	0.0315		fatti	0.0322
	japanes	0.0017		incid	0.0237		**ascorb**	0.0112
	nuclear	0.0012		**anticarcinogen**	0.0090		tast	0.0078
	epigallocatechingal	0.0012		invas	0.0086		hplc	0.0068
	volunt	0.0012		**adenocarcinoma**	0.0082		except	0.0068
	peopl	0.0012		critic	0.0056		convers	0.0063
	detail	0.0012		hypothesi	0.0043		predomin	0.0063
	hypothesi	0.0012		epidemiolog	0.0034		domin	0.0054
	malondialdehyd	0.0012		tumor	0.0030		fecal	0.0054
	center	0.0012		stratifi	0.0030		uniqu	0.0049
	prevent	0.0008		inhibitor	0.0022		basi	0.0044
	vitro	0.0008		final	0.0022		monitor	0.0039
	take	0.0008		modifi	0.0022		highlight	0.0039
	therapeut	0.0008		tract	0.0022		still	0.0034
	anticarcinogen	0.0008		event	0.0017		volatil	0.0034
	even	0.0008		continu	0.0017		peak	0.0029
41 (0.35)	**egcg**	0.6888	13 (0.29)	**polyphenol**	0.5760	48 (0.26)	**treatment**	0.4955
	epigallocatechingal	0.0831		inhibit	0.0696		**treat**	0.1810
	epigallocatechin	0.0311		calcium	0.0228		**therapi**	0.0211
	cell	0.0307		particular	0.0211		**efficaci**	0.0168
	gallat	0.0188		support	0.0206		untreat	0.0146
	synergist	0.0131		rich	0.0206		end	0.0141
	dosedepend	0.0078		**benefici**	0.0137		stain	0.0130
	inhibit	0.0057		oxidas	0.0109		challeng	0.0119
	progress	0.0045		better	0.0086		access	0.0103
	pharmacolog	0.0045		treatment	0.0069		side	0.0086
	kda	0.0045		purifi	0.0069		less	0.0081
	physiolog	0.0037		expect	0.0063		restor	0.0070
	synthas	0.0033		degrad	0.0057		along	0.0059
	mode	0.0029		intracellular	0.0057		immun	0.0049
	medium	0.0024		second	0.0051		name	0.0049
	nuclear	0.0020		undergo	0.0051		littl	0.0043
	matrix	0.0020		far	0.0051		diminish	0.0043
	obes	0.0020		consequ	0.0051		led	0.0038
	implic	0.0020		extens	0.0046		partial	0.0038
	recogn	0.0020		hypothes	0.0046		section	0.0038

Table 3 continued

Key topic	Core term	Beta value	Key topic	Core term	Beta value
7 (0.24)	**catechin**	0.4542	15 (0.23)	**black**	0.4475
	gallat	0.1145		**theaflavin**	0.0923
	epigallocatechin	0.0916		**oolong**	0.0725
	epicatechin	0.0662		red	0.0353
	egc	0.0517		benefit	0.0295
	ecg	0.0445		regard	0.0205
	galloyl	0.0178		unit	0.0199
	gallocatechin	0.0148		shift	0.0096
	epicatechingal	0.0110		basi	0.0096
	moieti	0.0085		remark	0.0090
	green	0.0059		lowest	0.0071
	dure	0.0038		physiolog	0.0064
	decaffein	0.0038		veri	0.0064
	brew	0.0034		make	0.0058
	circul	0.0025		benefici	0.0058
	detail	0.0025		countri	0.0058
	hypothesi	0.0025		enzymat	0.0051
	incorpor	0.0021		nutraceut	0.0051
	epitheli	0.0021		clear	0.0051
	understood	0.0021		good	0.0045
47 (0.23)	**cell**	0.4601	35 (0.22)	**gte**	0.4013
	apoptosi	0.0691		**green**	0.2015
	prolifer	0.0403		uptak	0.0433
	anticanc	0.0297		acut	0.0427
	inhibit	0.0283		prior	0.0185
	cytotox	0.0281		enrich	0.0139
	caspas	0.0248		prolong	0.0104
	death	0.0224		mark	0.0092
	viabil	0.0177		unknown	0.0087
	carcinoma	0.0174		convers	0.0081
	cycl	0.0165		reduct	0.0069
	assay	0.0163		note	0.0064
	apoptot	0.0148		oxygen	0.0064
	hepg	0.0146		hypothesi	0.0064
	antiprolif	0.0134		shortterm	0.0064
	antitumor	0.0118		calcul	0.0052
	arrest	0.0106		fulli	0.0052
	migrat	0.0099		greater	0.0046
	mcf	0.0099		leptin	0.0046
	cellular	0.0094		area	0.0040

the anticancer effect of EGCG (Rady et al. 2018). Topic 24 demonstrates the impact of chemoprevent in tea on cancer. Moreover, topic 35 shows a meaningful finding: Tea polysaccharides are often easily ignored in early year researches, because most of researches focus on tea polyphenols. Until recent years, the researches on tea polysaccharides have made a continual increase, especially with a comparison on tea polyphenols. Topic 14 shows that both phenols and flavonoids have antioxidant properties; at the same time, diphenylpicrylhydrazyl (DPPH) is often used to estimate antioxidant activity.

(4) Table 6 shows the term-frequency and normalized term-frequency in the three datasets. The same top

Table 4 Analytical results of top 10 topics using VEM ($K = 50$) in the 2010–2013

Key topic	Core term	Beta value	Key topic	Core term	Beta value	Key topic	Core term	Beta value
36 (0.48)	**green**	0.9218	31 (0.30)	**polyphenol**	0.5298	22 (0.28)	**rat**	0.4798
	beverag	0.0239		complex	0.0727		**diabet**	0.1649
	benefici	0.0091		**molecular**	0.0548		administr	0.0456
	featur	0.0064		hydrogen	0.0329		wistar	0.0288
	comparison	0.0059		vitro	0.0161		**treat**	0.0238
	monitor	0.0057		bond	0.0157		male	0.0232
	medium	0.0051		red	0.0148		theanin	0.0228
	short	0.0029		understand	0.0111		administ	0.0172
	area	0.0029		rich	0.0103		divid	0.0116
	literatur	0.0028		tree	0.0088		fed	0.0107
	allow	0.0024		benefici	0.0088		intraperiton	0.0099
	prove	0.0020		sinc	0.0078		serum	0.0093
	volunt	0.0019		better	0.0076		streptozotocin	0.0089
	expect	0.0017		stain	0.0073		root	0.0086
	act	0.0013		simul	0.0073		prevent	0.0085
	therebi	0.0011		mixtur	0.0072		inject	0.0083
	initi	0.0011		certain	0.0072		immunohistochem	0.0071
	conclud	0.0005		physiolog	0.0071		spraguedawley	0.0070
	compris	0.0001		elucid	0.0068		dysfunct	0.0069
	end	0.0001		basi	0.0067		stain	0.0057
38 (0.33)	**egcg**	0.5965	20 (0.29)	**black**	0.4967	7 (0.28)	dure	0.4830
	epigallocatechingal	0.0914		**theaflavin**	0.1259		least	0.0523
	green	0.0896		fiber	0.0336		acet	0.0503
	gallat	0.0301		declin	0.0254		loss	0.0399
	inhibit	0.0198		**oolong**	0.0248		advers	0.0361
	treatment	0.0194		recommend	0.0113		gastrointestin	0.0271
	therapeut	0.0111		solubl	0.0113		contrast	0.0197
	act	0.0090		nitrogen	0.0110		basi	0.0187
	quercetin	0.0086		area	0.0106		point	0.0159
	molecular	0.0071		characterist	0.0095		final	0.0140
	mous	0.0060		brew	0.0092		characterist	0.0139
	abund	0.0052		secondari	0.0092		regard	0.0138
	bioavail	0.0050		flavor	0.0091		second	0.0135
	dosedepend	0.0050		comparison	0.0089		stomach	0.0134
	synergist	0.0049		second	0.0085		intens	0.0133
	vitro	0.0044		treat	0.0085		taken	0.0132
	futur	0.0039		beverag	0.0083		extent	0.0107
	support	0.0034		cultiv	0.0083		throughout	0.0104
	biolog	0.0032		except	0.0081		antidiabet	0.0096
	egcginduc	0.0030		grow	0.0078		mark	0.0086

Table 4 continued

Key topic	Core term	Beta value	Key topic	Core term	Beta value
50 (0.27)	**cancer**	0.4773	44 (0.26)	**antioxid**	0.4550
	colon	0.0736		ethanol	0.0692
	chemoprevent	0.0656		renal	0.0479
	colorect	0.0611		scaveng	0.0463
	prevent	0.0409		dpph	0.0259
	incid	0.0360		prevent	0.0240
	carcinogenesi	0.0287		injuri	0.0231
	adenocarcinoma	0.0115		tradit	0.0223
	biolog	0.0111		assay	0.0189
	agent	0.0097		vivo	0.0170
	anticarcinogen	0.0087		**oxygen**	0.0158
	epidemiolog	0.0075		crude	0.0154
	peopl	0.0072		weight	0.0146
	modifi	0.0068		power	0.0141
	phytochem	0.0064		chines	0.0111
	reduct	0.0063		solubl	0.0109
	conclud	0.0062		rich	0.0093
	stomach	0.0052		hydroxyl	0.0085
	site	0.0051		neutral	0.0083
	lung	0.0051		organ	0.0080
30 (0.27)	**gte**	0.4203	24 (0.26)	**acid**	0.4544
	green	0.0817		**phenol**	0.1321
	injuri	0.0344		**flavonoid**	0.0590
	fat	0.0328		gallic	0.0435
	storag	0.0280		**amino**	0.0401
	sensori	0.0238		**quercetin**	0.0199
	inhibit	0.0230		vitamin	0.0147
	protein	0.0189		fatti	0.0143
	oxidas	0.0142		dietari	0.0128
	enrich	0.0126		plant	0.0122
	supplement	0.0116		treatment	0.0117
	prolong	0.0111		watersolubl	0.0092
	prior	0.0092		nutrit	0.0086
	dure	0.0091		absorb	0.0084
	cycl	0.0084		ingredi	0.0075
	lipid	0.0082		flavanol	0.0067
	train	0.0080		exert	0.0063
	seen	0.0075		gkg	0.0062
	still	0.0071		veri	0.0062
	gkg	0.0069		particular	0.0058

four terms in each dataset: Green, cell, EGCG, and cancer are found out. The empirical results show that green tea is the most use of teas for the purpose of health in managerial decision-making from wide studies, and the second is for the cell. In addition, in terms of the compounds of tea, it is found that EGCG is the maximal resource; the following concern is to the tea polyphenols, catechins, and metabolism. Importantly, it is worth noting that in recent years, tea polyphenols has decreased slightly, but conversely catechins has increased significantly, and metabolism has a significant increase. In managerial implications, the main effect of tea is focused on antioxidant activity and treatment for the best

Table 5 Analytical results of top 10 topics using Gibbs ($K = 50$) in the 2014–2017

Key topic	Core term	Beta value	Key topic	Core term	Beta value	Key topic	Core term	Beta value
20 (0.43)	**green**	0.8497	17 (0.29)	**catechin**	0.5564	24 (0.25)	**cancer**	0.4684
	seen	0.0069		**green**	0.0773		**breast**	0.0736
	clinic	0.0055		galloyl	0.0261		**tumor**	0.0402
	less	0.0049		metabolit	0.0189		**lung**	0.0395
	central	0.0028		moieti	0.0153		**prostat**	0.0388
	certain	0.0028		recoveri	0.0135		**colon**	0.0335
	except	0.0021		strong	0.0099		**ovarian**	0.0315
	restrict	0.0021		local	0.0099		**chemoprevent**	0.0268
	biolog	0.0021		comparison	0.0081		**colorect**	0.0201
	ecg	0.0021		advanc	0.0081		carcinogenesi	0.0194
	glycosid	0.0021		degrad	0.0072		implic	0.0134
	everi	0.0021		molecular	0.0063		agent	0.0121
	side	0.0021		complet	0.0054		complet	0.0087
	mrna	0.0014		consequ	0.0054		unit	0.0087
	array	0.0014		ring	0.0054		local	0.0047
	chlorophyl	0.0014		clarifi	0.0054		epidemiolog	0.0040
	disrupt	0.0014		elucid	0.0045		progress	0.0034
	divid	0.0014		male	0.0045		wiley	0.0034
	bodi	0.0014		meal	0.0045		promis	0.0034
	uniqu	0.0014		chronic	0.0045		molecular	0.0027
25 (0.32)	**egcg**	0.6124	4 (0.26)	**treatment**	0.5036	5 (0.25)	**acid**	0.4715
	cell	0.0784		**treat**	0.0960		**amino**	0.0706
	epigallocatechingal	0.0660		reduct	0.0250		**fatti**	0.0549
	promot	0.0148		intens	0.0173		volatil	0.0358
	molecul	0.0109		end	0.0144		gallic	0.0349
	abl	0.0101		exert	0.0125		dark	0.0288
	anticanc	0.0093		point	0.0116		aroma	0.0175
	mediat	0.0062		negat	0.0116		**quercetin**	0.0166
	exert	0.0055		untreat	0.0087		hplc	0.0070
	futur	0.0055		doe	0.0068		**phenylalanin**	0.0070
	stem	0.0055		futur	0.0068		gcms	0.0061
	tumor	0.0039		western	0.0068		reach	0.0061
	block	0.0039		safe	0.0068		even	0.0053
	enzymat	0.0039		diminish	0.0058		probabl	0.0053
	subsequ	0.0031		featur	0.0058		account	0.0053
	delta	0.0031		microscop	0.0058		degrad	0.0044
	phosphoryl	0.0031		accompani	0.0048		understood	0.0044
	link	0.0031		slight	0.0048		reason	0.0044
	least	0.0031		coli	0.0048		asian	0.0035
	biomark	0.0031		mark	0.0048		matter	0.0035

Table 5 continued

Key topic	Core term	Beta value	Key topic	Core term	Beta value
41 (0.23)	**protein**	0.4407	40 (0.21)	**cell**	0.3966
	anthocyanin	0.0467		**apoptosi**	0.0706
	abund	0.0415		prolifer	0.0521
	synthesi	0.0374		**cytotox**	0.0330
	upregul	0.0301		inhibit	0.0317
	proteom	0.0187		caspas	0.0299
	enzym	0.0166		assay	0.0244
	hplc	0.0146		death	0.0217
	comparison	0.0115		gml	0.0181
	blot	0.0115		carcinoma	0.0181
	downregul	0.0094		anticanc	0.0154
	crude	0.0073		viabil	0.0149
	albumin	0.0073		hepg	0.0145
	biochem	0.0063		mcf	0.0136
	enrich	0.0063		apoptot	0.0127
	facilit	0.0063		intracellular	0.0095
	revers	0.0052		antiprolif	0.0095
	synthas	0.0052		migrat	0.0095
	chain	0.0052		cycl	0.0091
	explain	0.0052		bax	0.0091
35 (0.22)	**polyphenol**	0.4372	14 (0.18)	**antioxid**	0.3034
	polysaccharid	0.0490		**phenol**	0.1478
	membran	0.0443		**flavonoid**	0.1209
	mitochondri	0.0434		scaveng	0.0539
	link	0.0198		**dpph**	0.0426
	promot	0.0189		assay	0.0348
	dosedepend	0.0170		mgg	0.0139
	toward	0.0132		abt	0.0139
	attract	0.0123		loss	0.0105
	chemic	0.0113		**diphenylpicrylhydrazyl**	0.0096
	solut	0.0104		antibacteri	0.0087
	prevent	0.0095		distribut	0.0079
	practic	0.0095		rich	0.0070
	enzymat	0.0095		green	0.0053
	administr	0.0085		nutrit	0.0053
	nutraceut	0.0085		phytochem	0.0053
	strong	0.0076		electron	0.0044
	act	0.0066		potent	0.0044
	oxid	0.0066		impli	0.0044
	morpholog	0.0066		cellular	0.0035

interests of healthcare considerations; especially, effect of treatment has received greater attention in recent years when compared to the other ingredients.

(5) Table 7 shows the analytical results of Jaccard index to illustrate high co-occurring relationships between terms and tea and represent the focus of research in

Table 6 The term-frequency and normalized term-frequency in the three datasets

Rank	2010–2017			2010–2013			2014–2017		
	Term	Freq.	Norm.	Term	Freq.	Norm.	Term	Freq.	Norm.
1	**green**	2971	1	**green**	1482	1	**green**	1489	1
2	**cell**	2041	0.6870	**cell**	1044	0.7045	**cell**	997	0.6696
3	**egcg**	1689	0.5685	**egcg**	897	0.6053	**egcg**	792	0.5319
4	**cancer**	1424	0.4793	**cancer**	720	0.4858	**cancer**	704	0.4728
5	catechin	1089	0.3665	polyphenol	544	0.3671	**catechin**	630	0.4231
6	polyphenol	1060	0.3568	inhibit	522	0.3522	acid	557	0.3741
7	acid	1035	0.3484	acid	478	0.3225	**treatment**	543	0.3647
8	antioxid	969	0.3262	rat	477	0.3219	antioxid	524	0.3519
9	inhibit	951	0.3201	**catechin**	459	0.3097	polyphenol	516	0.3465
10	treatment	940	0.3164	antioxid	445	0.3003	risk	500	0.3358
11	risk	923	0.3107	risk	423	0.2854	**metabol**	472	0.3170
12	rat	897	0.3019	**treatment**	397	0.2679	**protein**	471	0.3163
13	protein	855	0.2878	**protein**	384	0.2591	inhibit	429	0.2881
14	metabol	790	0.2659	black	364	0.2456	rat	420	0.2821
15	black	700	0.2356	gte	327	0.2206	**gene**	403	0.2707
16	mice	698	0.2349	**metabol**	318	0.2146	mice	402	0.2700
17	gte	695	0.2339	mice	296	0.1997	plant	402	0.2700
18	oxid	636	0.2141	intak	292	0.1970	gte	368	0.2471
19	diet	610	0.2053	oxid	290	0.1957	diet	354	0.2377
20	gene	603	0.2030	glucos	275	0.1856	oxid	346	0.2324

The 'Freq.' refers to frequency, and the 'Norm.' refers to normalization

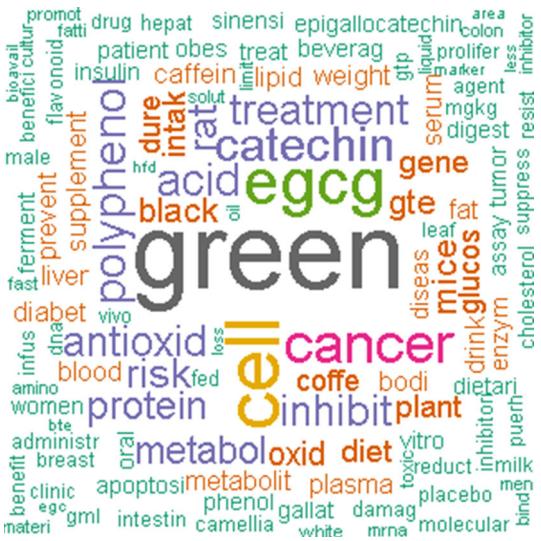

Fig. 9 Word cloud in the 2010–2017

Fig. 10 Word cloud in the 2010–2013

collected articles. The most of terms from the results of Jaccard index are consistent with higher term-frequency of topic modeling method. In the analytical results, "green" is much higher occurrence than other terms because it is co-occurred with tea and regarded as green tea. Furthermore, it is found that although EGCG has a high term-frequency, its Jaccard index is not high; this interesting lesson is valuable to further analysis or exploration in the subsequent work. Nevertheless, in differentiating implicitly the gaps for the 2010–2013 dataset and the 2014–2017 dataset, they have partially similar results on term-frequency from the results of Jaccard index. For example, tea polyphenols has a decreasing

Fig. 11 Word cloud in the 2014–2017

presentation, and treatment, catechins, antioxidant, and metabolism are increased oppositely. This phenomenon of analytical result also supports that researchers have more interests on treatment and metabolism for recent years.

5 Conclusions

This study proposed a hybrid model for integrating an overall literature review method, text mining techniques, and the LDA method to conduct an overview of systematic analyses to identify the effects of tea and health for healthcare benefits. A total of 2109 articles were collected from WoS and published between 2010 and 2017. These were catalogued into three datasets to implement various experiments for learning and benefiting the differentiation process to unearth managerial findings or some unexplained, contradictory, or counterintuitive facts. In accordance with the experimental procedures, the data preprocessing and topic modeling were first implemented and performed in an R environment. Afterward, perplexity was used to evaluate performance of two parameter estimation methods (VEM and Gibbs sampling) for using LDA with measuring different Ks. After the scheme for choosing the best topic model, topic distance was followed and used to select distinctive topics. Simultaneously, use the Jaccard index to evaluate consistency and examine co-occurrence relationships between terms for finding out the influence and the effectiveness for concerning tea and health.

For the practical application of the proposed method, this study is concerned with the issue of tea and healthcare from analysis for a literature review. In the empirical results, terms with both a high frequency and a high

Table 7 Analytical results of Jaccard index for the tea in the three datasets

Rank	2010–2017		2010–2013		2014–2017	
	Term	Jaccard	Term	Jaccard	Term	Jaccard
1	green	0.5380	green	0.5748	green	0.5098
2	polyphenol	0.2708	polyphenol	0.2982	treatment	0.2577
3	cell	0.2559	cell	0.2753	polyphenol	0.2493
4	treatment	0.2428	inhibit	0.2592	acid	0.2444
5	acid	0.2388	acid	0.2331	cell	0.2409
6	inhibit	0.2263	cancer	0.2259	metabol	0.2256
7	metabol	0.2136	treatment	0.2255	antioxid	0.2140
8	antioxid	0.2095	protein	0.2054	catechin	0.2114
9	catechin	0.2066	antioxid	0.2048	inhibit	0.2003
10	cancer	0.2040	catechin	0.2017	cancer	0.1866
11	protein	0.1940	metabol	0.2	protein	0.1854
12	egcg	0.1587	egcg	0.1816	oxid	0.1557
13	risk	0.1575	rat	0.1719	risk	0.1551
14	oxid	0.1540	black	0.1699	egcg	0.1404
15	black	0.1476	risk	0.1611	gene	0.1319
16	rat	0.1401	oxid	0.1525	plant	0.1298
17	mice	0.1207	intak	0.1387	diet	0.1210
18	diet	0.1187	mice	0.1228	mice	0.1195
19	gene	0.1140	glucos	0.1120	rat	0.1145
20	gte	0.0592	gte	0.0629	gte	0.0565

Jaccard index were mined out deeply, such as "green," "egcg," "catechin," "treatment," "cancer," "acid," and "cell," for the distinctive topics. Regarding the research findings from the empirical results, five focal points are highlighted. (1) From a technical viewpoint, Gibbs sampling does not display better performance than the VEM method for perplexities of an evaluation standard. (2) Terms with high frequency have high word-probabilities and high Jaccard index for the same topic. (3) Key terms related to tea compounds, green tea, and health or disease are identified as common topics. (4) The same top four terms in the experienced three datasets are green (tea), cell, EGCG, and cancer. For managerial decision-making, the main effect of tea is emphasized for its antioxidant activity and natural therapies. (5) The analytical result supports that researchers have greater interest in treatment and the metabolism with tea.

On the one hand, for the managerial implications, there are three key directions identified. (1) Based on search query results, the implication is that cancer is the most significant disease identified, and it was focused on the therapeutic effects or the natural impact of tea and health from the collected articles. Thus, tea is a useful and alternative treatment for serious illnesses, such as cancer. (2) Tea compounds describe the effects or treatments that were analyzed and discovered the key compositions of EGCG, polyphenols, and catechins. (3) The three datasets (2010–2013, 2014–2017, and 2010–2017) identified differentiating changes and gaps by studying research trends. For example, the analytical results from the 2014–2017 dataset showed that researchers were more focused on learning about tea as a treatment than in the 2010–2013 dataset. On the other hand, eight core concerns are concluded to the research application regarding this study's contribution and novelty. (1) It is rare to use LDA for topic modeling from real experience of limited literature analysis conducted for text mining to examine the effects of tea and health. (2) The empirical evidence suggests that tea has natural and important preventive impacts on disease treatments, especially for cancer. This study also confirmed from past research that tea could be taken as a botanical herb. (3) Regarding the research novelty, it fills a gap in objective knowledge-based text mining and topic modeling techniques with hybrid models on tea and health than previous research that is only based on traditional content analysis methods. (4) For this study's strengths, it achieves reductions of manpower cost and relative objectivity for using objective LDA method for topic modeling when compared to past studies. (5) For this study's significance, the proposed method is beneficial for efficient analysis from massive amounts of extant data to explore latent information to develop knowledge effectively, accelerate the research process quickly, improve understanding for new hypotheses objectively, and identify key questions for further research. (6) Conclusive remarks and research importance for this study, it offers the new rationale for the application in the medical field; and the three datasets identified their differentiations for studying research trends. (7) The advantages of the proposed method include effective hybrid method, objective text mining technique, and well-evaluation index with good application performance. (8) From the application research, this study is a good example of the various approaches used for topic modeling and text mining with effective evaluation methods similar to previous researches in (Vo and Ock 2015; Rashid et al. 2019; Bastani et al. 2019). As for its disadvantage, this study is limited because of its longitudinal time frame, and this requires further subsequent experiments in the future.

Although this study identified a few advances for tracking new research areas of tea and health or disease and for achieving positive results of sufficiently empirical evidence, it can still be challenged by future work: (1) Creating dictionaries of terms to heighten the standard of modeling topics, (2) making use of other methods like the study of hierarchical Dirichlet process (HDP) in statistics and other machine learning fields in (Xuan et al. 2019; Sangaiah et al. 2019c), examining and defining the best number for topics identified, and (3) adding other topic models into experiments conducted, such as a correlated topic model (CTM) (Blei and Lafferty 2005).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abbas A, Zhang L, Khan SU (2014) A literature review on the state-of-the-art in patent analysis. World Patent Inf 37:3–13. https://doi.org/10.1016/j.wpi.2013.12.006

Alam M-N, Almoyad M, Huq F (2018) Polyphenols in colorectal cancer: current state of knowledge including clinical trials and molecular mechanism of action. BioMed Res Int. Article ID 4154185, pp 1–29. https://doi.org/10.1155/2018/4154185

Alibasic A, Simsekler MCE, Kurfess T, Woon W-L, Omar MA (2020) Utilizing data science techniques to analyze skill and demand changes in healthcare occupations: case study on USA and UAE healthcare sector. Soft Comput 24:4959–4976. https://doi.org/10.1007/s00500-019-04247-1

Amado A, Cortez P, Rita P, Moro S (2018) Research trends on big data in marketing: a text mining and topic modeling based

literature analysis. Eur Res Manag Bus Econ 24:1–7. https://doi.org/10.1016/j.iedeen.2017.06.002
Arab L, Liu W, Elashoff D (2009) Green and black tea consumption and risk of stroke: a meta-analysis. Stroke 40(5):1786–1792
Balentine DA, Wiseman SA, Bouwens LCM (1997) The chemistry of tea flavonoids. Crit Rev Food Sci Nutr 37:693–704. https://doi.org/10.1080/10408399709527797
Bastani K, Namavari H, Shaffer J (2019) Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst Appl 127:256–271. https://doi.org/10.1016/j.eswa.2019.03.001
Benn JA (2015) Tea in China: a religious and cultural history. University of Hawai'i Press, Honolulu
Blei DM (2012) Probabilistic topic models. Commun ACM 55:77–84. https://doi.org/10.1145/2133806.2133826
Blei DM, Lafferty JD (2005) Correlated topic models. In: Proceedings of the 18th international conference on neural information processing systems. MIT Press, Vancouver, pp 147–154
Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993
Cabrera C, Artacho R, Gimenez R (2006) Beneficial effects of green tea—a review. J Am Coll Nutr 25:79–99. https://doi.org/10.1080/07315724.2006.10719518
Cao H (2013) Polysaccharides from Chinese tea: recent advance on bioactivity and function. Int J Biol Macromol 62:76–79. https://doi.org/10.1016/j.ijbiomac.2013.08.033
Cheng K, Chi N-N, Liu J-D (2019) Green tea extract for treatment of cancers: a systematic review protocol. Medicine 98(15):e15117. https://doi.org/10.1097/MD.0000000000015117
Choi HS, Lee WS, Sohn SY (2017) Analyzing research trends in personal information privacy using topic modeling. Comput Secur 67:244–253. https://doi.org/10.1016/j.cose.2017.03.007
Chrystal P (2014) Tea: a very British beverage. Amberley Publishing, Stroud
da Silva Pinto M (2013) Tea: a new perspective on health benefits. Food Res Int 53: 558–567. https://doi.org/10.1016/j.foodres.2013.01.038
Deerwester S, Dumais ST, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inform Sci 41:391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3c391:AID-ASI1%3e3.0.CO;2-9
Delen D, Crossland MD (2008) Seeding the survey and analysis of research literature with text mining. Expert Syst Appl 34:1707–1720. https://doi.org/10.1016/j.eswa.2007.01.035
Fan W, Wallace L, Rich S, Zhang Z (2006) Tapping the power of text mining. Commun ACM 49:76–82. https://doi.org/10.1145/1151030.1151032
Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Usama MF, Gregory P-S, Padhraic S, Ramasamy U (eds) Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, pp 1–34
Feldman R, Regev Y, Hurvitz E, Finkelstein-Landau M (2003) Mining the biomedical literature using semantic analysis and natural language processing techniques. BIOSILICO 1:69–80. https://doi.org/10.1016/S1478-5382(03)02330-8
Graham HN (1992) Green tea composition, consumption, and polyphenol chemistry. Prev Med 21:334–350. https://doi.org/10.1016/0091-7435(92)90041-F
Greenberg JA, Axen KV, Schnoll R, Boozer CN (2005) Coffee, tea and diabetes: the role of weight loss and caffeine. Int J Obes 29(9):1121–1129. https://doi.org/10.1038/sj.ijo.0802999
Hao T, Chen X, Li G, Yan J (2018) A bibliometric analysis of text mining in medical research. Soft Comput 22:7875–7892. https://doi.org/10.1007/s00500-018-3511-4
Hobbs JR, Walker DE, Amsler RA (1982) Natural language access to structured text. In: Proceedings of the 9th conference on computational linguistics—volume 1. Academia Praha, Prague, Czechoslovakia, pp 127–132. https://doi.org/10.3115/991813.991833
Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, Berkeley, CA, pp 50–57. https://doi.org/10.1145/312624.312649
Hosoda K, Wang M-F, Liao M-L, Chuang C-K, Iha M, Clevidence B, Yamamoto S (2003) Antihyperglycemic effect of oolong tea in type 2 diabetes. Diabetes Care 26(6):1714–1718. https://doi.org/10.2337/diacare.26.6.1714
Jain A, Manghani C, Kohli S, Nigam D, Rani V (2013) Tea and human health: the dark shadows. Toxicol Lett 220:82–87. https://doi.org/10.1016/j.toxlet.2013.04.010
Jiao L, Bi L, Lu Y, Wang Q, Gong Y, Shi J, Xu L (2018) Cancer chemoprevention and therapy using Chinese herbal medicine. Biol Proced Online 20(1):1–14. https://doi.org/10.1186/s12575-017-0066-1
Jing L-P, Huang H-K, Shi H-B (2002) Improved feature selection approach TFIDF in text mining. In: Proceedings of international conference on machine learning and cybernetics, vol 942, pp 944–946. https://doi.org/10.1109/ICMLC.2002.1174522
Kajima S, Tanaka Y, Uchiyama Y (2017) Japanese sake and tea as place-based products: a comparison of regional certifications of globally important agricultural heritage systems, geopark, biosphere reserves, and geographical indication at product level certification. J Ethnic Foods 4:80–87. https://doi.org/10.1016/j.jef.2017.05.006
Kao Y-H, Chang H-H, Lee M-J, Chen C-L (2006) Tea, obesity, and diabetes. Mol Nutr Food Res 50:188–210. https://doi.org/10.1002/mnfr.200500109
Kiselev V-I, Ashrafyan L-A, Muyzhnek E-L, Gerfanova E-V, Antonova I-B, Aleshikova O-I, Sarkar F-H (2018) A new promising way of maintenance therapy in advanced ovarian cancer: a comparative clinical study. BMC Cancer 18, Article number: 904. https://doi.org/10.1186/s12885-018-4792-9
Koch KR (2007) Gibbs sampler by sampling-importance-resampling. J Geod 81:581–591. https://doi.org/10.1007/s00190-006-0121-1
Koga T, Meydani M (2001) Effect of plasma metabolites of (+)-catechin and quercetin on monocyte adhesion to human aortic endothelial cells. Am J Clin Nutr 73:941–948. https://doi.org/10.1093/ajcn/73.5.941
Kohsaka R, Matsuoka H (2015) Analysis of Japanese municipalities with Geopark, MAB, and GIAHS certification: quantitative approach to official records with text-mining methods. SAGE Open 5:1–10. https://doi.org/10.1177/2158244015617517
Mahmood T, Naveed A, Khan B (2010) The morphology, characteristics, and medicinal properties of Camellia sinensis' tea. J Med Plant Res 4(19):2028–2033. https://doi.org/10.5897/JMPR10.010
Mair VH, Hoh E (2009) The true history of tea. Thames and Hudson, New York
Marcos-Pablos S, García-Peñalvo FJ (2020) Information retrieval methodology for aiding scientific database search. Soft Comput 24:5551–5560. https://doi.org/10.1007/s00500-018-3568-0
Martin MA, Goya L, Ramos S (2017) Protective effects of tea, red wine and cocoa in diabetes. Evidences from human studies. Food Chem Toxicol 109:302–314. https://doi.org/10.1016/j.fct.2017.09.015
Mietzner D, Reger G (2005) Advantages and disadvantages of Scenario approaches for strategic foresight. Int J Technol Intell Plan 1:220–239. https://doi.org/10.1504/IJTIP.2005.006516

Moro S, Cortez P, Rita P (2015) Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst Appl 42:1314–1324. https://doi.org/10.1016/j.eswa.2014.09.024

Munday R (2016) Tea: health effects. In: Caballero B, Finglas PM, Toldrá F (eds) Encyclopedia of food and health. Academic Press, Oxford, pp 273–278. https://doi.org/10.1016/B978-0-12-384947-2.00686-3

Nechuta S, Shu X-O, Li H-L, Yang G, Ji B-T, Xiang Y-B, Cai H, Chow W-H, Gao Y-T, Zheng W (2012) Prospective cohort study of tea consumption and risk of digestive system cancers: Results from the Shanghai Women's Health Study. Am J Clin Nutr 96:1056–1063. https://doi.org/10.3945/ajcn.111.031419

Neyestani TR, Shariatzade N, Kalayi A, Gharavi A, Khalaji N, Dadkhah M, Zowghi T, Haidari H, Shab-bidar S (2010) Regular daily intake of black tea improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. Ann Nutr Metab 57:40–49. https://doi.org/10.1159/000312666

Pastoriza S, Mesías M, Cabrera C, Rufián-Henares JA (2017) Healthy properties of green and white teas: an update. Food Funct 8:2650–2662. https://doi.org/10.1039/C7FO00611J

Rady I, Mohameda H, Rady M, Siddiqui I-A, Mukhtara H (2018) Cancer preventive and therapeutic effects of EGCG, the major polyphenol in green tea. Egypt J Basic Applied Sci 5(1):1–23. https://doi.org/10.1016/j.ejbas.2017.12.001

Rashid J, Shah S-M-A, Irtaza A (2019) Fuzzy topic modeling approach for text mining over short text. Inf Process Manag 56(6):102060. https://doi.org/10.1016/j.ipm.2019.102060

Salton G, McGill MJ (1986) Introduction to modern information retrieval. McGraw-Hill, New York

Sangaiah AK, Tirkolaee EB, Goli A, Dehnavi-Arani S (2019a) Robust optimization and mixed-integer linear programming model for LNG supply chain planning problem. Soft Comput. https://doi.org/10.1007/s00500-019-04010-6

Sangaiah AK, Suraki MY, Sadeghilalimi M, Bozorgi SM, Hosseinabadi AAR, Wang J (2019b) A new meta-heuristic algorithm for solving the flexible dynamic job-shop problem with parallel machines. Symmetry 11(2):165. https://doi.org/10.3390/sym11020165

Sangaiah AK, Medhane DV, Han T, Hossain MS, Muhammad G (2019c) Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics. IEEE Trans Ind Inf 15(7):4189–4196. https://doi.org/10.1109/TII.2019.2898174

Shelton S, Badejo E (2018) Does green tea reduce the risk of breast cancer? Evid-Based Pract 21:48. https://doi.org/10.1097/01.EBP.0000545084.51626.b7

Shen L, Song L-G, Ma H, Jin C-N, Wang J-A, Xiang M-X (2012) Tea consumption and risk of stroke: a dose–response meta-analysis of prospective studies. J Zhejiang Univ Sci B 13:652–662. https://doi.org/10.1631/jzus.B1201001

Tian L, Huang J (2019) Antioxidant effects of tea catechins on the shelf life of raw minced duck meat. Food Sci Technol 39(1):59–65. https://doi.org/10.1590/fst.25217

van Dieren S, Uiterwaal CSPM, van der Schouw YT, van der A DL, Boer JMA, Spijkerman A, Grobbee DE, Beulens JWJ (2009) Coffee and tea consumption and risk of type 2 diabetes. Diabetologia 52:2561–2569. https://doi.org/10.1007/s00125-009-1516-3

Vo D-T, Ock C-Y (2015) Learning to classify short text from scientific documents using topic models with various types of knowledge. Expert Syst Appl 42:1684–1698. https://doi.org/10.1016/j.eswa.2014.09.031

Wei W, Guo C, Chen J, Tang L, Sun L (2019) CCODM: conditional co-occurrence degree matrix document representation method. Soft Comput 23:1239–1255. https://doi.org/10.1007/s00500-017-2844-8

Wolfram S (2007) Effects of green tea and EGCG on cardiovascular and metabolic health. J Am Coll Nutr 26:373S–388S. https://doi.org/10.1080/07315724.2007.10719626

Xie X, Ge S, Hu F, Xie M, Jiang N (2019) An improved algorithm for sentiment analysis based on maximum entropy. Soft Comput 23:599–611. https://doi.org/10.1007/s00500-017-2904-0

Xu B, Lin H, Lin Y, Guan Y (2020) Integrating social annotations into topic models for personalized document retrieval. Soft Comput 24:1707–1716. https://doi.org/10.1007/s00500-019-03998-1

Xuan J, Lu J, Zhang G (2019) Cooperative hierarchical Dirichlet processes: superposition vs. maximization. Artif Intell 271:43–73. https://doi.org/10.1016/j.artint.2018.10.005

Yao LH, Jiang YM, Shi J, Tomas-Barberan FA, Datta N, Singanusong R, Chen SS (2004) Flavonoids in food and their health benefits. Plant Foods Hum Nutr 59:113–122. https://doi.org/10.1007/s11130-004-0049-7

Zahedi E, Saraee M (2018) SSAM: Toward supervised sentiment and aspect modeling on different levels of labeling. Soft Comput 22:7989–8000. https://doi.org/10.1007/s00500-017-2746-9

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.