The Research Trends of Text Classification Studies (2000–2020): A Bibliometric Analysis

Haoran Zhu¹ and Lei Lei²

Abstract
Text Classification (TC) is the process of assigning several different categories to a set of texts. This study aims to evaluate the state of the arts of TC studies. Firstly, TC-related publications indexed in Web of Science were selected as data. In total, 3,121 TC-related publications were published in 760 journals between 2000 and 2020. Then, the bibliographic information was mined to identify the publication trends, important contributors, publication venues, and involved disciplines. Besides, a thematic analysis was performed to extract topics with increasing/decreasing popularity. The findings showed that TC has become a fast-growing interdisciplinary area, and that emerging research powers such as China are playing increasingly important roles in TC research. Moreover, the thematic analysis showed increased interest in topics concerning advanced classification algorithms, performance evaluation methods, and the practical applications of TC. This study will help researchers recognize the recent trends in the area.

Keywords
text classification, bibliometric analysis, research trends, topic extraction, dependency relations

Introduction
Text Classification (TC), also known as Document Classification or Text Categorization, is the process of assigning several predefined categories to a set of texts, often based on its content (Jindal et al., 2015; Wang & Deng, 2017). With the advent of the era of big data, the enormous quantity and diversity of digital documents have made it challenging for TC. As a result, TC has attracted much attention in various areas.

The working procedures of TC comprise text pre-processing, feature extraction/selection, training, prediction, and performance evaluation. Texts are usually pre-processed with tokenization, lemmatization, or stemming, in preparation for text representation (Kowsari et al., 2019). A classical model for text representation is Vector Space Model (VSM), with Bag-of-Words (BoW) as a popular sub-type (Santos et al., 2018). More recently, newly proposed models such as those based on word embedding (e.g., Khatua et al., 2019; Stein et al., 2019; Turner et al., 2017) and topic modeling (e.g., Pavlinek & Podgorelec, 2017; Potha & Stamatatos, 2019) have gained popularity in text representation. In addition, as texts are often represented via high-dimensional matrices, dimensional reduction is needed to address feature collinearity and to save computational cost (Shah & Patel, 2016).

Dimensional reduction is achieved with steps such as feature selection and feature extraction. The two steps, feature selection and feature extraction, though both aiming at reducing the number of features, are different in that feature extraction generates new variables while feature selection removes noises without creating new features (Seyyedi & Minaeibidgoli, 2018). The most common methods for feature selection include Term Frequency-Inverse Document Frequency (TF-IDF), Chi-square Statistics, Information Gain, and Mutual Information (Sababb et al., 2017; Shah & Patel, 2016). As for feature extraction, two approaches are popular, that is, Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI) (Shah & Patel, 2016). Both PCA and LSI transform a large number of features into a smaller set while preserving most of the differences, to boost the efficiency of classification. Features refined through selection and extraction are fed into classifiers for training and prediction. Traditionally, the most popular classifiers...
include Naive Bayes, K Nearest Neighbour, Decision Tree, Random Forest, and Support Vector Machine (Aggarwal et al., 2018). Lately, deep-learning-based classifiers have achieved impressive results in TC as they are able to model complex non-linear relationships within data (Kowsari et al., 2019). Evaluation is the final step in TC. The performance of TC techniques is often evaluated with various metrics, such as precision, accuracy, recall/sensitivity, and specificity (Kowsari et al., 2019).

Meanwhile, TC techniques have been applied in various contexts such as web page classification, authorship attribution, knowledge management, and spam email detection. For example, Qi and Davison (2009), Kiziloluk and Ozer (2017), and Meadi et al. (2017) explored algorithms and features in web page classification; Li et al. (2017) and Saleh et al. (2017) focused on the application of semantics-based approaches in web page classification. In the field of authorship attribution, in addition to traditional unsupervised methods such as Burrows’ delta (Burrows, 2002), an increasing number of studies have employed machine learning based classification techniques and reported promising results (Ebrahimpour et al., 2013; Jockers et al., 2008; Posadasduran et al., 2017; Tsimboukakis & Tambouratzis, 2010). TC techniques are also important methods in knowledge management, such as content-based recommendation (Hawashin et al., 2019; Wijewickrema et al., 2019; Wu et al., 2020), patent classification (Kim et al., 2020), and information extraction (Al-Yahya, 2018). Besides, TC techniques have been frequently applied to the detection of unwanted messages, including short message spam, junk mails, and suspicious malignant mails (Ezpeleta et al., 2017; Hsiao & Chang, 2008; Mujtaba, Shuib, Raj, & Gunalan, 2018; Seyyedi & Minaeibidgoli, 2018).

With the growing number of publications on TC, it is important for researchers to have a generalized understanding of research in this field. A number of review studies have already been carried out. For instance, Aggarwal et al. (2018) and Kowsari et al. (2019) presented a general overview of TC algorithms; Manikandan and Sivakumar (2018) and Kadhim (2019) conducted surveys on machine-learning-based techniques for TC; Altinel and Ganiz (2018) reviewed the history and development of semantic approaches to TC; Shah and Patel (2016) compared existing methods for feature selection and extraction. However, to our knowledge, no research has been conducted to systematically review TC research with large-scale bibliographic data from a bibliometric perspective. The bibliometric method is an effective tool to analyze, both quantitatively and qualitatively, the literature and research trend concerning a specific research area (Falagas et al., 2006). It helps assess the progress of a research area, identify the most relevant and influential source of publications, recognize major authors and institutions, and uncover potential research topics (Song et al., 2019). Many bibliometric studies have been conducted on topics related to natural language processing and data mining. A typical work in this line of research was a bibliometric review of computational linguistics from a general perspective (Radev, 2016). Other works included studies that investigated the landscape of specific research areas such as topic modeling (Li & Lei, 2019), big data (Raban & Gordon, 2020), digital library (Ahmad et al., 2018), machine learning (Elalfy & Mohammed 2020; Santos et al., 2019), Internet of Things (Erfanmanesh & Abrizah, 2018), decision making (Zyoud & Fuchs-Hanusch, 2017a), and environmental studies (Zhang et al., 2020; Zheng et al., 2017; Zyoud & Fuchs-Hanusch, 2017b, 2020; Zyoud & Zyoud, 2021).

Given that no such research has been performed on TC, the present study aims to provide a bibliometric analysis of TC-related publications in the past two decades. The rest of the paper is organized as follows. Section 2 describes the data source and methods of data analysis for the study. Section 3 reports on our results from the four perspectives: (1) annual trends in publications, (2) active contributors at country, institution, and author levels, (3) publication sources and disciplines, and (4) topics with increasing or decreasing popularity. We discuss in Section 4 the major findings and implications.

Data and Methods

Data

The data was collected from Clarivate Analytics Web of Science (WoS) Core Collection on September 20, 2021. The search statement we used is as follows:

\[
TS = (\text{“text classification” OR “text categorisation” OR “text categorization” OR “document classification” OR “document categorisation” OR “classification of text” OR “categorisation of text” OR “categorisation of text” OR “categorisation of document” OR “classification of document”}) \text{ AND PY = (2000–2020) AND DT=(Article)}
\]

Several points should be noted about the search statement. First, we not only used the term text classification, but also included its synonyms such as text categorization and document classification. We also included keywords of different phrasal structures and word orders (e.g., not only text classification but also classification of text(s)) for a more complete retrieval.

Second, we followed common practice in bibliometric research and restricted document types to articles. Only research articles were considered in the present study for two reasons. First, research articles provide original research findings and thus are of higher value in bibliometric analysis than other document types (Geng et al., 2017; Song et al., 2019). Second, most research articles include abstracts, which provides us the opportunity to analyze the trends of the research themes in TC across the examined years, while other document types such as book reviews often lack abstracts.
Third, we searched bibliographic data from three sub-databases of WoS: Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI). WoS was chosen as the data source because it is arguably one of the most famous and comprehensive databases of bibliographic information in the world (Song et al., 2019), and it has been widely utilized in many previous studies (Cansun & Arik, 2018; Raban & Gordon, 2020; Zhang et al., 2020; Zhu, 2021).

These three sub-databases were chosen because they are among the most widely used data source for bibliometric studies across many fields (Cansun & Arik, 2018; Donner, 2017; Li & Lei, 2019; Lopezrobles, 2019).

Last, the starting year was set as 2000 because our university library started to purchase the WoS database in that year. We acknowledge that a perhaps better approach is to include TC-related publications before 2000. However, we believe such a possible limitation may not have much undue effect on our analysis because: (1) the number of publications before 2000 was relatively small and (2) our major interest is in the recent development and the state of the art of TC related research. Thus, publications in the past 20 years should suffice.

To summarize, the aforementioned query obtained a total of 3,121 research articles published in 760 journals contributed by 7,186 different authors (from 2,292 institutions in 88 countries/regions). The raw bibliographic data of the articles were downloaded for the follow-up analyses.

Methods

Methods for descriptive results. The descriptive results regarding the annual publication trends, the analyses of major contributors at author/institution/country/region levels, and the analysis of publication venues were obtained from the WoS website. To be specific, after the search was performed, we clicked the “Analyze Results” button on the result page for a descriptive analysis of the retrieved bibliographic data (https://support.clarivate.com/ScientificandAcademicResearch). These results are to be presented in subsections 3.1 to 3.5. It should be noted that WoS adopts the complete counting principle in results analysis (Vavryčuk, 2018). Given that different counting principles (e.g., complete counting and partial counting) may lead to different results, the raw bibliographic data of this study has been provided in the Supplemental Appendix in case readers are interested in investigating them with alternative methods.

Following previous bibliometric studies, we used a series of bibliometric indicators to measure both research productivity and research impact. For research productivity, we used publication counts, that is, the number of publications of a given year, author, institution, country, etc. For research impact, we analyzed the citation counts, that is, the times an article is cited. In addition, we also used the $H$-index in the author and journal analyses. The $H$-index is the number of articles ($N$) in the examined list of publications that have $N$ or more citations. The $H$-index was included in the present study because it reflects both productivity and impact, and thus can be complementary to traditional metrics such as citation counts (Teixeira da Silva & Dobrânszki, 2018).

In addition, to capture the temporal change in research productivity (e.g., the annual trends of TC-related publication counts), we applied the Mann-Kendall trend test, a recommended method for non-parametric time series analysis (Kisi & Ay, 2014; Zhu, 2021).

Methods for thematic analysis. We performed a diachronic thematic analysis based on the retrieved abstracts in order to analyze the hot and cold topics in TC-related research. The steps of data processing are described as follows.

Firstly, we extracted all the noun phrases from abstracts of the articles downloaded with the python package spaCy (see https://github.com/explosion/spaCy for more technical details). The spaCy extracts noun phrases based on the analyses of syntactic dependency relations in a text, and hence achieves high accuracy in noun phrase extraction (Zhu & Lei, 2022). For example, in processing the following sentence, spaCy would first parse the dependency relations between the word tokens (Figure 1), and then extract all the noun phrases (NPs) based on the parsed dependency relations. Thus, the extracted noun phrases from the examplar sentence are: *term weighting, text documents, vector space, proper weights, and terms.*

Note that some of the downloaded articles do not have abstracts and hence were excluded at this step. Thus, a total of 3,115 abstracts were used for the extraction of hot and cold topics.

Secondly, we filtered the noun phrases based on their frequency and range. Here, the frequency refers to the total occurrence of a noun phrase in all the abstracts, and the range refers to the number of abstracts where a noun phrase occurs. Note that the thresholds for frequency and range might be arbitrary. After several rounds of trial, we found that the thresholds used in Lei et al. (2020) could also be applied in the present study. That is, candidate noun phrases should appear in at least 20 abstracts with a frequency of at least 30. A closer look at the extracted high-frequency noun phrases showed that some are noises such as *this paper, this study,* and *our findings.* They should not be considered as TC-related research topics and were removed. As a result, 108 candidate noun phrases were selected for the follow-up analyses.

Thirdly, for a balance of the number of abstracts in different periods, we divided the abstracts across the time span 2000 to 2019 into three periods for the temporal analysis, that is, 2000 to 2009, 2010 to 2015, and 2016 to 2020. Lastly, we calculated and compared the normalized frequency of
candidate noun phrases in each of the three periods with the following equation:

\[
\text{Normalized Frequency} = \frac{\text{Raw Frequency}}{\text{Number of Abstracts in that period}}
\]

We then conduct a one-way Chi-square test to identify the hot topics (topics with increasing normalized frequency) and cold topics (topics with decreasing normalized frequency). The identified hot/cold topics and the results of the one-way Chi-square tests would be presented in Section 3.7.

**Results**

In this section, we present the results of the bibliometric analysis and discuss the implications.

**Annual Trends of Publications**

Figure 2 shows the annual trend in the number of TC-related publications. It can be seen that the number of TC-related publications rose from 31 in 2000 to 463 in 2020 (Results of Mann-Kendall Trend Test: \(S=9.25, p<.01\)). The over ten-fold increase indicates that the research area has attracted more attention during the past two decades. A close look at the annual number of publications reveals an obvious uptrend either from 2000 to 2006 (Results of Mann-Kendall Trend Test: \(S=29.25, p<.01\)), or from 2007 to 2020 (Results of Mann-Kendall Trend Test: \(S=14.333, p<.01\)). However, a dramatic decrease is found from 2006 (with 188 publications) to 2007 (with only 73 publications). The decrease may be explained by the fact that two journals, that is, *Lecture Notes in Computer Science* and *Lecture Notes in Artificial Intelligence*, were removed from the WoS Core Collection from 2006. Since the two journals contributed a large number of TC-related publications from 2000 to 2006 (see Section 3.5), their removal may account for the decrease in the number of TC-related publications in 2007.

**Authors**

Table 1 lists the 30 authors with at least eight TC-related publications. We also provide in Table 1 other indicators to evaluate these authors’ productivity and impact, including total number of citations, number of citations per paper, and H-index. Note that these indicators are calculated with only TC-related publications of the authors, rather than all articles the authors published.

In terms of the number of TC-related publications, the three most productive authors are Fabrizio Sebastiani, Irene Diaz, and Masao Fuketa. Fabrizio Sebastiani has wide-ranging research interests, including “boosting” methods, human inspection in text classification, and multilingual text classification (Berardi et al., 2014; Fernandez et al., 2016). Irene Diaz worked mainly on feature selection in the early stage, and later shifted to classification methods for practical purposes, such as for precision agriculture (e.g., Arango, Campos, et al., 2016; Arango, Diaz, et al., 2016; Diaz et al., 2017) and medical use (e.g., Nunez et al., 2017). Masao Fuketa’s works are mostly published in collaboration with El Sayed Atlam. Their research primarily focuses on the extraction and filtering of “Field Association Terms,” that is, words that are specific to documents in the same field, and the application of Field Association Terms in text classification (Atlam et al., 2011; Dorji et al., 2011; Tanaka et al., 2009).

In terms of the number of citations per paper, Xijin Tang stands out as the most influential scholar of all the authors. With eight TC-related publications, Tang has received a total
of 530 citations, with an average of 66.25 citations per paper. Over half of these citations are from two papers on text representation that Tang co-authored with Wen Zhang (Zhang et al., 2008, 2011).

From the perspective of the $H$-index, Fabrizio Sebastiani, Dino Isa, and Lam Hong Lee are the three most influential scholars. All of the three scholars have an $H$-index of 9. Besides Fabrizio Sebastiani, who has published the largest number of TC-related articles as aforementioned, Dino Isa and Lam Hong Lee have a considerable overlap of research interests. Both Dino Isa and Lam Hong have devoted much of their work to the application and enhancement of classifiers such as SVM.

It should be pointed out that not all influential researchers in the field are included in Table 1, since some researchers publish relatively fewer in number but higher quality articles. For example, Zhihua Zhou from Peking University has published only seven TC-related publications, but has received 1,893 citations, with each article cited 270 times. It should also be noted that the ranking in Table 1 only reflects the efforts that the researchers devoted to the field of TC, hence it is not necessarily a reflection of the recognition researchers have earned in academia at large.

**Institutions**

The 24 most productive institutions in TC-related research are listed in Table 2. An interesting observation (see Table 2) is that Asian institutions contribute prominently to TC-related research. To be specific, 17 of the 24 most productive institutions are based in Asia, and the top 4 institutions are all Asian ones. In particular, Chinese universities research institutions have occupied important positions in the list of most productive institutions.

**Countries/Regions**

In total, TC-related publications originated from 88 countries/regions. The most productive countries/regions with more than 100 TC-related publications are listed in Table 3. The mainland part of China is the most productive country/region with 854 TC-related publications, followed by the USA which has 613 TC-related publications. The two countries combined accounted for over 45% of all TC-related publications. Other productive countries/regions with over 100 TC-related publications include South Korea (165), Australia (133), Canada (132), India (128), Spain (125), Japan (123), Taiwan (122), Germany (110), and Italy (110). In addition, we plotted (in Figure 3) the temporal trends in the numbers of TC-related publications of the six most productive countries/regions. As is illustrated in Figure 3, The mainland part of China accounted for the most of the increase of TC-related publications in the past two decades.

**Journals**

A total of 760 journals published TC-related publications from 2000 to 2020. We list in Table 4 all the journals with
more than 20 TC-related publications. The three largest publishing outlets in the past 20 years are Lecture Notes in Computer Science (266), Expert Systems with Applications (150), and Lecture Notes in Artificial Intelligence (144). Note that Lecture Notes in Computer Science and Lecture Notes in Artificial Intelligence have been excluded from WoS Core Collection since 2007. Therefore, Expert Systems with Applications is now the largest publishing source for TC-related publications of all journals currently indexed by WoS Core Collection.

In addition to the number of TC-related publications, the quality of TC-related publications in these publishing sources is also worth exploring. To achieve this, we investigate the $H$-index and of each journal using all TC-related publications it has published from 2000 to 2020. The ranking of the journals according to $H$-index of TC-related publications is illustrated in Table 5. The quantile information (based on WoS Journal Citation Report 2020) of these journals is also provided in Table 5. As measured by the $H$-index, the three most influential journals in the research field of text classification are Expert Systems with Applications ($H$-index: 40), IEEE Transactions on Knowledge and Data Engineering ($H$-index: 31), and Information Processing & Management ($H$-index: 30).

### Subject Categories

TC-related publications spread across 137 WoS subject categories. Table 6 shows the subject categories with more than 40 TC-related publications. It can be seen from Table 6 that the majority of TC-related articles are published in the field of Computer Sciences. Meanwhile, TC has drawn the attention from many other disciplines, including Engineering, Library Science, Management Science, Language and Linguistics, and Biotechnology.

### Thematic Changes

In this section, we report on the identified hot and cold topics in TC-related research. The hot and cold topics as well as

---

**Table 1.** Authors with at Least Eight TC-Related Publications.

| Author                        | Affiliation                                      | Number of publications | $H$-index | Total citations | Citations per paper |
|-------------------------------|--------------------------------------------------|------------------------|-----------|-----------------|---------------------|
| Sebastiani, Fabrizio          | National Research Council                        | 18                     | 9         | 348             | 19.33               |
| Fuketa, Masao                 | Tokushima University                             | 13                     | 6         | 119             | 9.15                |
| Diaz, Irene                   | University of Oviedo                             | 13                     | 5         | 107             | 8.23                |
| Esuli, Andrea                 | National Research Council                        | 12                     | 7         | 114             | 9.5                 |
| Montanes, Elena               | University of Oviedo                             | 12                     | 5         | 138             | 11.5                |
| Atlam, El Sayed               | Tanta University                                 | 12                     | 4         | 59              | 4.92                |
| Uysal, Alper Kursat           | Eskisehir Technical University                   | 11                     | 7         | 566             | 51.45               |
| Morita, Kazuhiro              | Tokushima University                             | 11                     | 5         | 70              | 6.36                |
| Ranilla, Jose                 | University of Oviedo                             | 11                     | 4         | 89              | 8.09                |
| Bi, Yaxin                     | Ulster University                                | 10                     | 7         | 180             | 18                  |
| Montes-y-Gomez, Manuel        | National Institute of Astrophysics, Optics and Electronics (INAOE) | 10                     | 7         | 123             | 12.3               |
| Wu, Jia                       | Macquarie University                             | 10                     | 6         | 184             | 18.4                |
| Bouguila, Nizar               | Concordia University                              | 10                     | 6         | 197             | 19.7                |
| Combarro, Elias F.            | University of Oviedo                             | 10                     | 4         | 99              | 9.9                 |
| Isa, Dino                     | University of Nottingham                         | 9                      | 9         | 376             | 41.78               |
| Lee, Lam Hong                 | Quest International University Perak             | 9                      | 9         | 376             | 41.78               |
| Wei, Chih-Ping                | National Taiwan University                       | 9                      | 7         | 160             | 17.78               |
| Ganiz, Murat Can              | Marmara University                               | 9                      | 7         | 188             | 20.89               |
| Altincay, Hakan               | Eastern Mediterranean University                 | 9                      | 5         | 101             | 11.22               |
| Zuo, Wanli                    | Southeast University                             | 9                      | 4         | 81              | 9                   |
| Theeramunkong, Thanaruk        | Thammast University                              | 9                      | 4         | 82              | 9.11                |
| Chen, Hsinchun                | University of Arizona                            | 8                      | 8         | 256             | 32                  |
| Jiang, Liangxiao              | China University of Geosciences                  | 8                      | 7         | 228             | 28.5                |
| Tang Xijin                    | Chinese Academy of Sciences                      | 8                      | 7         | 530             | 66.25               |
| Li, Chenghua                  | Chonbuk National University                      | 8                      | 6         | 130             | 16.25               |
| Li, Tao                       | Florida International University                 | 8                      | 6         | 269             | 33.63               |
| Iglesias, Eva L.              | University of Vigo                               | 8                      | 5         | 103             | 12.88               |
| Park, Jonghun                 | Seoul National University                        | 8                      | 5         | 137             | 17.13               |
| Wang, Youwei                  | Shandong Agricultural University                 | 8                      | 4         | 57              | 7.13                |
| Aoe, Jun-ichi                 | Tokushima University                             | 8                      | 3         | 67              | 8.38                |
the results of the one-way Chi-square tests are presented in Table 7.

Hot topics. Hot topics in TC-related research can be grouped into five categories.

Feature-related topics. For example, feature and feature extraction experienced significant increases, while feature selection and feature vector show noticeable, though not significant increases. This shows the growing importance of pre-processing features before they are applied to classification.

High-dimensionality of feature space has been a long-standing challenge in text classification; therefore, researchers have been working on eliminating noises in features to preserve the most informative features (Liu et al., 2014; Seyyedi & Minaeibidgoli, 2017). The results presented here indicate that techniques for feature selection/extraction have gained, and will probably continue to draw attention in future TC-related research.

Algorithm-related topics. Many topics in this category pertain to methods of text representation, for example,
Figure 3. Trends in the numbers of TC-related publications of the six most productive countries/regions.

Table 4. Journals With More Than 20 TC-Related Publications.

| Source titles                                          | Number of publications | Percentage of 3,121 |
|--------------------------------------------------------|------------------------|---------------------|
| Lecture notes in computer science                      | 266                    | 8.523               |
| Expert systems with applications                       | 150                    | 4.806               |
| Lecture notes in artificial intelligence               | 144                    | 4.614               |
| IEEE access                                            | 115                    | 3.685               |
| Information processing & management                    | 95                     | 3.044               |
| Neurocomputing                                         | 66                     | 2.115               |
| Knowledge-based systems                                | 58                     | 1.858               |
| IEEE transactions on knowledge and data engineering    | 56                     | 1.794               |
| Journal of biomedical informatics                      | 54                     | 1.730               |
| Information sciences                                   | 41                     | 1.314               |
| Journal of machine learning research                   | 40                     | 1.282               |
| Pattern recognition letters                            | 40                     | 1.282               |
| Journal of the American Society for information science and technology | 37                     | 1.186               |
| Neural computing applications                          | 35                     | 1.121               |
| Knowledge and information systems                      | 34                     | 1.089               |
| Pattern recognition                                    | 34                     | 1.089               |
| Journal of information science                         | 31                     | 0.993               |
| Multimedia tools and applications                      | 30                     | 0.961               |
| Applied sciences basel                                 | 28                     | 0.897               |
| Applied soft computing                                 | 28                     | 0.897               |
| Applied intelligence                                  | 27                     | 0.865               |
| Journal of intelligent fuzzy systems                   | 27                     | 0.865               |
| Machine learning                                       | 27                     | 0.865               |
| Bmc bioinformatics                                     | 26                     | 0.833               |
| Information retrieval                                  | 25                     | 0.801               |
| Journal of intelligent information systems             | 25                     | 0.801               |
| Journal of the American medical informatics association | 25                     | 0.801               |
| IEEE transactions on pattern analysis and machine intelligence | 22                     | 0.705               |
| IEICE transactions on information and systems          | 22                     | 0.705               |
| Intelligent data analysis                              | 21                     | 0.673               |
model, word embedding, Convolutional Neural Networks (CNN), and Latent Dirichlet Allocation (LDA). These algorithms are recently developed techniques for the representation of texts, and are claimed to overcome the limitations of traditional representation methods such as BoW with information mined at deeper levels. For example, Al Moubayed et al. (2017) argued that topic modeling approaches reflect contextual information, Stein et al. (2019) and Wang et al. (2019) found that word embedding models could reveal semantic information that boosts the accuracy of classification tasks, and Yao (2019) and Cheng (2019) showed that CNN model may facilitate learning abstract relations...
and hidden features. Other terms in this category are either widely accepted methods for text classification, such as machine learning, or promising approaches based on neural networks, such as Extreme Learning Machine (ELM).

**Evaluation-related topics.** Terms in this category include performance evaluation measures such as F-measure/F-score, accuracy, efficiency, sensitivity, and recall. These terms represent different dimensions in evaluating the classification results. The fact that all these terms have seen a noticeable rise in frequency demonstrates the increasing importance of performance evaluation in TC-related research. Accuracy and recall (also termed sensitivity) were widely used and extensively discussed in earlier studies (e.g., Azzini & Ceravolo, 2006; Sordo & Zeng, 2005; Zahedi & Sorkhi, 2013), and are thus considered classical indicators to measure classification performance. F-measure or F-score, though underused earlier, have received more attention in recent years as their normalized frequency has more than tripled from periods 1 to 3. There are still other evaluation metrics in use such as Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC). However, AUC and ROC are filtered out by low normalized frequency and are not included in Table 7, perhaps due to their potential limitations when applied as classification evaluation measures (Muschelli, 2019; Wald & Bestwick, 2014).

**Application-related topics.** Topics in this category suggest that TC can be applied to various fields, ranging from sentiment analysis, machine translation, to authorship attribution,
since these are in essence classification tasks. For example, the majority of sentiment analysis studies focus on either sentiment polarity, which determines whether a text is positive or negative, or sentiment subjectivity classification, which defines whether a text is subjective or objective (Ortigosa-Hernandez et al., 2012). In a similar vein, researchers in the field of literary stylistics, who noticed the need of combining quantitative means with traditional qualitative analysis, have introduced a series of TC algorithms to style-based authorship attribution (Koppel et al., 2009; Stammatatos, 2009).

Another important task of natural language engineering (NLE), that is, *machine translation*, is closely linked with cross-lingual text classification (Garcia et al., 2017), and thus has appeared more frequently in TC-related research in recent years. It is interesting to note that TC techniques have also gained popularity in medicine and clinical diagnosis, as manifested in the increased use of *patient* and *disease* across the three periods. A closer look shows that TC techniques are often employed by medical researchers to increase the accuracy of diagnosis (Krebs et al., 2019; Sullivan et al., 2014), improve clinical treatment and care (Liu & Wang, 2018; Nii et al., 2012), and analyze the feedback of patients (Liu & Chen, 2019).

**Other topics (interdisciplinary).** Some terms in this category pertain to the data sources of TC-related research, such as *social media, twitter/tweet, and wikipedia*. Studies that use such type of data are often claimed to be based on “big data.” Some other terms, such as *corpus/corpora* and *semantics*, have also been more frequently used, which reveals the interplay between TC and linguistics. In general, these terms demonstrate the interaction of TC with other areas such as data science, information technology, and linguistics. Such a finding provides further evidence to the findings in Section 3.6 that TC has received much attention from various disciplines.

**Cold topics.** Cold topics, that is, terms that have experienced a decrease in normalized frequency are relatively small in number, and can be roughly divided into three types. The first type are three classifiers: *KNN, Rocchio, and Naïve Bayes*. These classifiers share the feature of being straightforward and easy-to-understand, but are limited in low classification accuracy, high sensitivity to noise, and inability in deeper semantics identification. Hence, although these classifiers are frequently used in earlier studies, they have been outperformed by recently proposed approaches such as neural network and word embedding (e.g., Amanpreet, 2019; Bani-Hani & Khasawneh, 2019; Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018). The second type are terms related to the internet, such as *web, web page, and web document*. It should be noted that, as previously mentioned, some other internet-related terms such as *social media, twitter/tweet, and wikipedia* have been significantly more frequently used in the past decades and hence identified as hot topics. When the internet-related hot and cold topics are compared, it is obvious that the cold topics are more general terms, while the hot ones are often the description of more specific websites. Internet-related nouns usually appear in TC-related publications as data source (e.g., Cheng & Chen, 2019; Hafez, 2017; Kazemian & Ahmed, 2015; Wang et al., 2017). This might indicate that, with the emergence and thriving of fine-grained, vertical websites, TC-related research now tends to employ data from certain types of websites that can better serve their specific research purposes, rather than from the cyberspace indiscriminately. Besides, the findings show that topics related to hierarchical classification (*hierarchical classification*) or training (*training set*) are also decreasing in frequency, which shows that these topics have received less discussion in TC-related publications.

**Discussion and Implications**

Based on 3,121 publications collected from WoS, the present study attempts to present a comprehensive overview of the research landscape and the latest development of TC research. It is, to the best of our knowledge, the first study that presents a comprehensive review of TC research using bibliometric methods. In particular, we adopted a novel method for thematic analyses, making use of dependency-based topic extraction and trend analysis. Our study revealed four points of interest in TC research.

The first point is that emerging research powers are playing an increasingly visible role in the field of TC research. Results of the institution analysis show that 10 of the 25 most productive institutions in TC research are based in the mainland part of China. Results of the country analysis show that, in terms of publication numbers, the mainland part of China is the most productive country/region and has contributed to more than a quarter of TC publications since 2000. More importantly, from a diachronic perspective, China has made remarkable progress in TC research, especially since 2015. As discussed earlier, most of the increment in TC publications in the past 5 years originates from China. Apart from China, other developing countries such as India also have made important contributions both to the number of publications and to the increment in TC publications since 2000. The rise of China and other emerging research powers in TC research may be explained from two perspectives. First, along with economic development, these countries/regions have given more impetus to academic output with increased investment in scientific research (Lei & Liao, 2017). Second, the researchers in these countries/regions are motivated to publish more research articles in high-quality journals, in order to win recognition in international academia, and to cope with the publish or perish pressure (Lee, 2014). These factors would result in more international publications from these emerging research powers in both natural and social sciences, including TC research.

The second point is the interdisciplinary nature of TC research. Our findings show that the majority of TC-related
publications are in the field of Computer Science. Meanwhile, TC techniques have also gained popularity in Engineering and Social Sciences such as Library Science and Management Science. In particular, many TC-related publications are contributed by researchers from the fields of Biochemistry and Biotechnology, which traditionally seem not directly or closely relevant to TC techniques. Such a phenomenon may be explained by the wide and increasing use of natural language processing (NLP) techniques in biological sequences processing (Badal et al., 2018; Buchan & Jones, 2020; Islam et al., 2018; Le et al., 2019). To be specific, many biological sequences that play fundamental roles in life, such as Deoxyribonucleic Acid (DNA) chains and Protein sequences are formed by small molecules with intricate structures and complex grammars, similar to how texts are formed by words or n-grams (Huang & Yu, 2016; Islam et al., 2018; Srivastava & Baptista, 2016). Hence, it has been proposed that biologists may employ techniques in NLP or computational linguistics for the analyses of biological sequences (Gimona, 2006).

It is also worth noting that the Language Linguistics category has contributed a considerable number of TC publications in recent years. This indicates that techniques and algorithms from TC research may find wide applications in humanities and social sciences. For example, researchers in literary science and linguistics have employed TC as an effective tool in facilitating research such as the attribution of authorship and the interpretation of literary styles (Zhu et al., 2020). As a result of such an interaction between different fields, TC has evolved into an interdisciplinary research area.

The third point of interest is that TC has attracted attention from industries. Our analysis of the most productive institutions shows that the presence of commercial enterprises is noticeable in the list of most productive author institutions. While the majority of these productive institutions are universities and research institutes, two business companies, International Business Machines (IBM) and Microsoft are on the list. Both IBM and Microsoft are leading enterprises with global impact in the information technology industry, and both have research branches that are dedicated to practical problems, industrial challenges, and technical innovations. The identification of the two enterprises on the list indicates that TC-related research outputs may have wide applications in industries.

Finally, the results of thematic change analysis reveal important development patterns in TC research, which may help facilitate our understanding of the trends and the state of the arts in the field. It is shown that, during the past two decades, many topics have received increasing attention, especially those related to classification features, new algorithms, performance evaluation methods, and the practical applications of TC in other disciplines. Noticeably, state-of-the-art models and methods (topic model, word embedding, and CNN) have been introduced to the domain of TC and witnessed increasing popularity. It may be predicted that research on these topics will continue to draw attention in the recent future. In contrast, a few topics have experienced decreased interest, including some traditional algorithms which used to be in wide use in TC studies (e.g., k-nearest neighbor, roc- chio, and naive Bayes). These observed changing patterns of TC research may provide researchers with useful implications in topic choices, research design, algorithm optimization, and the interpretation of research findings.

We acknowledge that the study has some limitations which should be addressed in future research. Firstly, the data employed in the present study is limited to WoS Core Collection. Future studies may consider including data from other databases such as Scopus to further verify the findings of the present study. Secondly, although the search statement we used was able to retrieve a considerable number of TC-related publications, it may be difficult to ensure that the search results are exhaustive. Some publications relevant to TC research may have not been retrieved. Lastly, our method for topic extraction included human judgment. Although the researchers closely double-checked the results, subjectivity is impossible to be avoided. Thus, future research may consider modifying the algorithm for topic identification to address the issue of subjectivity.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by an MOE (Ministry of Education of China) Foundation Project of Humanities and Social Sciences: Linguistic Complexity-based Research on Text Classification (Grant No. 21YJC740085).

ORCID iD
Lei Lei https://orcid.org/0000-0002-3366-1855

Supplemental Material
Supplemental material for this article is available online.

References
Aggarwal, A., Singh, J., & Gupta, K. (2018). A Review of different text categorization techniques. International Journal of Engineering and Technology, 7, 11–15.
Ahmad, K., Jian Ming, Z., & Rafi, M. (2018). Assessing the digital library research output: Bibliometric analysis from 2002 to 2016. The Electronic Library, 36(4), 696–704.
Al-Yahya, M. (2018). Stylometric analysis of classical Arabic texts for genre detection. The Electronic Library, 36(5), 842–855.
Altinel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6), 1129–1153. https://doi.org/10.1016/j.ipm.2018.08.001
Amanpreet, K. (2019). Machine learning-based novel approach to classify the shoulder motion of upper limb amputees. Biocybernetics and Biomedical Engineering, 39(3), 857–867.
Arango, R. B., Campos, A. M., Combarro, E. F., Canas, E. R., & Diaz, I. (2016). Mapping cultivable land from satellite imagery with clustering algorithms. *International Journal of Applied Earth Observation and Geoinformation, 59*, 99–106.

Arango, R. B., Diaz, I., Campos, A. M., Canas, E. R., & Combarro, E. F. (2016). Automatic arable land detection with supervised machine learning. *Earth Science Informatics, 9*(4), 535–545.

Atlam, E., Morita, K., Fuketa, M., & Aoe, J. (2011). A new approach for Arabic text classification using Arabic field-association terms. *Journal of the Association for Information Science and Technology, 62*(11), 2266–2276.

Azzini, A., & Ceravolo, P. (2006) Evolutionary ANNs for improving accuracy and efficiency in document classification methods. In B. Gabrys, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-based intelligent information and engineering systems. KES 2006. Lecture Notes in Computer Science* (Vol. 4253, pp. 1111–1118). Springer.

Badal, V. D., Kundrotas, P. J., & Vakser, I. A. (2018). Natural language processing in text mining for structural modeling of protein complexes. *BMC Bioinformatics, 19*(1), 1–10.

Bani-Hani, D., & Khasawneh, M. T. (2019). A recursive general regression neural network (R-GRNN) oracle for classification problems. *Expert Systems with Applications, 135*, 273–286.

Berardi, G., Esuli, A., & Sebastiani, F. (2014). Optimising human inspection work in automated verbatim coding. *International Journal of Market Research, 56*(4), 489–512.

Buchan, D. W., & Jones, D. T. (2020). Learning a functional grammar of protein domains using natural language word embedding techniques. *Proteins, 88*(4), 616–624.

Burrows, J. (2002). ‘Delta’: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287.

Cansun, S., & Arik, E. (2018). Political science publications about Turkey. *Scientometrics, 115*(1), 169–188.

Cheng, C., & Chen, H. (2019). Sentimental text mining based on an additional features method for text classification. *PLoS One, 14*(6), e0217591.

Cheng, Y., Ye, Z., Wang, M., & Zhang, Q. (2019). Document classification based on convolutional neural network and hierarchical attention network. *Neural Network World, 29*(2), 83–98.

Diaz, I., Mazza, S. M., Combarro, E. F., Gimenez, L. I., & Gaiad, J. E. (2017). Machine learning applied to the prediction of citrus production. *Spanish Journal of Agricultural Research, 15*(2), e0205.

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics, 113*(1), 219–236.

Dorji, T. C., Atlam, E., Yata, S., Fuketa, M., Morita, K., & Aoe, J. (2011). Extraction, selection and ranking of field association (FA) terms from domain-specific corpora for building a comprehensive FA terms dictionary. *Knowledge and Information Systems, 27*(1), 141–161.

Ebrahimpour, M., Putnins, T. J., Berryman, M. J., Allison, A., Ng, B. W., & Abbott, D. (2013). Automated authorship attribution using advanced signal classification techniques. *PLoS One, 8*(2), e54998.

Elalfy, E. M., & Mohammed, S. (2020). A review of machine learning for big data analytics: Bibliometric approach. *Technology Analysis and Strategic Management, 32*(7), 1–22. https://doi.org/10.1080/09537325.2020.1732912

Erfanmanesh, M., & Abrizah, A. (2018). Mapping worldwide research on the Internet of Things during 2011–2016. *The Electronic Library, 36*(6), 979–992.

Ezpeleta, E., Garitano, I., Zurutuza, U., & Hidalgo, J. M. (2017). Short messages spam filtering combining personality recognition and sentiment analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 25*, 175–189.

Falagas, M. E., Karavasilou, A. I., & Bliziotis, I. A. (2006). A bibliometric analysis of global trends of research productivity in tropical medicine. *Acta Tropica, 99*(2), 155–159.

Fernandez, A., Esuli, A., & Sebastiani, F. (2016). Lightweight random indexing for polylingual text classification. *Journal of Artificial Intelligence Research, 57*, 151–185.

Garcia, M. A., Rodriguez, R. P., & Rifin, L. A. (2017). Wikipedia-based cross-language text classification. *Information Sciences, 406–407*, 12–28. https://doi.org/10.1016/j.ins.2017.04.024

Geng, Y., Chen, W., Liu, Z., Chiu, A. S., Han, W., Liu, Z., & Cui, X. (2017). A bibliometric review: Energy consumption and greenhouse gas emissions in the residential sector. *Journal of Cleaner Production, 159*, 301–316.

Gimona, M. (2006). Protein linguistics: A grammar for modular protein assembly? *Nature Reviews Molecular Cell Biology, 7*(1), 68–73.

Hathliyan, N. F., & Hafez, A. M. (2017). Subjective text mining for Arabic social media. *International Journal on Semantic Web and Information Systems, 13*(2), 1–13.

Hawashin, B., Alzubi, S., Kanan, T., & Mansour, A. (2019). An efficient semantic recommender method for Arabic text. *The Electronic Library, 37*(2), 263–280. https://doi.org/10.1108/EL-12-2018-0245

Hsiao, W., & Chang, T. (2008). An incremental cluster-based approach to spam filtering. *Expert Systems with Applications, 34*(3), 1599–1608.

Huang, H., & Yu, C. (2016). Clustering DNA sequences using the out-of-place measure with reduced n-grams. *Journal of Theoretical Biology, 406*, 61–72.

Islam, S. M., Heil, B. J., Kearney, C. M., & Baker, E. J. (2018). Protein classification using modified n-grams and skip-grams. *Bioinformatics, 34*(9), 1481–1487.

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: Literature review and current trends. *Webology, 12*(2), 1–28.

Jockers, M. L., Witten, D., & Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunk centroid classification. *Literary and Linguistic Computing, 23*(4), 465–491.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review, 52*, 273–292. https://doi.org/10.1007/s10462-018-09677-1

Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. *Expert Systems with Applications, 42*(3), 1166–1177.

Khatua, A., Khatua, A., & Cambria, E. (2019). A tale of two epistemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing and Management, 55*(1), 247–257.

Kim, J., Yoon, J., Park, E., & Choi, S. (2020). Patent document clustering with deep embeddings. *Scientometrics, 123*(2), 563–577.
term weighting schemes for text classification. Applied Soft Computing, 58, 193–206.

Saleh, A. I., Rahmawy, M. F., & Abulwafa, A. E. (2017). A semantic based web page classification strategy using multi-layered domain ontology. World Wide Web, 20(5), 939–993.

Santos, B. S., Steiner, M. T., Fenerich, A. T., & Lima, R. H. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. Computers & Industrial Engineering, 138, 106120. https://doi.org/10.1016/j.cie.2019.106120

Santos, F. T., Domingues, M. A., Sundermann, C. V., De Carvalho, V. O., Moura, M. F., & Rezende, S. O. (2018). Latent association rule cluster based model to extract topics for classification and recommendation applications. Expert Systems with Applications, 112, 34–60.

Seyyedi, S. H., & Minaeibidgoli, B. (2017). Enhancing effectiveness of dimension reduction in text classification. International Journal on Artificial Intelligence Tools, 26, 1750008:1–1750008:21. https://doi.org/10.1142/S0218213017500087

Seyyedi, S. H., & Minaeibidgoli, B. (2018). Estimator learning automata for feature subset selection in high-dimensional spaces, case study: Email spam detection. International Journal of Communication Systems, 31(7), e3541.

Shah, F. P., & Patel, V. (2016). A review on feature selection and feature extraction for text classification [Conference session]. International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, pp. 2264–2268. https://doi.org/10.1109/WiSPNET.2016.7566545

Song, Y., Chen, X., Hao, T., Liu, Z., & Lan, Z. (2019). Exploring two decades of research on classroom dialogue by using bibliometric analysis. Computers & Education, 137, 12–31.

Sordo, M., & Zeng, Q. (2005). On sample size and classification accuracy: A performance comparison. In J. L. Oliveira, V. Maojo, F. Martín-Sánchez, & A. S. Pereira (Eds.), Biological and medical data analysis (pp. 193–201). Springer.

Srivastava, S., & Baptista, M. S. (2016). Markovian language model of the DNA and its information content. Royal Society Open Science, 3(1), 2052–2054.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the Association for Information Science and Technology, 60(3), 538–556.

Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. Information Sciences, 471, 216–232.

Sulieman, L. M., Gilmore, D., French, C., Cronin, R. M., Jackson, G. P., Russell, M., & Fabbri, D. (2017). Classifying patient portal messages using convolutional neural networks. Journal of Biomedical Informatics, 74, 59–70.

Sullivan, R., Yao, R., Jarrar, R., Buchhalter, J., & Gonzalez, G. (2014). Text classification towards detecting misdiagnosis of an epilepsy syndrome in a pediatric population. Proceedings of the AMIA Annual Symposium, pp. 1082–1087, Washington, DC, USA.

Szymanski, J., & Kawalec, N. (2019). An analysis of neural word representations for Wikipedia articles classification. Cybernetics and Systems, 50(2), 176–196.

Tanaka, A., Atlam, E., Morita, K., Tsukuda, Y., Fuketa, M., & Aoe, J. (2009). Relevant estimation among fields using field association words. Journal of Computer Applications in Technology, 35(2), 296–306.

Teixeira da Silva, J. A., & Dobrászski, J. (2018). Multiple versions of the h-index: Cautionary use for formal academic purposes. Scientometrics, 115(2), 1107–1113. https://doi.org/10.1007/s11192-018-2680-3

Tsimboukas, N., & Tambouratzis, G. (2010). A comparative study on authorship attribution classification tasks using both neural network and statistical methods. Neural Computing and Applications, 19(4), 573–582.

Turner, C. A., Jacobs, A., Marques, C. K., Oates, J. C., Kamen, D. L., Anderson, P. E., & Obeid, J. S. (2017). Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Medical Informatics and Decision Making, 17(1), 1–11.

Vavryčuk, V. (2018). Fair ranking of researchers and research teams. PLoS One, 13(4), e0195509. https://doi.org/10.1371/journal.pone.0195509

Wald, N. J., & Bestwick, J. P. (2014). Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test. Journal of Medical Screening, 21(1), 51–56.

Wang, H., & Deng, S. (2017). A paper-text perspective: Studies on the influence of feature granularity for Chinese short-text-classification in the Big Data era. The Electronic Library, 35(4), 689–708.

Wang, Q., Ma, S., & Zhang, C. (2017). Predicting users’ demographic characteristics in a Chinese social media network. The Electronic Library, 35(4), 758–769.

Wang, Y., Zeng, Y., Tang, J., & Xu, B. (2019). Biological neuron coding inspired binary word embeddings. Cognitive Computation, 11(5), 676–684.

Wijewickrema, M., Petras, V., & Dias, N. (2019). Selecting a text similarity measure for a content-based recommender system: A comparison in two corpora. The Electronic Library, 37(3), 506–527. https://doi.org/10.1080/EL-08-2018-0165

Wu, D., Wu, L., Palmer, A., Kinshuk, D., & Zhou, P. (2020). Automatic evaluation of online learning interaction content using domain concepts. The Electronic Library, 38(3), 421–445. https://doi.org/10.1108/EL-09-2019-0223

Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Medical Informatics and Decision Making, 19(Suppl 1), 71. https://doi.org/10.1186/s12911-019-0781-4

Zahedi, M., & Sorkhi, A. G. (2013). Improving text classification using domain concepts. The Electronic Library, 38(3), 689–708.

Zahedi, M., Sorkhi, A. G. (2010). Improving text classification in the Big Data era. Knowledge Based Systems, 21(8), 879–886.

Zahedi, M., Sorkhi, A. G., & Zhang, X. (2008). Text classification based on multi-word with support vector machine. Knowledge Based Systems, 21(8), 879–886.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF-IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758–2765.

Zhang, Y., Pu, S., Lv, X., Gao, Y., & Ge, L. (2020). Global trends and prospects in microplastics research: A bibliometric analysis. Journal of Hazardous Materials, 400, 123110. https://doi.org/10.1016/j.jhazmat.2020.123110

Zheng, T., Li, P., Shi, Z., & Liu, J. (2017). Benchmarking the scientific research on wastewater-energy nexus by using bibliometric analysis. Scientific and statistically-based web page classification strategy using multi-layered domain ontology. World Wide Web, 20(5), 939–993.
Zhu, H. (2021). Home country bias in academic publishing: A case study of the New England journal of medicine. *Learned Publishing, 34*(4), 578–584. https://doi.org/10.1002/leap.1404

Zhu, H., & Lei, L. (2022). A dependency-based machine learning approach to the identification of research topics: A case in COVID-19 studies. *Library Hi Tech, 40*(2), 495–515. https://doi.org/10.1108/LHT-01-2021-0051

Zhu, H., Lei, L., & Craig, H. (2020). Prose, verse and authorship in dream of the red chamber: A stylometric analysis. *Journal of Quantitative Linguistics, 28*(4), 1–17. https://doi.org/10.1080/09296174.2020.1724677

Zyoud, S. H., & Fuchs-Hanusch, D. (2017a). A bibliometric-based survey on AHP and TOPSIS techniques. *Expert Systems with Applications, 78*, 158–181. https://doi.org/10.1016/j.eswa.2017.02.016

Zyoud, S. H., & Zyoud, A. H. (2021). Coronavirus disease-19 in environmental fields: A bibliometric and visualization mapping analysis. *Environment, Development and Sustainability, 23*(6), 8895–8923. https://doi.org/10.1007/s10668-020-0104-5

Zyoud, S. H., & Fuchs-Hanusch, D. (2017b). Estimates of Arab world research productivity associated with groundwater: A bibliometric analysis. *Applied Water Science, 7*(3), 1255–1272. https://doi.org/10.1007/s13201-016-0520-2

Zyoud, S. H., & Fuchs-Hanusch, D. (2020). Mapping of climate change research in the Arab world: a bibliometric analysis. *Environmental Science and Pollution Research, 27*(3), 3523–3540. https://doi.org/10.1007/s11356-019-07100-y