TRENDS AND PATTERNS OF TEXT CLASSIFICATION TECHNIQUES: A SYSTEMATIC MAPPING STUDY

Maw Maw[1], Vimala Balakrishnan[2*], Omer Rana[3], Sri Devi Ravana[4]

[1,2,4]Faculty of Computer Science and Information Technology, University of Malaya,
50603 Kuala Lumpur, Malaysia

[3]School of Computer Science and Informatics, Cardiff University, Wales, United Kingdom

Email: mawmaw@siswa.um.edu.my[1], vimala.balakrishnan@um.edu.my[2*](corresponding author),
RanaOF@cardiff.ac.uk[3], sdevi@um.edu.my[4]

DOI: https://doi.org/10.22452/mjcs.vol33no2.2

ABSTRACT

Due to the mass availability of textual data on Web, text classification (TC), classifying texts into predetermined sets becomes a spotlight for researchers. A number of TC applications have been proposed yet very few studies reported an overview of TC research area in a proper and systematic manner. This paper aims to provide an overview of TC research trends and gaps by structuring and analyzing research patterns, encountered problems and problem-solving methods in TC. In other words, this study highlights problem types, data sources, choice of language of text and types of applied techniques in TC. An intensive systematic study is conducted by applying guidelines proposed by Petersen and colleagues in 2007. In this paper, ninety-six literatures from five electronic databases from 2006 to 2017 were systematically reviewed and followed each and every step properly in accordance with systematic mapping study. Nine main problems in TC research area were identified and significant findings which highlighted the evolution of TC research within the past 12 years were investigated. Different from other review articles, this paper highlighted issues and technical gaps of TC area in a useful and effective manner.

Keywords: Machine learning techniques, Text classification, Text categorization, Natural language processing (NLP), Systematic mapping (SM)

1.0 INTRODUCTION

TC is an important segment of text mining [1] as well as the vital area of research in Natural Language Processing (NLP) [2]. TC can be defined as a process that assigns a given document to a set of pre-defined categories based on its textual contents and extracted features [3]. Generally TC process is based on four main phases: preprocessing/document representation phase, feature extraction, feature selection/feature transforming phase and finally machine learning/classification phase [4]. TC task can be accomplished by human experts as the need to understanding terms and knowledge processing necessitates it [5]. The tremendous growth of digitized text on Web has prompted researchers to focus on text related tasks including categorizing, summarizing, clustering and classifying of the texts into specific classes automatically. Challenges in TC exist due to the complexity of natural languages and unstructured text format. Moreover, it is impossible to apply manual classification on billions of text documents as it is a time consuming and labor intensive process [6], [7].

Evidently, numerous automatic TC techniques have been proposed to cover inadequacy of a specific technique in terms of performance, speed and usability [2], [8]. The most common and standard method in TC is to represent each data item by frequencies of a word or words and train data with different classifier models such as Support Vector Machine (SVM), Naïve Bayes (NB) and decision trees etc. [9]. Despite the robustness of state-of-the-art approaches, many studies pointed out the weaknesses of some of the traditional machine learning approaches and proposed ways to achieve better performances. For example, in [10], the authors proposed a model with standard kernel functions to upgrade the performance of SVM. Nevertheless, some state-of-the-art approaches yielded worse results when applied on languages with scarce resources such as Chinese and Arabic [11][5][12] . Hence, underlying problems within traditional machine learning techniques have led to newer and more effective techniques to emerge as a mean to cover the flaws [13][14]. One main challenge in text processing is to deal with semantic and syntactic structures of human languages. Therefore, recent trends in TC had moved towards exploitation of deep learning approaches such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) etc. which provided more accurate results [15], and worked well in text processing tasks without the need of prior semantic knowledge of specific languages[16].

TC is an active and broad research area and its techniques were applied in diverse domains. Recently, it has received more attention from researchers as rapid growth of digitized data demands a need for more efficient TC techniques. Consequently, several primary studies of different TC techniques have been reported such as in [17][18] and a good number of surveys and reviews of these techniques such as in [4][19] have been conducted. Many of these studies focused on performance measurements of selected classifiers when applied to specific datasets. Additionally, most of them also emphasized on listing the available text classifiers and; describing features of those techniques [4], [20]. However, scattered published works in TC makes it difficult for practitioners and future researchers to access up-to-date coverage of the field of interest. It has been identified that only minimum attempts of systematic reviews of reported techniques have been conducted. Additionally, very few studies provided an overview of TC in a well-structured manner [21]. Aiming to fill this gap, this study provides a concise overview of TC trends and gaps. Unlike Systematic Literature Review (SLR) studies which require a deep analysis of specific outcomes and experimental designs of the literatures[22], the main goal of a Systematic Mapping (SM) study is to yield an overview of a specific research area[23] without the requirement of an in-depth analysis of the literatures. It focuses on systematic classification of selected literature, thematic analysis and determining the maturity of publications in selected years [23]. To the best of our knowledge, this study will be the first SM study on TC techniques based on the problems and limitations encountered by researchers in TC. This study would highlight a broader view of existing TC techniques and, would serve as a useful reference for future researches with regards to TC problems.

The main body of the article is structured into 6 sections; starting with the literature review of the topic in Section 1, discussion of related works in Section 2, introducing stages of SM study as core methodology applied in this paper in Section 3, providing analysis of key findings in accordance with research questions and classification schemes in Section 4, describing the limitations and potential future works in Section 5 and concluding with summary of the entire investigation made based on the authors' views in Section 6.

2.0 RELATED WORKS

TC research domain is a well-developed area of research. Though SM study within TC techniques have not been reported, however, a number of articles that emphasized on in-depth literature reviews had reported in past years. One study reported and highlighted every phase of TC and commonly applied algorithms and methods [4]. Review papers often discuss literature of a specific domain in-depth, for example, one study performed a systematic literature review (SLR) on theory and methods of document classifications and text mining approaches [22]. The authors discussed a great deal of details document representation phase and machine learning phase of TC processes, listing most commonly used approaches in each phase. Similarly, another SLR was reported [13] focusing on investigation of commonly applied feature selection and document representation methods, and providing list of most applied datasets. In [24], authors analyzed a review on TC algorithms by comparing among K's Nearest Neighbor (KNN), SVM and NB in terms of performance with accuracy of 86.3%, 86.3% and 73.4% respectively. Another interesting survey focused on investigating and comparing the effectiveness of three machine learning algorithms: SVM, NB and decision tree. The authors identified that SVM performed well and yielded better accuracy compared to the two other algorithms [19]. In 2005, a survey on text categorization techniques was reported [20] but it only listed existing text categorization techniques with no review of its existing gaps and limitations. Most of previous studies provided a similar pattern of literature review of existing TC techniques and how they work and were focusing on comparing various machine learning algorithms. Despite a systematic literature review in TC research [13] was published, it was observed that there were less attempt made to research gaps and trends of existing TC techniques from a high level perspective (i.e. classification of overall TC research area of various aspects such as problem types, type of datasets and types of proposed solutions etc.). These gaps motivated us to work on SM study by classifying different significant categories of TC techniques.

3.0 METHODOLOGY

This study applied SM process as the main methodology, as introduced in 2007 by Petersen et al. [23] with an updated guidelines in 2015 [21]. The basic idea of SM is to develop a classification scheme for the selected literatures. The technique encompasses the formulation of research questions, search for relevant papers, selected paper screenings, key wording of abstracts and data extraction and mapping as shown in Fig 1. The following sub-sections elaborate each of these steps in detail.

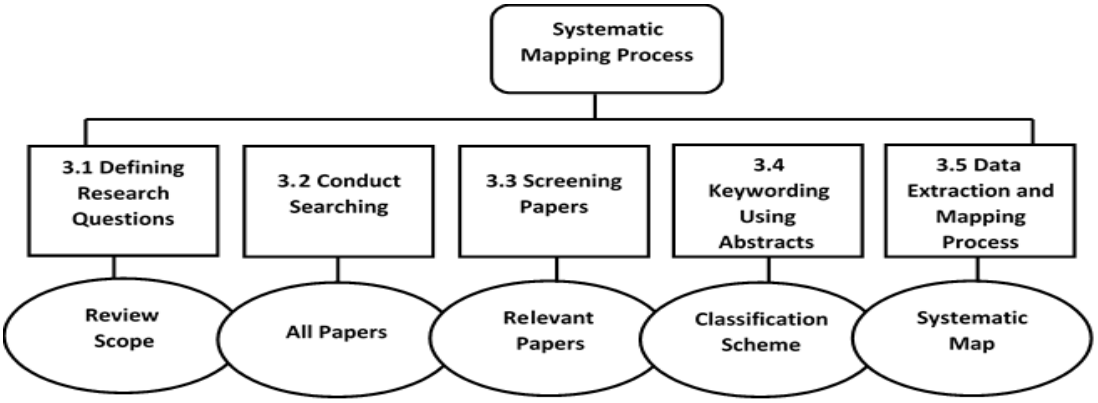

Fig. 1: Stages of SM process [23]

3.1 Formulating Research Questions (RQs)

This step sets scope of the whole investigating process. Two main research questions that embody our goals were defined as follows:
RQ1: How did trends and patterns of TC change within the past 12 years (i.e. 2006 – 2017)?
RQ1.1: How was the research growth of TC techniques in terms of publication distribution within the past 12 years?
RQ1.2: What were the highly focused TC phases in solving identified TC problems within the past 12 years?
RQ2: What are the technical gaps in TC within the past 12 years (i.e. 2006 – 2017)?
RQ2.1: What kind of TC problems did researchers try to solve within the past 12 years?
RQ2.2: What were the solutions proposed by researchers to fill the technical gap(s) in the past 12 years?

In reference to RQ1, research trends of TC area were studied by investigating the research purposes or changes of technical trends and publication frequency. Regarding to RQ2, the research gap or weak points in TC research area were identified by examining prominent issues and types of problems in TC area and type of solutions.

3.2 Search Criteria

In search process, several further steps are required to attain a qualified set of literature including defining keywords for building a search string, determining specific electronic databases and setting the search period. First, the term "text classification" was applied as main keyword and searched in selected electronic databases. To cover all literatures in text classification research area, the term "text" was rephrased with "document" and "classification" with the term "categorization" and searched. Similarly, the term "applications" was interchangeably applied in search with the term "techniques". Five high-impact digital databases were selected due to availability of high quality papers and popularity among researchers. The search period was limited from 2006 to 2017 as 12-year-period can be considered as ample coverage to demonstrate TC research trends in up-to-date manner. The number of literature was eliminated by selecting only journals and conference articles written in English in the preliminary search. The summary of search criteria is organized in Table 1.

Table 1 Summary of the search criteria

No.	Category	Details information
1	Applied keywords	Text classification techniques; document classification techniques; text categorization techniques; document categorization techniques; TC techniques; text classification applications
2	Search string	(((Text OR Document) AND (Classification OR Categorization)) AND (Techniques OR Applications))
3	Electronic databases	IEEExplore; ACM Digital Library; Springer; Science Direct; Google Scholar
4	Search period	2006 to 2017 (12 years)

3.1.3 Screening of Literature

In screening phase, most relevant studies which would closely address the pre-defined research questions were filtered by applying inclusion and exclusion criteria. Papers were initially evaluated by checking titles and abstracts. Introductions were screened when only titles and abstracts could not provide enough information to decide inclusion or exclusion of specific literatures. Table 2 shows criteria for decision of inclusion and exclusion of articles.

Table 2: Inclusion and exclusion criteria for screening the most relevant literatures

No	Inclusion criteria	No.	Exclusion criteria
	A literature was included if:		**A literature was excluded if:**
1.	it highlights one or more problems/weaknesses of existing TC techniques with proposed solution(s)	1.	it does not identify any problems in existing TC techniques and does not propose any solution
2.	it emphasizes on digital TC problems and techniques	2.	it emphasizes on hand-written TC problems rather than digital TC
		3.	it focuses on images and videos rather than on digital texts
		4.	it focuses only on comparative evaluation of performance of existing common TC techniques by applying a specific dataset

As TC is a broad research area, a bulk of articles and conference proceedings were publicly available. From five chosen electronic databases, total of 14,016 articles were collected by applying query described in section 3.2. The statistical results of searching process are shown in Fig 2.

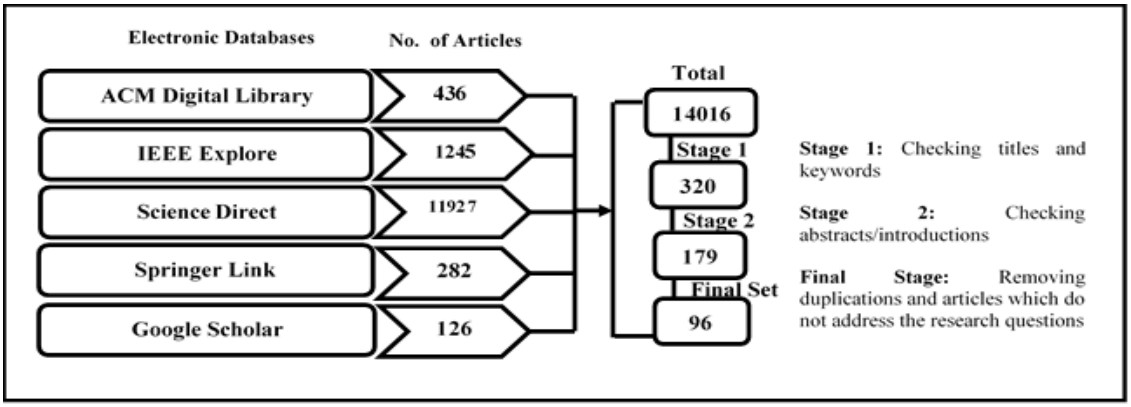

Fig. 2: Summary of the literature screening steps

Initial huge amount of resources was reduced to 320 by checking article titles and keywords in stage 1. After eliminating articles with unclear abstracts and introductions, 179 articles remained in stage 2. Many studies were excluded in this stage because problems related to TC techniques were not stated or stated incompletely. For last stage, articles were examined thoroughly based on criteria specified in Table 2. Articles which do not comply with inclusion criteria and are duplicated were excluded in this stage. Finally, 96 papers were filtered as final list for further analysis.

3.4 Building a Classification Scheme

Building a classification scheme is based on keywording strategy. Applying keywording strategy, abstracts were studied to obtain prominent keywords and to extract categories. In [25], authors highlighted that there are two traditional ways of building up a classification scheme. Bottom-up approach is a way of building a scheme by reading selected articles and extracting significant categories while top-down approach is a way to structure a scheme using general knowledge of a specific field. A more effective way of building a classification scheme appeared to be hybrid approach, which combines both top-down and bottom-up approaches [25]. Therefore, it was decided to build the

scheme in a hybrid method. Using this approach, classification properties from reading abstracts as well as from our general knowledge were extracted. Finally, a classification scheme was built with following categories.

(a) Problem type/research purpose: This includes type of loopholes or weaknesses of previously proposed systems or approaches in TC techniques. In other words, we would like to identity reasons or research purpose(s) of why a specific research is necessary to be carried out. **(b) Type of proposed solution:** This investigates the type of proposed solutions whether they are combined approach (i.e. a combination of existing state-of-the-art approaches); direct approach (i.e. applied only one existing state-of-the-art approach); novel approach (i.e. an original work); modified approach (i.e. proposed a solution by making changes on existing state-of-the-art approaches); extended approach (i.e. added a supplement solution to existing approaches)**(c) Applied TC phases:** This determines the phase(s) in which the proposed solutions fall into, namely, document representation, feature extraction, feature selection or classifying (or machine learning phase) to solve a specified TC problem. **(c) Dataset:** This identifies sources of dataset on which experiments were conducted. Type of data set is also classified either as publicly available datasets or manually extracted datasets. Publicly available datasets are regarded as benchmarks datasets and are downloadable free of charge for TC research while extracted datasets are prepared by researchers themselves. **(d) Language:** This characterizes language or languages of text applied by researchers in selected studies. **(e) Ability to compare to other approaches:** This category investigates the fact that whether the authors had shown the ability to compare their approaches with other existing work(s) or to state-of-the-art approaches to check the strength of proposed solutions.

3.5 Data extraction and a systematic map creation

After setting up a classification scheme, the authors analyzed final selected papers and extracted necessary information based on classification scheme and research questions. A systematic map which visualizes the specified problems with the type of techniques is shown in section 4.1.7.

4.0 RESULTS AND DISCUSSION

This section presents all the findings obtained based on the formulated RQs and the classification scheme of SM. Salient findings are discussed in detail and responses to the RQs are specifically elaborated in section 4.2.

4.1 Results

This section reports the major findings of the systematic mapping based on the RQs.

4.1.1 Literature search results

Conference proceedings were considered in our study for two reasons: first, the duration which takes to publish articles in conference proceedings is shorter compared to journal articles, hence, findings and works provided in conference papers are deemed to be more up-to-date. Second, some of these papers were unique in terms of highlighting TC issues and proposed methods [5][26][27]. The distribution of publication within the last 12 years of selected sources is shown in Fig 3.

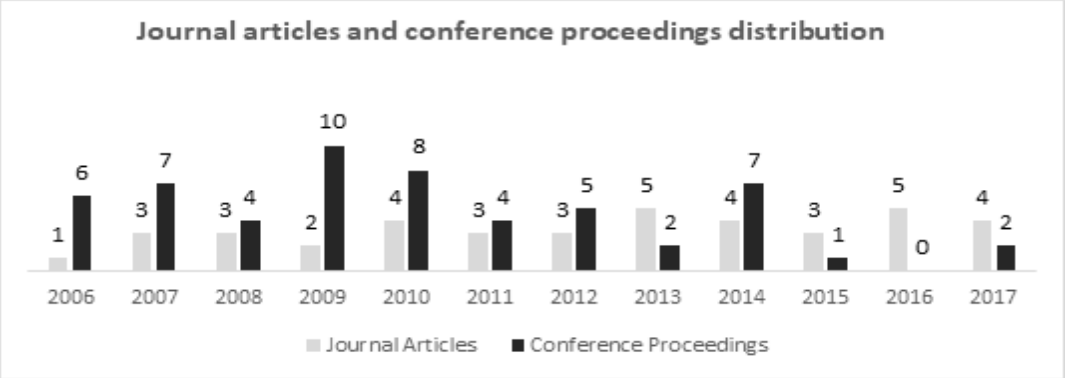

Fig. 3: Publishing summary from 2006 to 2017

Generally, research interest in TC was high since 2006 until 2017 whereby rate of conference publications was generally higher. The years 2013 and 2016 had the highest rate in publication of journal articles. In the selected 96 studies, 54 were from conference proceedings and 42 were published in journals. In terms of electronic databases (see

Fig 4), it was found that majority of selected articles were from IEEE followed by Elsevier (Science Direct). Furthermore, rate of journal publication was gradually increased and was stable until 2011 except 2009. In 2009 there was, a big number of difference between two types of publication as related statistical data shows that conference article publication was five times higher than journal article publication.

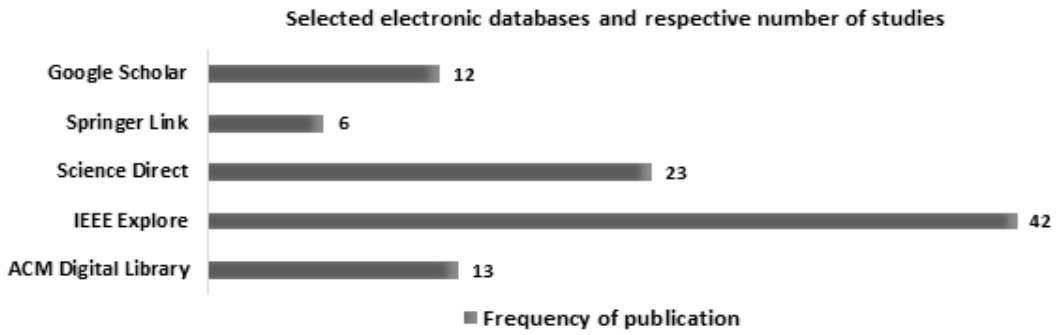

Fig. 4: Electronic databases and respective number of selected studies

4.1.2 Problem type/research purpose

Nine types of common problems or research purposes in TC area were identified. They are: **(a) Performance enhancement:** This problem refers to classification accuracy of existing systems. In other words, researchers want to solve the matters which effect performance of existing techniques by proposing new approaches. Examples are in [10], [28]–[30]. **(b) Malfunction of existing techniques:** This problem concerns weaknesses or loopholes of existing approaches. Researchers point out gaps of current techniques which do not work well in a specific area and try to solve that problem by proposing new methods such as in [26], [31], [27], [32]–[35]. **(c) Insufficiency of resources:** This is closely related to limitation of a specific area than to be referred to as a problem. Researchers face difficulties in the specific area of researches due to a lack of necessary resources, for example, unavailability of corpus or labeled datasets in a specific language. Examples include the works of [5], [36], [37]. **(d) Reduction of human effort requirement:** In TC, labeling the documents by human annotators is a necessary step, but it is a huge burden on human experts as they need to handle a large number of documents. It is also a time consuming, costly and labor-intensive task. So, researchers want to reduce the human involvement by proposing alternative ways such as in [7], [38], [17], [39], [40]. **(e) Reduction of storage requirement/cost:** Some studies point out the larger the text documents to be classified, the more storage is needed and the cost becomes higher, for example, previous works of [2], [8], [18]. **(f) Enhancing functions of a specific area with TC techniques:** This refers to studies that use TC techniques to enhance other functions in a particular domain. For example, researchers believe that operating flow in online health community can be improved by applying TC approaches, for example [41], [42]. **(g) Improving the all-round development:** Some studies proposed techniques to improve the all-round development including reducing computational time, cost, classification accuracy as in [43], [44]. **(h) Scarce research on a specific area of study:** This is also closely related to the research purpose. Some researchers proposed techniques on a specific research area on which a little attention was paid such as in [45], [46]. **(i) Investigating effects of a specific technique:** This refers to studies which explore the usefulness of a specific technique on a selected research area such as in [47], [48]. Table 6 summarizes different problem types and number of studies on solving specific problems. Detailed studies and respective references are shown in Appendix A.

Table 3: Problem type(s) or research purpose(s) of each selected literature

ID	Type of problems/research purpose	Number of studies
#PT1	Performance enhancement	34
#PT2	Malfunction of existing techniques	31
#PT3	Insufficiency of resources	5
#PT4	Reduction of human effort requirement	8
#PT5	Reduction of storage requirement/cost	5
#PT6	Enhancing functions of a specific area with text classification techniques	7
#PT7	Improving all-round development	4
#PT8	Scarce researches on a specific area of study	2
#PT9	Investigating effects of a specific technique	7

4.1.3 Type of proposed solution

A number of approaches or solutions were proposed in past 12 years. Based on type of proposed solutions, type of approaches can be categorized into five main groups: direct approach, combined approach, novel approach, modified approach and extended approach. Types of proposed solutions and respective number of studies are comparatively illustrated in Fig 6.

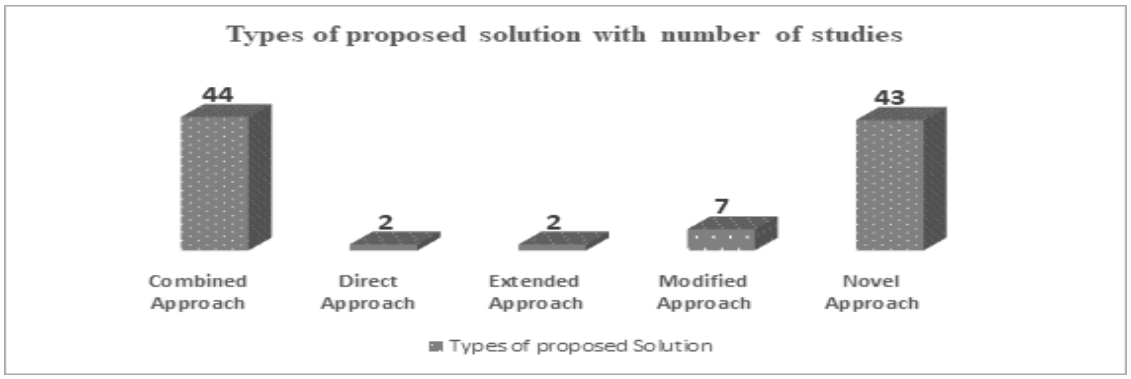

Fig. 6: Number of studies with respective type of proposed solution

4.1.3 Applied TC phase(s)

All proposed techniques were based on four main TC steps: data representation, feature selection, feature extraction, and classification. Fig 7 shows the number of studies which emphasized on either of four major TC phases from 2006 to 2017.

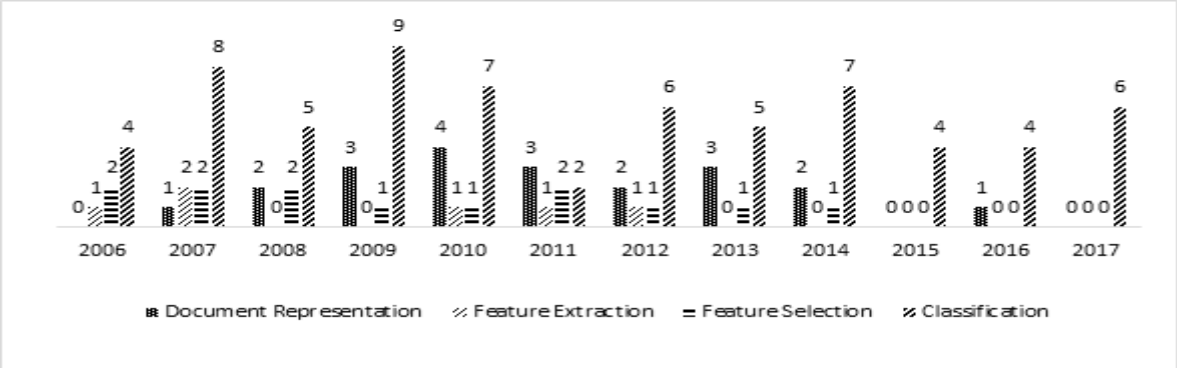

Fig. 7: Statistics of focused TC phases from 2006 - 2017

4.1.4 Dataset

Aiming for evaluating and comparing performance, researchers conduct experiments on selected data sources. Fifty percent of studies conduct text classification experiments on one data source while another half is conducting on more than one data sources. In fact, it cannot be determined the performance level until it works well on varieties of datasets. When dealing with experiments, two different types of datasets were mainly applied in previous studies: publicly available dataset and extracted datasets. Data need to be extracted when data sources are rarely available. However, some studies conducted experiments on both types of datasets [32], [49], [50]. Over 70% of studies utilized publicly available datasets and only14 studies used datasets extracted from different sources. Seven studies applied both type of datasets to prove the performance of their approaches. It was found that 40 different data sources in different languages were applied in previous literatures, mostly from Reuter-21578 and 20 Newsgroup.

4.1.5 Language

By reviewing selected studies, results show that 12 different languages of text were applied to experimental processes of proposed techniques. Out of 96 studies, two studies experimented on five languages: English, French, Italian, German, and Spanish [44], [51] By observing statistics from Fig 8.

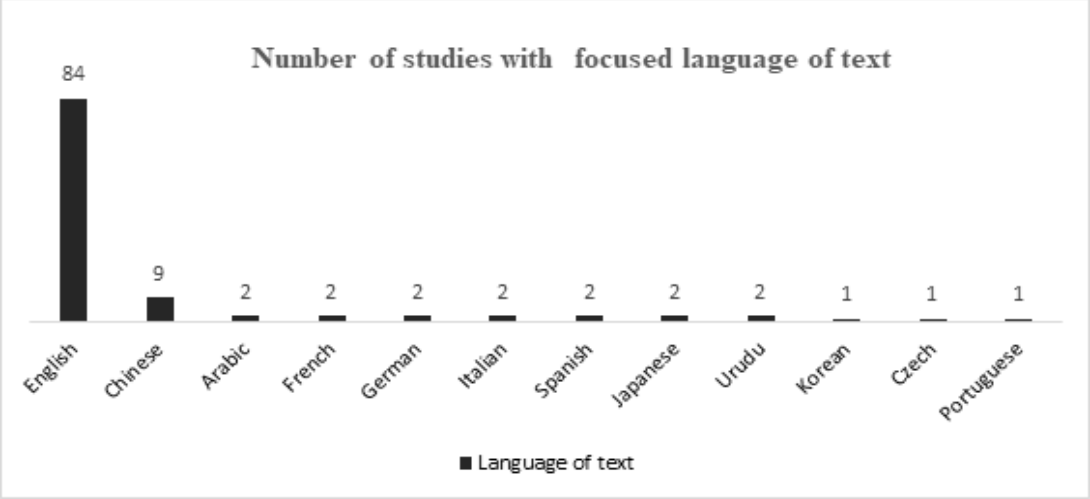

Fig. 8: Graph of applied languages of text in literatures

4.1.6 Ability to compare with other approaches

Nature of TC techniques often prompts researchers to focus on the strength of their proposed systems over other state-of-the-art approaches. Researchers typically compare the performance of the proposed systems to one or more contemporary approaches to prove their systems work better than existing ones. In this way, the significance of new proposed techniques becomes more prominent. By analyzing selected studies, it is found that 66% of all literatures have proved performance of their approaches are better by testing on existing systems while 34% did not conduct any comparisons to other mechanisms.

4.1.7 Systematic map

In this section, a bubble plot is created to report research gaps and future trends in each TC phase according to nine identified problem types. This provides a more visualized perspective to help future researchers understand phases of TC that have been studied and areas that have gaps. A systematic map of identified problems and four main TC phases in which researchers have attempted to solve by proposing new approaches is illustrated in Fig 9.

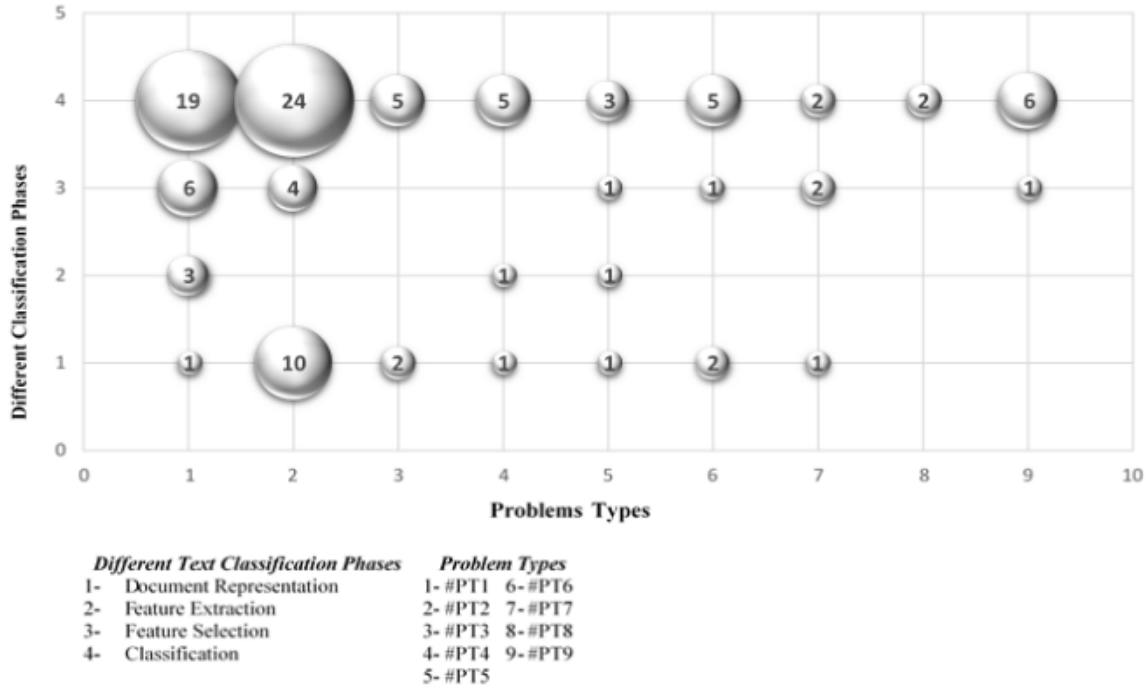

Fig. 9: Systematic map on identified problems and focused areas

4.2 Discussion

This section presents the discussion on all the RQs and sub-RQs stated in Section 3.

RQ1: How did trends and patterns of TC change within the past 12 years (i.e. 2006 – 2017)?

Research trends and patterns of TC area are identified in terms of research growth and how applied techniques have evolved in TC area. Based on findings, TC research area has been an active and trendy research area over the last 12 years as there was not any publishing gap in any year from 2006 until 2017.

Manual data labeling is an important step in TC problem, and it becomes a challenge when a large amount of data items are involved [3], [38]. Therefore, finding ways to reduce human involvement has been a crucial problem in TC research as labeling by human annotators is costly, and their decisions can be biased in some cases. Nevertheless, human involvements are important due to the research demands of classification tasks in trending research areas. Some research areas such as classification of sarcastic sentences, detecting hate speech and dangerous speech and tracking act of terror, etc. are subjective in nature, hence the knowledge of human experts in the labeling processes are required.

Problem relating to insufficient resources includes labeled negative data and corpuses which are essential for multilingual TC or cross language TC are not available easily. This problem is more closely concerned on classification problems with different language of text rather than English. According to reports in section 4.1.5, within 12 years, English was mostly focused language of text in experiments of TC techniques. Based on different structure and formation of different natural languages, extensive preprocessing tasks are required to perform for dealing with some natural languages. For example, Chinese text segmentation is necessary prior to other text processing tasks when working with Chinese digitized texts [5]. This fact leads to emerge novel approaches to take over of existing approaches as in [11]. In earlier years, Bag of Word (BOW) approach was popular for using as feature set but most studies identified weak points in BOW approach[52] and proposed new techniques to use semantic knowledge, ontologies and corpus-based knowledge in text categorization processes [10], [53]. It can be said that it was a change of trend in feature selection phase from traditional BOW methods to semantic cooperated approaches. Yet those approaches still could not give satisfied results in some area [32]. In fact, TC researches are mainly connected with Natural Language Processing (NLP) tasks and Information Retrieval research area as it has to deal with natural language of text. Natural languages have complexity in structure of sentences and usages some of which could not feed into existing classifiers without extra pre-processing tasks such as part of speech (POS) tagging, dependency parsing, segmentation of the words and machine translation. Depending on data size and growth in classification domains, the accuracy in some of existing approaches is reduced[15]. Based on flaws of existing preprocessing

techniques and obstacles in dealing with natural languages understanding tasks, researchers were seeking for techniques which could win high accuracy with less complexity. In this way, technical trends in TC switched from traditional machine learning to Deep Learning (DL) techniques since 2016. In 2016 and 2017, 9 out of 11 selected studies were applying deep learning approaches. Some of DL techniques have proven ability to work better than other state-of-the-art approaches. For example, Convolutional Neural Network (CNN) performs better without prior syntactic or semantic knowledge of a sentence [54].

RQ2: What are the technical gaps in TC research area within the past 12 years?

TC has four main phases: document representation or data preprocessing, feature extraction, feature selection and the classification or machine learning phase. The results in section 4.1.3 showed that most researchers proposed methods to fill the gap in the classification or machine learning phase. To be specific, a total of 67 (i.e. 71%) studies proposed solutions for classification phase while 21(i.e. 22%) studies were dedicated to solutions to document representation phase. This is probably because the classification phase is considered as the most vital phase of the whole TC process, and common issues such as data imbalance and overfitting are highly related to this phase. Therefore, this probably explains the spike in studies related to the classification/machine learning phase.

Most research works were concerned about the performance of the entire classification task as higher performance rate indicates the usefulness of the proposed solution. As shown in section 4.1.2, 35% of the literatures aimed to improve the performance in classifying texts, using a number of combined, modified, direct, extended and novel TC techniques to solve the underlying issues, as illustrated in Fig 6. Of these, most studies proposed novel approaches and combined approaches, with 43 and 44 studies, respectively. However, 34% of total studies failed to do any comparable testing with existing approaches. Additionally, there were very minimum evidence of evaluating proposed mechanisms on different language of text. Therefore, this fact could be considered as a gap in TC research area. It was observed that researchers were equally paying attention to both problems of insufficiency of resources and reducing human effort requirement. Due to increased amount of digitized text, previous researchers had attempted to reduce the memory usage and computational cost [8], [18], [55], [56]. Generally, the performance of a TC system is measured based on accuracy in classification and recall rate. On the other hand, length of computation time, cost-effectiveness and minimum demand of system requirements are also the vital factors to be considered in measuring performance of a TC system. However, this seems to be a challenging task to researchers as a few techniques were proposed to improve all-round development (i.e., enhancing performance, reducing memory usage and computational cost) of a specific area [44], [57]. An interesting observation from the systematic map is the lack of studies on feature extraction techniques in the past 12 years.

Another gap in TC research area was less research works emerged with different languages of text rather than English as discussed in previous section 4.2.1. Ninety percent of all studies emphasized text written in English while less than 10% of Chinese texts were applied in experiments of TC problems [56], [58], [59]. Although other languages such as Arabic, French and German are common languages, only a few studies had conducted on these languages [46][12]. One reason could be the scarcity of resources such as ontologies, data preprocessing tools and due to unavailability of labeled data sets in non-English texts [46]. This fact would still leave some gaps in TC research area especially when dealing with text from languages other than English.

5.0 LIMITATIONS AND FUTURE WORK

Though this study has attempted to cover and include all related articles in research, due to limited electronic databases in article search, some important studies from other databases may have been missed.

According to our findings, research on different languages of text is limited even among common languages such as Arabic and Chinese. To fill this gap, future studies should aim towards in-depth analysis of technical gaps in multilingual TC techniques, and the importance of availability of language resources in TC problems. Furthermore, it would be interesting to apply DL approaches on languages which were less applied in TC researches, and to compare the results obtained from experiments with English texts.

6.0 CONCLUSION

This paper reported trends, gap and research patterns of TC techniques based on problems and research correlation of previously proposed TC phases and associated problems. SM study was conducted by searching resources from five different electronic databases from 2006 to 2017 and adopting guidelines of [23]. As a general remark, regardless of the types of the article, research maturity in TC has developed since 2006 based on the frequency of publication

distribution. Nine different types of problems/research purposes of previous studies were identified. Different varieties of datasets, usage of text languages and focus area of proposed techniques were reported in our study. It was determined that majority of previous studies had focused mainly on the classification phase of TC process to solve performance problems and to overcome the weaknesses of existing techniques. Additionally, a comprehensive discussion of TC trends and gap in twelve years were identified and further elaborated to provide a summarized overview for the benefits of future researchers and practitioners.

ACKNOWLEDGEMENT

The authors would like to thank and acknowledge the support provided by the Ministry of Education under research grant reference number: Fundamental Research Grant Scheme 2019 [FP109-2018A].

REFERENCES

[1] J. J. Adeva, J. M. P. Atxa, M. U. Carrillo, and E. A. Zengotitabengoa, "Automatic text classification to support systematic reviews in medicine," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1498–1508, 2014.

[2] L. Shi, J. Zhang, E. Liu, and P. He, "Text Classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines," *Third Int. Conf. Nat. Comput. (ICNC 2007)*, vol. 1, no. 5, pp. 7–10, 2007.

[3] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-Scale Text Categorization by Batch Mode Active Learning," *Proc. 15th Int. Conf. World Wide Web - WWW '06*, pp. 633–642, 2006.

[4] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Trans. Comput.*, vol. 4, no. 8, pp. 966–974, 2005.

[5] X. Luo, W. Ohyama, T. Wakabayashi, and F. Kimura, "Automatic chinese text classification using character-based and word-based approach," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 329–333, 2013.

[6] F. Clarizia, F. Colace, M. De Santo, L. Greco, and P. Napoletano, "A new text classification technique using small training sets," *2011 11th Int. Conf. Intell. Syst. Des. Appl.*, pp. 1038–1043, 2011.

[7] F. Tchiegue, R. Li, and S. Ma, "A web text classification technique for unlabeled training samples," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2015–Novem, pp. 437–440, 2015.

[8] C. F. Tsai, Z. Y. Chen, and S. W. Ke, "Evolutionary instance selection for text classification," *J. Syst. Softw.*, vol. 90, no. 1, pp. 104–113, 2014.

[9] R. Angelova and G. Weikum, "Graph-based text classification: learn from your neighbors," *Proc. 29th Annu. Int.*, pp. 485–492, 2006.

[10] B. Altinel, M. Can Ganiz, and B. Diri, "A corpus-based semantic kernel for text classification by using meaning values of terms," *Eng. Appl. Artif. Intell.*, vol. 43, pp. 54–66, 2015.

[11] Y.-W. Chen, J.-L. Wang, Y.-Q. Cai, and J.-X. Du, "A method for Chinese text classification based on apparent semantics and latent aspects," *J. Ambient Intell. Humaniz. Comput.*, vol. 6, no. 4, pp. 473–480, 2015.

[12] M. Hadni, A. Lachkar, and S. A. Ouatik, "A new and efficient stemming technique for Arabic Text Categorization," *Proc. 2012 Int. Conf. Multimed. Comput. Syst. ICMCS 2012*, pp. 791–796, 2012.

[13] R. Jindal, "Techniques for text classification : Literature review and current trends," vol. 12, no. 2, pp. 1 28, 2015.[Online]. Available: http://www.webology.org/2015/v12n2/a139.pdf. [Accessed Feb. 16, 2018]

[14] P. Wang, J. Hu, H. J. Zeng, and Z. Chen, "Using Wikipedia knowledge to improve text classification," *Knowl. Inf. Syst.*, vol. 19, no. 3, pp. 265–281, 2009.

[15] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "HDLTex: Hierarchical Deep Learning for Text Classification," 2017. [Online]. Available: arXiv:1709.08267v2. [Accessed Feb. 16, 2018].

[16] X. Zhang and Y. LeCun, "Text Understanding from Scratch," 2016. [Online]. Available: arXiv:1502.01710v5. [Accessed Feb. 16, 2018].

[17] N. (2008). Karman, S. S., & Ramaraj, "Similarity -Based Techniques for Text Document Classification.pdf." MedWell Online, pp. 58–62, 2008.

[18] L. Liu and Q. Liang, "A high-performing comprehensive learning algorithm for text classification without pre-labeled training set," *Knowl. Inf. Syst.*, vol. 29, no. 3, pp. 727–738, 2011.

[19] G. S. Chavan, S. Manjare, P. Hegde, and A. Sankhe, "A Survey of Various Machine Learning," vol. 15, no. 6, pp. 288–292, 2014.

[20] Meenakshi and S. Singla, "Review Paper on Text Categorization Techniques," no. April, pp. 139–143, 2015.

[21] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Inf. Softw. Technol.*, vol. 64, pp. 1–18, 2015.

[22] B. Baharudin, L. H. Lee, and K. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification," *J. Adv. Inf. Technol.*, vol. 1, no. 1, pp. 4–20, 2010.

[23] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering," *12th Int. Conf. Eval. Assess. Softw. Eng.*, vol. 17, pp. 1–10, 2007.

[24] V. C. Gandhi and J. a Prajapati, "Review on Comparison between Text Classification Algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 1, no. 3, pp. 1–4, 2012.

[25] A. Seriai, O. Benomar, B. Cerat, and H. Sahraoui, "Validation of Software Visualization Tools: A Systematic Mapping Study," *2014 Second IEEE Work. Conf. Softw. Vis.*, pp. 60–69, 2014.

[26] L. Zhang, Y. Li, Y. Xu, D. Tjondronegoro, and C. Sun, "Centroid training to achieve effective text classification," *DSAA 2014 - Proc. 2014 IEEE Int. Conf. Data Sci. Adv. Anal.*, pp. 406–412, 2014.

[27] X. Zhang and W. Xiao, "Clustering based two-stage text classification requiring minimal training data," *Comput. Sci. Inf. Syst.*, vol. 9, no. 4, pp. 1627–1643, 2012.

[28] H. H. Malik, D. Fradkin, and F. Moerchen, "Single pass text classification by direct feature weighting," *Knowl. Inf. Syst.*, vol. 28, no. 1, pp. 79–98, 2011.

[29] J. Yun, L. Jing, J. Yu, and H. Huang, "A multi-layer text classification framework based on two-level representation model," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 2035–2046, 2012.

[30] S. H. Lu, D. A. Chiang, H. C. Keh, and H. H. Huang, "Chinese text classification by the Na??ve Bayes Classifier and the associative classifier with multiple confidence threshold values," *Knowledge-Based Syst.*, vol. 23, no. 6, pp. 598–604, 2010.

[31] B. S. Harish, S. V. Aruna Kumar, and S. Manjunath, "Classifying text documents using unconventional representation," *2014 Int. Conf. Big Data Smart Comput. BIGCOMP 2014*, pp. 210–216, 2014.

[32] Y. Wan, T. He, and X. Tu, "Conceptual Graph Based Text Classification," *Prog. Informatics Comput. (PIC), 2014 Int. Conf.*, pp. 104–108, 2014.

[33] K. A. Vidhya and G. Aghila, "Hybrid text mining model for document classification," *2010 2nd Int. Conf. Comput. Autom. Eng. ICCAE 2010*, vol. 1, pp. 210–214, 2010.

[34] R. D. Goyal, "Knowledge Based Neural Network for Text Classification," *2007 IEEE Int. Conf. Granul. Comput. (GRC 2007)*, no. 1, pp. 542–547, 2007.

[35] V. Polychronopoulos, N. Pendar, and S. R. Jeffery, "QuIET: A Text Classification Technique Using Automatically Generated Span Queries," *2014 IEEE Int. Conf. Semant. Comput.*, pp. 52–59, 2014.

[36] T. Peng, W. Zuo, and F. He, "SVM based adaptive learning method for text classification from positive and unlabeled documents," *Knowl. Inf. Syst.*, vol. 16, no. 3, pp. 281–301, 2008.

[37] L. Shi, X. Ma, L. Xi, Q. Duan, and J. Zhao, "Rough set and ensemble learning based semi-supervised algorithm for text classification," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 6300–6306, 2011.

[38] A. C. Köning and E. Brill, "Reducing the human overhead in text categorization," *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, no. Figure 1, pp. 598–603, 2006.

[39] F. Colace, M. De Santo, L. Greco, and P. Napoletano, "Text classification using a few labeled examples," *Comput. Human Behav.*, vol. 30, pp. 689–697, 2014.

[40] H. Han, Y. Ko, and J. Seo, "Using the revised EM algorithm to remove noisy data for improving the one-against-the-rest method in binary text classification," *Inf. Process. Manag.*, vol. 43, no. 5, pp. 1281–1293, 2007.

[41] J. Huh, M. Yetisgen-Yildiz, and W. Pratt, "Text classification for assisting moderators in online health communities," *J. Biomed. Inform.*, vol. 46, no. 6, pp. 998–1005, 2013.

[42] C. H. a. Koster and J. G. Beney, "Phrase-based document categorization revisited," *Proc. PAIR 2009 Work. CIKM 2009*, pp. 49–55, 2009.

[43] B. B. Kiranagi, "A Symbolic Approach for Text Classification Based on Dissimilarity Measure," pp. 104–108, 2010.

[44] Y. Guo and M. Xiao, "Transductive representation learning for cross-lingual text classification," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 888–893, 2012.

[45] A. R. Ali and M. Ijaz, "Urdu text classification," *Proc. 6th Int. Conf. Front. Inf. Technol. - FIT '09*, p. 1, 2009.

[46] M. M. Al-Tahrawi and S. N. Al-Khatib, "Arabic Text Classification Using Polynomial Networks," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 27, no. 4, pp. 437–449, 2015.

[47] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification," *Inf. Sci. (Ny).*, vol. 236, pp. 109–125, 2013.

[48] T. Gonçalves and P. Quaresma, "Using linguistic information to classify Portuguese text documents," *7th Mex. Int. Conf. Artif. Intell. - Proc. Spec. Sess. MICAI 2008*, pp. 94–100, 2008.

[49] R. Nanculef, I. Flaounas, and N. Cristianini, "Efficient classification of multi-labeled text streams by clashing," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5431–5450, 2014.

[50] Y. Song, D. Zhou, J. Huang, I. G. Councill, H. Zha, and C. L. Giles, "Boosting the Feature Space : Text Classification for Unstructured Data on the Web," no. 1, 2006.

[51] M. Amini, C. Goutte, and N. Usunier, "Combining Coregularization and Consensus-based Self-Training for Multilingual Text Categorization," pp. 475–482, 2010.

[52] X. Wang and R. Bai, "Applying RDF ontologies to improve text classification," *Proc. 2009 Int. Conf. Comput. Intell. Nat. Comput. CINC 2009*, no. 2, pp. 118–121, 2009.

[53] H. Yamakawa, J. Peng, and A. Feldman, "Semantic enrichment of text representation with wikipedia for text classification," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, no. 1, pp. 4333–4340, 2010.

[54] M. Amajd, "Text Classification with Deep Neural Networks.", 2017. [Online]. Available: http://ceur-ws.org/Vol-1989/paper27.pdf. [Accessed Feb.16, 2018].

[55] R. L. Liu, "Interactive high-quality text classification," *Inf. Process. Manag.*, vol. 44, no. 3, pp. 1062–1075, 2008.

[56] H. Kong, X. Hao, C. Zhang, S. Wang, X. Tao, Y. Hu, and I. Technology, "EFFICIENT KNN TEXT CATEGORIZATION BASED ON MULTIEDIT AND CONDENSING TECHNIQUES," no. August, pp. 19–22, 2007.

[57] S. Hingmire and S. Chakraborti, "Topic Labeled Text Classification : A Weakly Supervised Approach," *Proc. SIGIR 2014*, pp. 385–394, 2014.

[58] Z. Su, W. Song, D. Meng, and J. Li, "A new associative classifier for text categorization," *Proc. 2008 3rd Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2008*, pp. 291–295, 2008.

[59] S. Zheng, Y. Yang, H. Wu, and W. Liu, "Chinese Text Classification Using Key Characters String Kernel," *Semant. Knowl. Grid, 2009. SKG 2009. Fifth Int. Conf.*, pp. 113–119, 2009.

[60] L. Chen, G. Guo, and K. Wang, "Class-dependent projection based method for text categorization," *Pattern Recognit. Lett.*, vol. 32, no. 10, pp. 1493–1501, 2011.

[61] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 690–701, 2009.

[62] K. Kim, B. S. Chung, Y. Choi, S. Lee, J. Y. Jung, and J. Park, "Language independent semantic kernels for short-text classification," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 735–743, 2014.

[63] D. Miao, Q. Duan, H. Zhang, and N. Jiao, "Rough set based hybrid algorithm for text classification," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9168–9174, 2009.

[64] W. Zhang, X. Tang, and T. Yoshida, "TESC: An approach to TExt classification using Semi-supervised Clustering," *Knowledge-Based Syst.*, vol. 75, pp. 152–160, 2015.

[65] X. Z. D. Zhao, "Data Editing," no. 20080440260, pp. 1–4, 2010.

[66] X. Luo and N. Zincir-Heywood, "Incorporating Temporal Information for Document Classification," *2007 IEEE 23rd Int. Conf. Data Eng. Work.*, pp. 780–789, 2007.

[67] Z. Wang and X. Sun, "Iterative kernel discriminant analysis algorithm for document classification," *Proc. - 2009 Int. Conf. Inf. Eng. Comput. Sci. ICIECS 2009*, pp. 2–5, 2009.

[68] S. J. Lee and J. Y. Jiang, "Multilabel text categorization based on fuzzy relevance clustering," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 6, pp. 1457–1471, 2014.

[69] M. Suzuki and S. Hirasawa, "Text categorization based on the ratio of word frequency in each categories," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2003, pp. 3535–3540, 2007.

[70] D. A. Ostrowski, "A framework for the classification of unstructured data," *ICSC 2009 - 2009 IEEE Int. Conf. Semant. Comput.*, pp. 373–377, 2009.

[71] B. Rujiang and L. Junhua, "A Hybrid Documents Classification Based on SVM and Rough Sets," *2009 Int. e-Conference Adv. Sci. Technol.*, no. 1, pp. 18–23, 2009.

[72] F. Thabtah, W. Hadi, H. Abu-Mansour, and L. McCluskey, "A new rule pruning text categorisation method," *2010 7th Int. Multi-Conference Syst. Signals Devices, SSD-10*, 2010.

[73] M. S. Ahmed, L. Khan, and M. Rajeswari, "Using Correlation Based Subspace Clustering for Multi-label Text Data Classification," *Tools with Artif. Intell. (ICTAI), 2010 22nd IEEE Int. Conf.*, vol. 2, 2010.

[74] D. S. Mansjur, T. S. Wada, and B. H. Juang, "Using kernel density classifier with topic model and cost sensitive learning for automatic text categorization," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 1086–1090, 2009.

[75] H. Y. Wang, "Using weight-retouching and under-sampling SVM approaches for text categorization on imbalanced data," *2009 Int. Conf. E-bus. Inf. Syst. Secur. EBISS 2009*, pp. 1–4, 2009.

[76] H. Guan, J. Zhou, and M. Guo, "A Class-Feature-Centroid Classi fi er for Text Categorization," pp. 201–210, 2009.

[77] S. Dey, "A Multi-classifier System for Text Categorization," *Proc. 2011 ACM Symp. Res. Appl. Comput. (RACS '11)*, pp. 325–329, 2011.

[78] L. Rocha, F. Mourão, A. Pereira, M. A. Gonçalves, and W. Meira Jr, "Exploiting temporal contexts in text classification," *Proceeding 17th ACM Conf. Inf. Knowl. Manag.*, pp. 243–252, 2008.

[79] K. W. Kong, "Data Loss Prevention based on Text Classification in Controlled Environments," *ICISS. Lecture Notes in Computer Science*, Springer, Cham. Vol. 10063, pp. 131-150, 2016.

[80] M. Usman and S. Ayub, "Urdu Text Classification using Majority Voting," vol. 7, no. 8, pp. 265–273, 2016.

[81] S.-B. K. S.-B. Kim, K.-S. H. K.-S. Han, H.-C. R. H.-C. Rim, and S. H. M. S. H. Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466, 2006.

[82] L. Özgür and T. Güngör, "Text classification with the support of pruned dependency patterns," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1598–1607, 2010.

[83] W. Li, D. Miao, and W. Wang, "Two-level hierarchical combination method for text classification," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2030–2039, 2011.

[84] B. Goertzel and J. Venuto, "Accurate SVM Text Classification for Highly Skewed Data Using Threshold Tuning and Query-Expansion-Based Feature Selection," *2006 IEEE Int. Jt. Conf. Neural Netw. Proc.*, pp. 1220–1225, 2006.

[85] W. X. Xiao and X. Zhang, "Active transductive KNN for sparsely labeled text classification," *6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst. SCIS/ISIS 2012*, pp. 2178–2182, 2012.

[86] G. Arevian, "Recurrent neural networks for robust real-world text classification," *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. WI 2007*, no. 2, pp. 326–329, 2007.

[87] P. Achananuparp, X. Zhou, X. Hu, and X. Zhang, "Semantic Representation in Text Classification Using Topic Signature Mapping," *Neural Networks*, pp. 1035–1041, 2008.

[88] P. S. M. Ibañez, "Semi-Supervised Text Classification Using Enhanced KNN Algorithm," *Educacion*, vol. 53, no. 9, pp. 266–276, 2013.

[89] R. Zhao and K. Mao, "Supervised Adaptive-Transfer PLSA for Cross-Domain Text Classification," *2014 IEEE Int. Conf. Data Min. Work.*, pp. 259–266, 2014.

[90] R. Ahmad, S. Ali, and D. H. Kim, "A Multi-Agent system for documents classification," *ICOSST 2012 - 2012 Int. Conf. Open Source Syst. Technol. Proc.*, pp. 28–32, 2012.

[91] W. Lucia and E. Ferrari, "EgoCentric : Ego Networks for Knowledge-based Short Text Classification," *in Proc.of 23rd ACM Int. Conf. on Information and Knowledge Management, CIKM'14, Shanghai, China, November 3-7, 2014*, ACM, 2014. pp. 1079–1088.

[92] L. Lenc and P. Král, "Deep Neural Networks for Czech Multi-label Document Classification," pp. 1–12, 2017. [Online]. Available: arXiv:1701.03849v2. [Accessed Feb. 16, 2018].

[93] A. Conneau, H. Schwenk, Y. Le Cun, and L. Barrault, "Very Deep Convolutional Networks for Text Classification," *Eacl*, vol. 1, no. 2001, pp. 1107–1116, 2017.

[94] Y. Ko and J. Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques," *Inf. Process. Manag.*, vol. 45, no. 1, pp. 70–83, 2009.

[95] A. Onan, S. Korukoglu, and H. Bulut, "LDA-based Topic Modelling in Text Sentiment Classification : An Empirical Analysis," *Int. J. Comput. Linguist. Appl.*, vol. 7, no. 1, pp. 101–119, 2016.

[96] O. Frunza, D. Inkpen, and S. Matwin, "Building systematic reviews using automatic text classification techniques," *COLING '10 Proc. 23rd Int. Conf. Comput. Linguist. Posters*, no. August, pp. 303–311, 2010.

[97] V. N. Garla and C. Brandt, "Ontology-guided feature engineering for clinical text classification," *J. Biomed. Inform.*, vol. 45, no. 5, pp. 992–998, 2012.

[98] A. R. Dengel, "Learning of pattern-based rules for document classification," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, no. Icdar, pp. 123–127, 2007.

[99] Y. L. Y. Li, H. L. H. Lin, and Z. Y. Z. Yang, "Two Approaches for Biomedical Text Classification," *2007 1st Int. Conf. Bioinforma. Biomed. Eng.*, no. 60373095, pp. 310–313, 2007.

[100] Y. Yoon and G. G. Lee, "Two scalable algorithms for associative text classification," *Inf. Process. Manag.*, vol. 49, no. 2, pp. 484–496, 2013.

[101] M. Sato, R. Orihara, Y. Sei, Y. Tahara, and A. Ohsuga, "Japanese Text Classification by Character-level Deep ConvNets and Transfer Learning CHARACTER-LEVEL ConvNet," vol. 2, no. Icaart, pp. 175–184, 2017.

[102] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep Learning for Extreme Multi-label Text Classification," *Proc. 40th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '17*, pp. 115–124, 2017.

[103] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," pp. 1–9, 2016.[Online]. Available: arXiv:1509.01626v3. [Accessed Feb. 16, 2018].

Appendix A: Identified problems/research purposes with respective references

No.	Type of Problem	Number of Studies	Reference
1	Performance Enhancement	34	[10], [11], [28], [42][29][30][60][49] [61][62][63][64][39][50][59][65][66][67][68][69][70][71][12][72][6][73][74] [75][76][77][9][78][79][80]
2	Malfunction of existing techniques	31	[73],[74][7][82][83][40][84][85][52][50][26][31][27][32][33][34][35][86][53][87][88][89][71][90][58][91][55] [54][15][92][93]
3	Insufficiency of resources	5	[36][37][5][12][51]
4	Reducing human effort requirement	8	[38][17] [39][40][7][6][3] [94]
5	Reducing storage requirement/cost	6	[2], [18][8][55][56][95]
6	Enhancing functions of a specific area with text classification techniques	7	[96][1], [97][41][98][42], [99]
7	Improving all-round development	4	[43], [44], [57], [100]
8	Scarce researches on a specific area of Study	2	[45], [46]
9	Investigating effects of specific technique	8	[47], [48][101][102][103][79][95][16]