Computers in Human Behavior Reports 7 (2022) 100187

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Interest identification from browser tab titles: A systematic literature review

Mirko Farina [a], Maxim Kostin [a], Giancarlo Succi [b,*]

[a] Innopolis University, Universitetskaya ul., 1 Innopolis, 420500, Tatarstan, Russia
[b] Dipartimento di Informatica - Scienza e Ingegneria, Università di Bologna, Bologna, Italy

ARTICLE INFO

Keywords:
Software engineering
User interests identification
Non-verbal communication
Empirical studies
Systematic literature review

ABSTRACT

Modeling and understanding users interests has become an essential part of our daily lives.

A variety of business processes and a growing number of companies employ various tools to such an end. The outcomes of these identification strategies are beneficial for both companies and users: the former are more likely to offer services to those customers who really need them, while the latter are more likely to get the service they desire.

Several works have been carried out in the area of user interests identification. As a result, it might not be easy for researchers, developers, and users to orient themselves in the field; that is, to find the tools and methods that they most need, to identify ripe areas for further investigations, and to propose the development and adoption of new research plans.

In this study, to overcome these potential shortcomings, we performed a systematic literature review on user interests identification. We used as input data browsing tab titles. Our goal here is to offer a service to the readership, which is capable of systematically guiding and reliably orienting researchers, developers, and users in this very vast domain.

Our findings demonstrate that the majority of the research carried out in the field gathers data from either social networks (such as Twitter, Instagram and Facebook) or from search engines, leaving open the question of what to do when such data is not available.

1. Introduction

Modeling and understanding users interests has become an essential part of our daily lives (Kumar & Reinartz, 2012), becoming a crucial tool for marketing (Kim et al., 2010; Trusov et al., 2010), urban development (Ferdous, 2015), food services (Zeng et al., 2016), education (Robertson & Jones, 2009), use of social media (Zhang et al., 2021) and several other different domains or areas. A variety of business processes and a growing number of companies also employ user interests identification techniques on large scales (Mobasher, 2005; Montgomery et al., 2004; Wasim et al., 2011).

The outcomes of these identification strategies or procedures are typically beneficial for both companies and users: the former are more likely to offer services to those customers who really need them, while the latter are more likely to get the service they desire (Lilien & Rangaswamy, 2004).

Early attempts to individuate people's interests were carried out at the very beginning of the computer era -in the late sixties- and subsequently formalized for the first time by Hansen in 1971 (Hansen, 1972). This formalization played a very important role, as it allowed the field to evolve and to progress further. In particular, in the HCI community the very first work specifically devoted to classifying users was produced by Mac an Airchinnigh (Mac, 1982). This work originated a sequence of important followups and discussions, which were aptly summarised by Kofler in a landmark paper in 1986 (Koffler, 1986).

Given the intriguing and multifaceted research implications but also the strategic business relevance of this subject, studies on how to identify users' interests thrived. Research works focused on several different domains, including, e.g., spreadsheets (Bishop & McDaid, 2008), navigation systems (Wen et al., 2013), and social networks (Antelmi, 2019), but also considered locations of users (Preo ţ iuc-Pietro and Cohn, 2013) and preferred user interfaces (Yuan et al., 2020). At the same time, the approaches used for analysing users' behaviours also evolved. There are, nowadays, several different techniques based on a variety of approaches

* Corresponding author.
E-mail address: g.succi@unibo.it (G. Succi).

https://doi.org/10.1016/j.chbr.2022.100187
Received 13 July 2021; Received in revised form 26 February 2022; Accepted 1 March 2022
Available online 6 June 2022

(such as recommender systems (Hazrati, 2020), process analysis (Zhuang, 2017), analysis of web usage (Apaolaza, 2013), evaluating inputs to voice interfaces (Jung et al., 2020), and many more).

All of above shows that research on this area is indeed very active. However, it also demonstrates that current research is performed under extremely wide conditions in terms of breath, scope, and even reach. This can make it challenging for researchers, developers, and potential users to orient themselves and identify what they exactly need. For this reason, we think that a systematic literature review on a focused area could be extremely beneficial for the field.

In this study, we thus performed a systematic literature review on user interests identification, using as input browsing tab titles. Browsing activities (such as browsing tab titles) contain rich sources of information and are very telling of users' behaviours – they typically contain information about what people read and seek as well as about how they spend a significant part of their life (Ruiz et al., 2014). In this context it is worth noting that a recent study estimated that the world's population devotes (on average) at least 3 h to online activities per day (Johnson, 2021). Real online activity, though, is believed to be higher than 3 h per day, as the above-mentioned estimate included children, people in countries with limited internet access, and various other factors potentially affecting the estimate itself. Moreover, interest identification has become a topic of heated debates even in business magazines and specialized IT websites that have recently highlighted the importance of using browsing activities to understand the behaviour of end users – see for instance a very recent discussion followed by a panel on Forbes (Huntinghouse et al., 2021; Williams, 2021), or a discussion specifically on tabbed browsing on Criteo (Pruett, 2020).

In this research work we show that:

- the work we conducted is original, as we did not identify any other systematic literature review performed on the same topic in recent years, and
- our work has the potential to offer a service to our readership, as it provides a very valuable contribution towards orienting researchers, developers, and users in the vast amount of existing studies carried out on user interests identification. In particular, we claim that our work is helpful because it can allow researchers, developers, and users to: i. more easily locate the tools and methods they need, ii. identify the most promising open areas for future investigations, and iii. propose and develop new strategies for the implementation of such investigations.

Before we proceed any further, we would like to note that we decided to restrict our analysis to scientific journals and conference articles. This was done to ensure the integrity as well as the accuracy and reliability of our findings. However, since the topic of our investigation is huge, we needed to organise our work systematically and coherently. For this reason, we decided to perform a systematic literature review on the selected research topic (identifying user interests from browsing activities, such as browsing tab titles).

According to Fink (2019) (Fink, 2019), a Systematic Literature Review (SLR) is "a systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners". A SLR then offers to the reader a vast and comprehensive analysis of the research conducted in the field, while critically synthesizing research in ways that can be beneficial for a larger audience. SLRs are very important tools as they offer a baseline for scientific progress and for this reason are increasingly used by researchers (Brereton et al., 2007; Kitchenham et al., 2009).

This SLR is organized as follows. Section 2 describes the protocol used in this SLR. Section 3 presents the results of the SLR. Section 4 contextualises and critically discusses our findings. Section 5 offers an interpretation of our results in relation to the research questions characterising this systematic literature review. Section 6 evaluates potential issues with this study as well as various other potential threats to its validity, while Section 7 draws some conclusion and outlines plans for future research.

2. SLR protocol development

This review was performed in accordance with the protocol defined by Kitchenham (2004) (Kitchenham, 2004) (Fig. 1) and in conformity with the "Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020,"[1] which provides a guidance on organizing the review and helps reporting performed work and describing research findings (Brereton et al., 2007). The PRISMA checklist (Moher et al., 2009) is supplied in Table 13 as an appendix to this work.

2.1. Research questions

Any SLR starts out by individuating a series of research questions that guide and inform its development. This systematic literature review attempted to answer the following research questions:

RQ_1: What are existing approaches for the identification of users' interests?
RQ_2: What technologies, tools, and methods can be used to perform interests identification?
RQ_3: How effective and reliable are such tools and methods?
RQ_4: Is the information that we gather from the browser tab name sufficient to perform interests identification?

The motivation for 1 was to gain an understanding of what existing research has produced in terms of general approaches to user interests identification. In particular, we were interested in finding text-based identification approaches. Addressing this research question allowed us to understand what is the state of the art on this topic at the time of writing.

The motivation for 2 was to find out which methods and approaches to interest identification are the most popular in the relevant industry at the time of writing.

The motivation for 3 was to rank such methods and approaches in terms of effectiveness and reliability.

The motivation for 4 was to understand whether it would possible to apply approaches and methods, found in 1 and 2, to perform user identification when having only a small amount of data. Knowing this information is important because it would allow us to understand what is the smallest possible amount of data needed to identify users' interests. This, in turn, would be instrumental to decrease model training time, while preserving prediction quality.

2.2. Literature search process

This section specifies the search process (keywords, search queries, and databases used), as well as the inclusion and exclusion criteria adopted in our Systematic Literature Review. This section also contains an evaluation scheme for our findings and various other relevant methodological considerations.

2.2.1. Search resources

In order to individuate benchmark papers in the field, we conducted a manual search through different databases, using a set of potentially relevant keywords. The keywords selected for our initial searches were:

- interests identification
- text classification
- topic modeling

[1] http://www.prisma-statement.org.

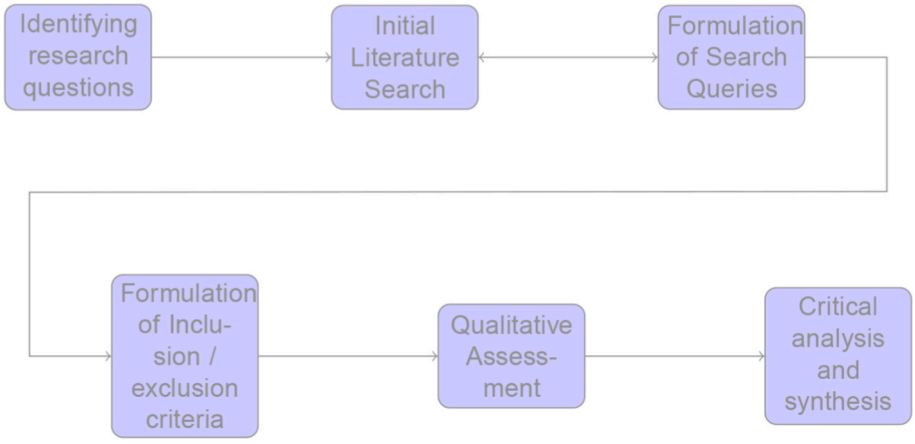

Fig. 1. Steps involved in a Systematic Literature Review, according to Kitchenham (2004) (Kitchenham, 2004).

- machine learning

The following digital repositories were screened for our initial searches:

- ACM Digital Library
- Arxiv.org
- Elsevier products (Scopus and ScienceDirect)
- Google Scholar
- IEEExplore Digital Library

We believe that the list of databases used for our searches, although ameliorable, is quite comprehensive, hence reliable and scientifically sound.

2.2.2. Search queries

Having collected a list of benchmark papers, as it is customary in any SLR, we generalised our keywords and determined key topics, which were useful to inform and guide our subsequent searches. The first key topic we individuated was "user interests identification via browser tab name". However, this key topic was too specific in its scope. In most cases in fact, the search results obtained by searching databases with it were closer or coincident to those found for 'user identification' or 'user classification'; hence, to broaden our searches, we decided to use a wider set of topics that fell into a similar category. In particular, we decided to focus not only on *users' interest identification*, but also on text and title classification. We decided to do so because there are a lot of studies related to topic identification on texts datasets (Hong & Davison, 2010; Hu et al., 2014; Wallach, 2006) and sentiment analysis on short texts datasets (Dos Santos & Gatti, 2014; Hassan & Mahmood, 2017; Kiritchenko et al., 2014; Wang et al., 2016a). The idea was therefore that data for user interests identification could be collected in the same way as for topic identification or sentiment analysis. Moreover, since this work is related to textual data, it seemed essential to understand which natural language processing (NLP) techniques were most appropriate for usage (Oshikawa et al., 1811; Otter et al., 2021; Qiu et al., 2020, pp. 1–26). This process allowed us to determined a final list of potentially relevant key topics. These included:

- user interests identification
- user personality classification
- text classification
- short text classification
- title classification
- NLP approaches

Further, we converted each of the aforementioned topics into proper (more structured) search queries. Since only unlabeled data was available, it was necessary to add a key word, *clustering*, to the search queries produced. As a result, all searches were performed using OR-combinations and AND-combinations of the following:

- (interest* **OR** user* interest*) **AND** (identification)
- (topic* **OR** key topic*) **AND** (identification)
- (user* **OR** user* personality) **AND** (classification **OR** clustering)
- (title **OR** short text **OR** text) **AND** (classification **OR** clustering)
- (text **OR** search queries **OR** short text) **AND** pre-processing
- (NLP **OR** natural language processing) **AND** (approache* **OR** method* **OR** algorithm*)

The queries were then edited to comply with the search capabilities of each system, but overall the semantics was retained. The search was performed in both titles and abstracts and the results of our query were stored in our reading log. The whole process is described in detail in subsection *'Search Results by Sources'* below.

2.3. Inclusion and exclusion criteria

Having performed our initial searches, we then formulated our inclusion and exclusion criteria (Patino & Ferreira, 2018). The formulation of inclusion/exclusion criteria is a necessary step for any systematic literature review (Kitchenham et al., 2009). Inclusion and exclusion criteria typically help identifying relevant studies for inclusion or filtering out improper or irrelevant sources. To determine which papers among those searched and initially gathered would be included in our review, we employed the following inclusion criteria:

- Papers found using search queries specified above;
- Papers related to at least two of the above-mentioned key topics;
- Papers written in English;
- Papers published on or after 2005. [This year was chosen as the starting point of our Systematic Literature Review as methods applied earlier than this year are mostly rule-based methods and their usage is nowadays deprecated];
- Peer-review papers indexed by reputable publishers (such as ACM Digital Library, Scopus, ScienceDirect, IEEExplore Digital Library)

The adoption of the criteria above-mentioned helped us obtaining a set of high-quality papers that were unlikely to be prone to specificity or biases. The inclusion of these papers in our literature log was instrumental to gain a reliable state of the art knowledge over the researched area.

In our Systematic Literature Review we also used the following exclusion criteria, which we report below for completeness.

- Those papers that did not match at least one of the inclusion criteria specified above were automatically excluded from the review;
- Duplicates found during the search process were also excluded from the review;
- We evaluated the first 150 papers listed on the databases for every search query we used, more on this below;
- We considered only peer-reviewed publications that appeared in either journals or conference proceedings. Hence, books and patents were excluded from this review. Peer-reviewed publications in journals and conference proceedings are notoriously harder to get, so we thought that considering only such publications could increase the quality of our systematic literature review (this is why we excluded patents and grey literature). There aren't many books published on the topic, so we preferred not to focus on them at all.

2.4. Search results by sources

In this subsection we specify for the reader the exact process (search queries used for each of the selected databases), that led to the preparation of our reading log, hence, to the extraction of relevant data. During the search process, in the attempt to streamline our findings and make our task easier, while preserving scientific soundness, we considered the first 150 results in each search. Naturally, a lot of the items initially considered were not relevant for this study and below we provide a detailed description of the process of inclusion/exclusion (data extraction) of the various items we initially gathered.

Searching through ACM DL:

1. By applying the selected queries we retrieved 900 papers, 28 of them were deemed as potentially relevant.
2. All papers were written in English without duplicates.
3. All papers were "archival publications" in scientific journals. As a consequence, neither grey literature nor non-peer reviewed sources appeared in the search.
4. 1 paper was excluded because it failed to meet the second inclusion criteria.
5. 13 papers were excluded because they failed to meet the fourth inclusion criteria.
6. 1 publication was excluded, since it was a book.
7. As a result, 13 papers were included into the log for further analysis.

Searching through IEEExplore Digital Library:

1. By applying the selected queries we retrieved 900 papers, 23 of them were deemed as potentially relevant.
2. All papers were written in English without duplicates.
3. All papers were "archival publications" in scientific journals. As a consequence, neither grey literature nor non-peer reviewed sources appeared in the search.
4. 2 papers were excluded because they failed to meet the second inclusion criteria.
5. 8 papers were excluded because they failed to meet the fourth inclusion criteria.
6. 2 publication were excluded, since their content was out of scope.
7. As a result, 11 papers were included into the log for further analysis.

Searching through Elsevier (Scopus and ScienceDirect):

1. By applying the selected queries we retrieved 1800 papers (900 for each database we considered), 28 of them were deemed as potentially relevant; however, we couldn't get access to seven of them, which were then automatically excluded.
2. 3 papers were not written in English and, as a consequence, were excluded.
3. All papers were "archival publications" in scientific journals. As a consequence, neither grey literature nor non-peer reviewed sources appeared in the search.
4. 2 papers were excluded because they failed to meet the second inclusion criteria.
5. 3 papers were excluded because they failed to meet the fourth inclusion criteria.
6. As a result, 13 papers were included into the log for further analysis.

Papers selected through Google Scholar and Arxiv.org:

1. Papers selected with the help of these two repositories were not searched by using any of the search queries specified above; rather, they were included after manual searches and a process of snowballing. These are the second most common ways of including relevant references in a review, according to Petersen et al. (2015) (Petersen et al., 2015).
2. At the end of the manual search on these two additional databases, 37 papers were retrieved. Five of them were excluded because they failed to meet the fifth inclusion criteria. Two more papers were excluded from Arxiv.org results because they were duplicate, and 6 papers were excluded from Google Scholar, because of the following reasons: out of scope, patent description, or duplicate. Consequently, 5 papers from Arxiv.org and 19 papers from Google Scholar were included in our reading log.

As a result of these detailed searches, we included in the reading log of our Systematic Literature Review a total of 61 papers. We note that we performed two rounds of searches for this SLR. The first was carried out between 08.2020 and 12.2020. The second was conducted between 10.2021 and 11.2021; that is, after we received feedback on our manuscript from the reviewers. We note that the results of this second search, which was more accurate and precise, were described in this section.

The PRISMA flow chart diagram shown in Fig. 2 below represents the process of inclusion/exclusion we described above, visually for the reader.

2.5. Quality assessment

Next, we generated a set of questions that we applied to each of the papers selected for inclusion. We then assigned to each question a score, so that each paper's quality could be objectively assessed. All researchers involved in the study actively participated in the process. This was done to: i. maximise perceived objectivity, and ii. minimize the possibility of subjective interpretations and/or mistakes.

If "yes" was given as an answer, then the paper would receive 1 point. If "partially" was given as an answer, then the paper would receive 0.5 points. If "no" was given as an answer, then the paper would receive 0 points.

Each paper's score was then computed so that an assessment about its quality could be performed. The set of questions we used to assess the papers' quality are listed below:

1. Were the objectives and the research questions clearly specified?
   - 1 point if the objectives and research questions were explicitly stated;
   - 0.5 points if the goals of the paper and its research questions were sufficiently clear but could be improved;
   - 0 points if no objectives were stated, if the research questions were hard to determine, or if they didn't relate to the research being carried out.
2. Were the objectives and the research questions presented satisfied?
   - 1 point if the research questions were answered precisely, or if the objectives were clearly satisfied;

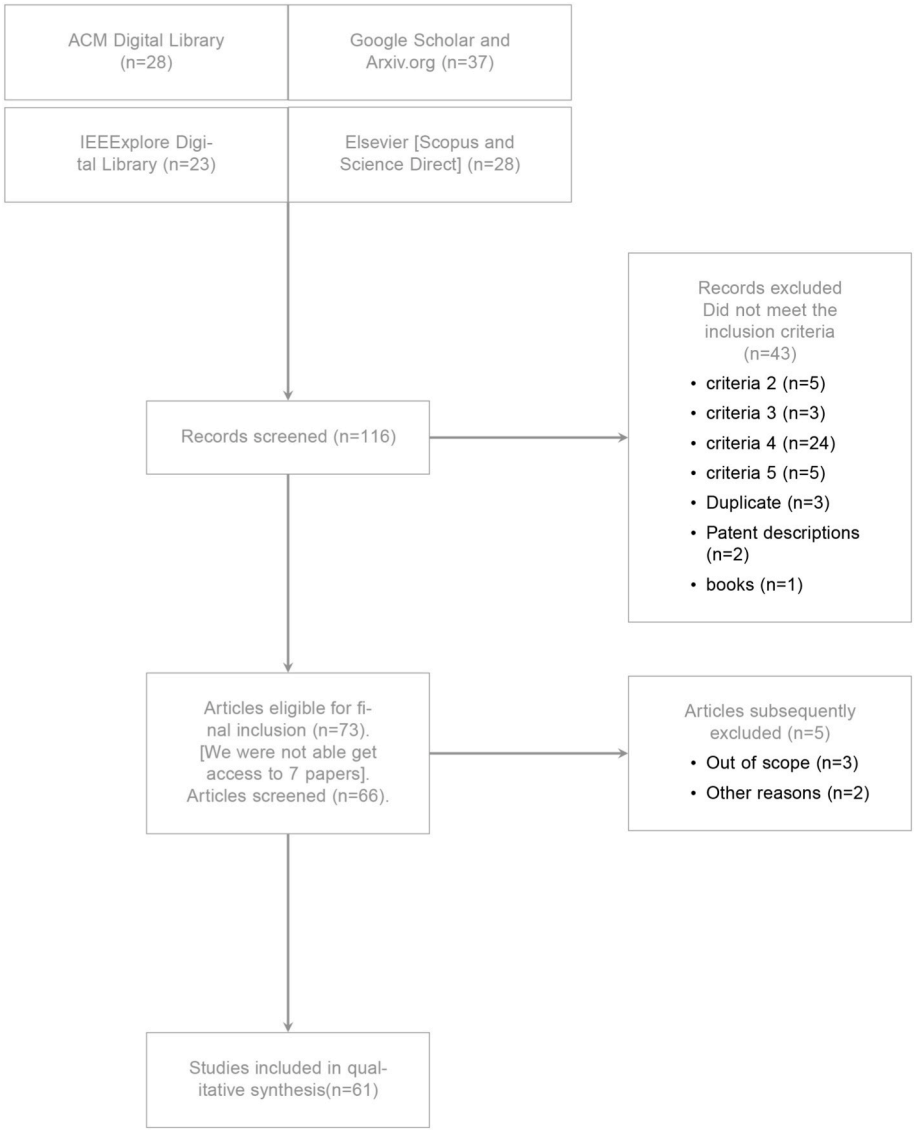

Fig. 2. Prisma flow chart diagram.

- 0.5 points if the research questions were partially answered, or if the objectives were partially achieved (with some deviations, for example);
- 0 points if the research questions remained unanswered, or if the objectives were unrelated to the research performed.

3. Was the research process transparent and reproducible?
   - 1 point if the paper specified the methodology as well as the technologies used and the data gathered, or if all the necessary steps and sources needed to reproduce the research were transparently available to the reader;
   - 0.5 points if minor details were lacking (for example, a dataset is not readily available);
   - 0 points if it was impossible to restore the sequence of actions, or if other critical details (such as an algorithm or technologies used) were missing.
4. Were the results evaluated critically and comprehensively?
   - 1 point if the authors of the paper provided a critical, balanced, and fair analysis of their results;
   - 0.5 points if the results were only partly (sufficiently) scrutinized and a comprehensive critical analysis was missing;
   - 0 points if the authors did not evaluate their results.
5. Was the conclusion sound? In other words, was it grounded on the results?
   - 1 point if the results provided evidence for the conclusion, or if the conclusion was logical and sound;
   - 0.5 points if the results could only partially justify the conclusion;
   - 0 points if the conclusion was overstated or if it couldn't be justified by the results presented in the paper.
6. Are there comparisons with alternatives?
   - 1 point if a comparison with other solutions is offered, with advantages and limitations clearly stated;
   - 0.5 points if the comparison was offered but it was not comprehensively discussed;
   - 0 points if no comparison was provided.

Thus, the total score assigned to each paper could range from 0 to 6. The results of this quality assessment procedure can be found on Tables 5 and 6 below. This process of quality assessment was performed on the 61 papers selected for inclusion in the reading log and it was instrumental to determine, evaluate, and further assess -as noted above- the overall quality of each of the papers we included in this study. As the tables below demonstrate, the vast majority of the papers included in our study were of high quality, as expected. Hence, we can infer that the

results we gathered from their analysis are scientifically sound.

However, the main goal of this Systematic Literature Review was not to simply evaluate the papers *per se*; rather, its aim was to collect and structure, in a meaningful way, information for future usage. For this reason, a paper questionnaire was compiled and then used as a structured tool to retrieve relevant information (such as the number of datasets and algorithms as well as the metrics used in the selected papers). The questionnaire included the following items or questions:

- What is the main idea of the study?
- What are the datasets used in the study?
- How was the datasets pre-processed?
- Which methods were used for clustering?
- What are the results of the applied methods and how were they evaluated?

The results of the questionnaire are available for consultation below. Please refer to Tables 8–11.

3. Results

3.1. Search sources overview

Table 1 shows the advantages and limitations of the databases used for the search process. This is included as a general tool for the reader to help her gauge and assess the potential benefits and limitations related with each database used.

Table 2 shows the distribution of papers by databases. See also Fig. 3 below for a graphically more appealing way representing this information.

We reiterate that all the articles included in our Systematic Literature Review were gathered from reputable journals and/or from proceedings of important conferences. A skeptical reader may at this point object that Fig. 3 above and Table 2 below are somehow misleading, because some papers could be found across different databases. As noted in section 2.3 above, we would like to emphasise here that we automatically excluded duplicates from our searches. Thus, in Fig. 3 above and Table 2 below, we only represented papers, which we attributed to a single repository. The attribution was subjective in character and yet followed the following rule. The paper was attributed to one database (e.g. Google Scholar) rather than to the other (e.g., Scopus) based on the chronological order of the searches we performed. We think that while this is perhaps not an ideal practice, it does not constitute a significant problem for this work. We thank the reviewer for pressing us on this important issue.

Table 1
Databases' advantages and limitations.

Database	Advantages	Limitations
1. ACM Digital Library	A source of peer-reviewed papers in IT	Paywall
2. Arxiv.org[a]	Open access, large collection of papers	Papers are very likely to be not peer-reviewed, sometimes of poor quality. However, some authors do upload their studies here after they have been peer-reviewed or published in other journals. For this reason, this database is worth checking
3.Elsevier (Scopus and ScienceDirect)	Scientific databases with effective and precise search tools and analytic statistics	Most of its journals as well as search features are behind paywall
4. Google Scholar	This database is helpful in finding open access articles. It provides papers' citations data, which is a useful indicator of an article's popularity within a given scientific community.	Some papers (those not peer-reviewed) can be problematic, a few more are behind a paywall
5. IEEExplore Digital Library	Published papers are all peer-reviewed by leading experts in the field	Paywall

[a] Relevant studies, initially retrieved in Arxiv.org, were further checked to ensure adherence to the peer-review criterion we considered for inclusion. Please refer to subsection 2.3 above.

Table 2
Papers distribution by databases.

Source	Quantity	Percentage
ACM Digital Library	13	21.3
Elsevier products (Scopus and ScienceDirect)	13	21.3
IEEExplore Digital Library	11	18.0
Arxiv.org	5	8.2
Google scholar	19	31.2
Total	61	100

3.2. Studies classification

This section presents the statistics related to the quality and content of the papers included in the systematic literature review. This analysis is instrumental for getting a better understanding of how the selected studied could be further classified or clustered.

Five major topics were identified during the review process (Table 3). It is worth noting in this context that the following topics: i. "interests identification", ii. "classification based on text" and iii. "classification based on short text" have emerged as 'hot topics' in the last 15 years or so. This may be due to the growth of services such as Twitter and Instagram (Arora et al., 2019), coupled with the change in the model of content consumption (the overall focus has moved from long texts to short texts) by users (Anger & Kittl, 2011), (Shi et al., 2014).

Fig. 4 shows the distribution of key topics by years of publication.

Fig. 5 below shows the distribution of papers by years of publication. The same data is reproduced in Table 4, where -for simplicity-the distribution is shown over a period of 5-years.

The number of relevant works increased by 3 times in the second period (2010–2014) and almost doubled in the third period (2015-now) (Table 4). The overall quality metrics of research, described in 2.5, have noticeably improved over the years (Table 6). Table 5 highlights the distribution of papers by quality scores ranges.

Table 7, shows the geographical distribution of the papers included in our reading log (we took the first author's affiliation as an indicator for this measure).We note that in most cases the affiliation coincided with the author's nationality. However, in a few cases those two differed.(Fig. 6) presents the information contained in (Table 7) graphically. We note a predominance of papers from China and The USA; however, many other countries are represented. This confirms the cross-cultural significance of our work.

With respect to the contents of the studies included, eight training model datasets were found to be most frequently used (Table 8, Table 9). Frequency in usage is for a dataset to have been mentioned in more than four of the papers included in the present review. Based on this criterion, Twitter can be assumed to be the most popular one. However, datasets belonging to the category "Other", which have been mentioned in less than three papers, constitute about 42.6% of the total number of datasets. This may well illustrate the diversity of datasets in use in the field.

Several datasets can be found in one paper, and one dataset can be found in several papers. The number of datasets occurrences were therefore summed up. We obtained 94 occurrences in total for both text classification datasets and short text classification datasets. Please refer to Table 8 and Table 9. The percentages displayed on the tables below are obtained by dividing the number of occurrences for each individual

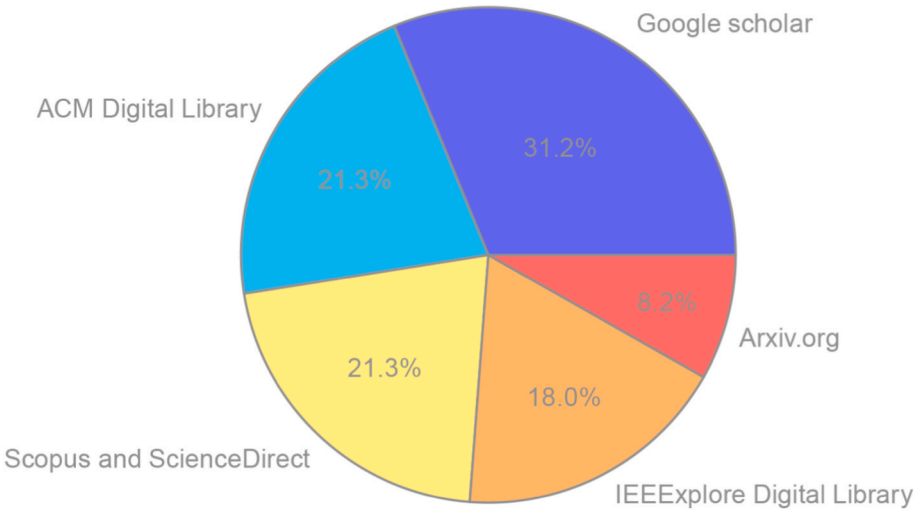

Fig. 3. Papers distribution by databases - [visually more appealing representation].

Table 3
Key topics distribution.

Topic	The earliest publication date	The latest publication date	Number of papers	Share of papers
User interests identification	2006	2017	3	4.9
User personality classification	2012	2017	4	6.6
Key topic identification based on text	2005	2020	23	37.7
Sentiment classification based on short text	2007	2020	23	37.7
NLP approaches	2014	2019	8	13.1

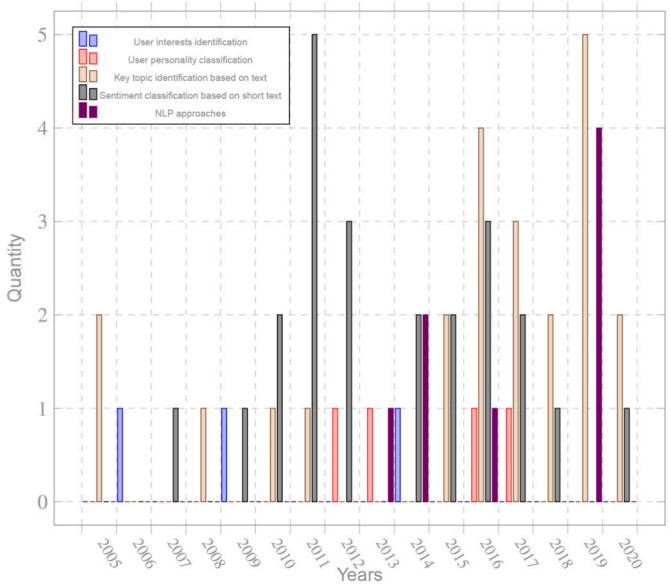

Fig. 4. Key topic distribution by year.

dataset by the combined number of occurrences displayed on both tables (94).

Table 10 lists the algorithms used for training models and the statistics of their usage. We observed that NN (Neural Network) is the most frequently mentioned algorithm. We concluded that it can probably be considered as the most widely used in the field.

The discrepancy between the total number of papers included in the

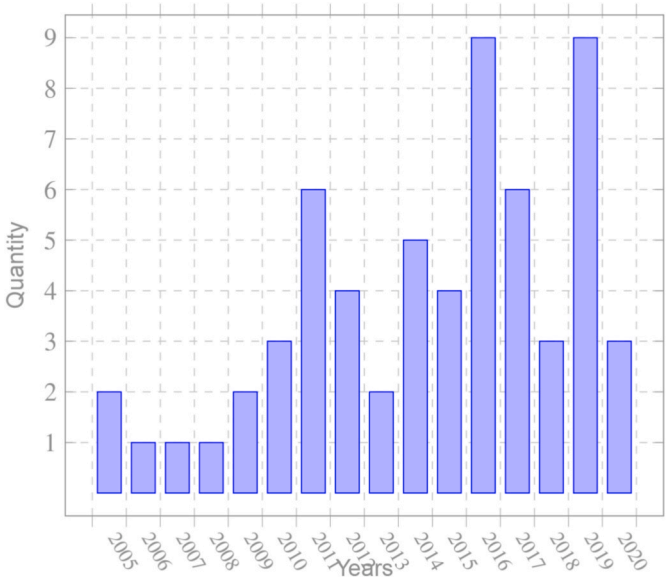

Fig. 5. Paper distribution by years.

Table 4
Papers distribution over a 5-years period.

Years	Quantity	Percentage
2005–2009	7	11.5
2010–2014	20	32.8
2015-now	34	55.7

Table 5
Quality assessment - statistics.

QA score	Quantity	Percentage
0–3.5	8	13.1
4–5	21	34.4
5.5–6	32	52.5

Table 6
Quality mean values by year.

Years	Quantity	Average Quality
2005–2009	7	4.4
2010–2014	20	4.9
2015-now	34	5.3

Table 7
Papers distribution by location.

Country	Quantity	Percentage
USA	23	37.7
China	22	36.1
Australia	3	4.9
India	3	4.9
UK	2	3.3
Austria	1	1.6
Germany	1	1.6
Greece	1	1.6
France	1	1.6
Jordan	1	1.6
Indonesia	1	1.6
Singapore	1	1.6
Sweden	1	1.6
Total	61	100

Table 8
Datasets mentioned in papers, used for text classification and clustering.

Dataset	Quantity	Percentage	Number of items	Number of classes
Google Snippet dataset	8	8.5	> 1M	> 8
AG's news corpus	6	6.4	> 1M	4
Reuters-21578	5	5.3	11367	82
Sogou news corpus.	5	5.3	2909551	5
DBPedia ontology dataset	5	5.3	4.23M	685
Others	23	24.5	> 10k	> 2

Table 9
Datasets mentioned in papers, used for short text classification and clustering.

Dataset	Quantity	Percentage	Number of items	Number of classes
Twitter[a]	15	16	> 10k	> 2
TREC	5	5.3	5500	6
Yahoo! Answers dataset	5	5.3	4483032	10
Others	17	18.1	> 10k	> 2

[a] The number of elements and classes varies from study to study and depends on the selected dataset and the purpose of the study.

Table 10
Algorithms used in the papers included.

Algorithms	Quantity	Percentage
Neural network (NN)	13	38.2
Latent Dirichlet Allocation (LDA)	7	20.6
K-means	5	14.7
Support Vector Machine (SVM)	5	14.7
K-nearest neighbors (k-NN)	4	11.8

SLR (61) and the number displayed in Table 10 (34) is due to the fact that some papers present comparisons of several algorithms (Curiskis et al., 2020; Rafeeque & Sendhilkumar, 2011), others introduce new algorithms (Yang et al., 2019), or use the same algorithm. This means that each paper does not necessarily discuss one algorithm.

Table 11
Metrics used in the papers included.

Metric	Number of mentions	Percentage
Accuracy	26	34.2
F_1-score	23	30.3
Mutual information (MI)	11	14.5
Topic coherence	5	6.6
Recall and precision	4	5.2
Others	7	9.2
Total	76	100

Table 11 and Fig. 7 display the statistics relative to the efficiency metrics mentioned in the papers selected for inclusion. The metrics are generally used to evaluate the quality of the prediction in a quantitative way, thus allowing us to compare different methods.

The metrics' quantities in Table 11 do not coincide with the total number of papers selected for inclusion in the SLR. The reason for that is that some papers did not contain proper metrics evaluations and others opted to use *several metrics simultaneously* to take into account different aspects of prediction.

"Others" in Table 11 refers to the metrics that were used in less than three papers. These are:

- area under ROC curve (AUCROC),
- entropy and purity,
- Kullback-Leibler (KL) divergence,
- error rate.

Having presented our results, analyzed and explained their statistical significance, in the next section of our systematic literature review, we go to contextualise them. In other words, we analyse and synthesise them critically.

4. Discussion

4.1. Available datasets used in selected papers

We observed that "Twitter," "TREC," and "Yahoo! Answers dataset" are the most used datasets in short text classification domains (Table 9). This might be due to the fact that there are a lot of data with short texts or these datasets have enough data and are labeled really well. In terms of text classification, "Google snippets dataset," "AG's news corpus," "Reuters-21578," "Sogou news corpus," and "DBPedia ontology dataset" are the most used (Table 8). News sites, as well as encyclopedia entries, are commonly used for text classification tasks because they offer a broad set of topics and (or) an abundance of texts.

Other datasets mentioned in the papers we selected for this study are "Amazon reviews" (Zhang & LeCun, 2015, pp. 1–9) and "Yelp reviews" (Zhang & LeCun, 2015, pp. 1–9); and "20 Newsgroups" (Lai et al., 2015; Yao et al., 2019; Zhong, 2005) dataset.

Sometimes researchers used private datasets (Yu et al., 2012) or collected data manually; for instance, by crawling internet resources (Banerjee et al., 2007; Jin et al., 2011; Yin & Wang, 2014).

In some cases researchers appeared to combine both approaches, e.g., (Wang et al., 2014). That is, the researchers created a dataset for training and used real queries from a Bing query flow over the course of 5 h as the source of the test dataset.

4.2. Text preprocessing and vectorization techniques

Before we move on to explain algorithms, we feel we should spend a few words discussing text pre-processing and text vectorization techniques, as analyzed in the papers we selected.

Pre-processing simply means that before fitting the relevant text to the specific model, the text should prepared. The most popular techniques used for this task are:

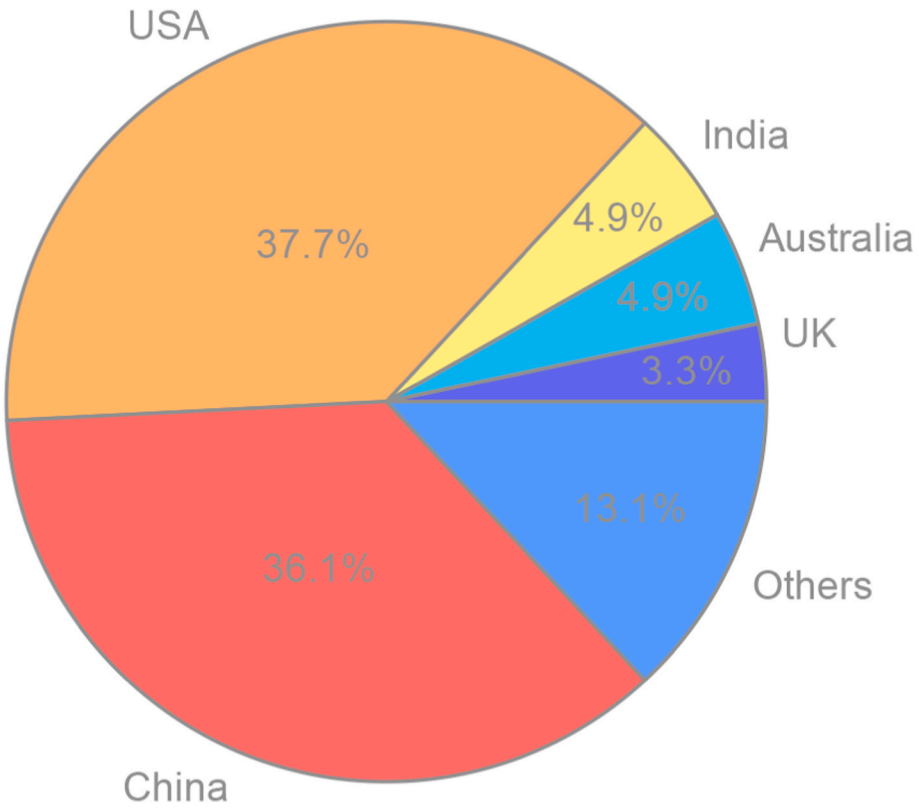

Fig. 6. Papers distribution by location - [visually more appealing representation].

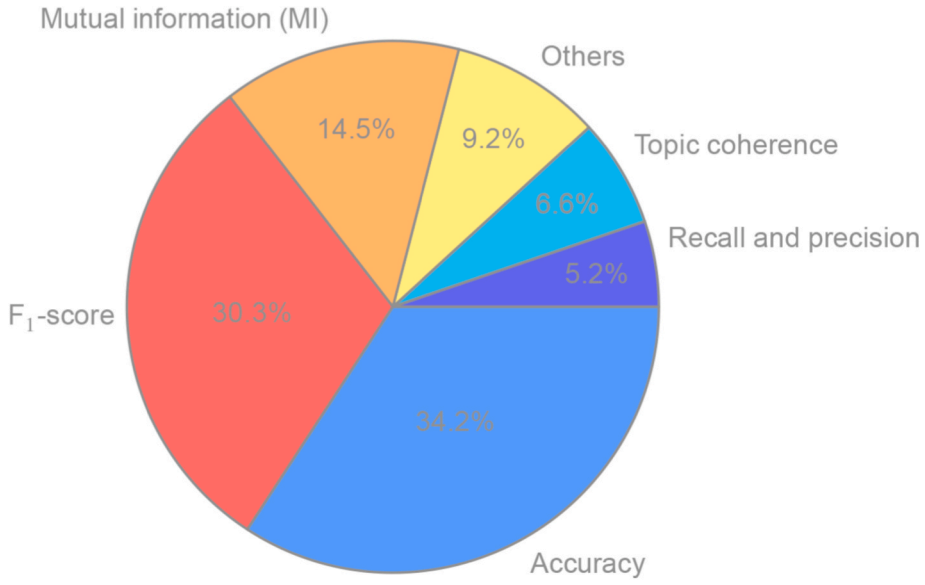

Fig. 7. Metrics used in the papers included.

- tokenization
- lowercasing
- stop-word elimination
- stemming
- lemming

We are aware that the list above is not an exhaustive one; however, we note that it can work as a good approximation. Moreover, applying these techniques may either significantly increase performance, as is shown in (Duwairi & El-Orfali, 2014; Yu et al., 2012), or dramatically decrease it (Wang et al., 2014, 2015, 2016b; Xu et al., 2017; Yin & Wang, 2014), as techniques' efficiency significantly vary from research to research and strongly depends on data and on its quality. ***Text pre-processing*** is reported to have been the first step of protocol in the majority of papers related to text classification, this step is usually followed by ***text vectorization***. The most popular techniques for text vectorization are:

- Bag of Words (BoW),
- Term frequency

- Inverse document frequency (TF-IDF), and
- word embedding.

However, we note that applying the same vectorization methods to short texts can give completely different results, often leading to considerably worse performance. Data sparsity is a well-known limitation for both TF-IDF (Wang et al., 2016b; Xu et al., 2017; Yin & Wang, 2014) and BoW (Wang et al., 2014, 2015), and it only deteriorates when these methods are applied to short texts.

This limitation is the reason behind the creation of new approaches to processing short text.

Such approaches include those based on expanding and enriching the context of data from Wikipedia (Banerjee et al., 2007; Hu et al., 2008), as well as those that aim to enrich short text with its translation (Tang et al., 2012).

In (Wang et al., 2014) the authors suggested to replace BoW with a "Bag of Concepts" (BoC) approach. In particular, researchers defined a *concept* as a set or a class of entities within a domain such that words belonging to similar classes get similar representations.

For the same task, Lee and Dernoncourt (Lee & Dernoncourt, 2016) suggested using either recurrent neural network (RNN) or convolutional neural network (CNN) to generate a vector representation for each short text, whereas Wang and Zhiguo in (Wang et al., 2016b) suggested CNN or long short-term memory (LSTM) models for the same purpose.

Clearly, as we have just seen, there are plenty of approaches to text vectorization. Among all these approaches, however, BoW and TF-IDF remain the most popular ones in the field to date.

4.3. Classification algorithms

We observed that the most popular algorithm used for training models is neural network (NN) (Table 10). However, our literature analysis revealed that this approach has become widely used only as recently as 2015. Before NN became dominant, several other algorithms were in use. For example, the simplest approach for text classification is using k-means algorithms (including k-means, k-means++ and fuzzy c-means). K-means algorithms were used in (Li & Zhang, 2007; Tang et al., 2012; Xu et al., 2017) as a main algorithm for classification, along with different pre-processing techniques. There are two main advantages in using this algorithm. These are: i. speed of convergence, and ii. simplicity. However, there are also significant limitations and shortcomings, including iii. sensitivity to the choice of initial centers, and iv. the requirement to know the number of classes *a priori*. This is why the usage of such algorithms gradually diminished over the years.

Another approach frequently mentioned in the selected literature is LDA (Latent Dirichlet allocation) (Blei et al., 2003). It was used in (Chen et al., 2011; Xiong et al., 2018; Zhang & Zhong, 2016) as a main algorithm to derive latent topics from short text. This method has the same advantages as the k-means algorithm, which we specified above. Its main limitation, however, especially important for the purpose of our study, is that LDA noticeably under-performs with short texts. This is due to the probabilistic nature of the model and to sparsity problem of the data. In order to overcome this problem the authors of (Zuo et al., 2016) suggested to utilize a Pseudo-document-based Topic Model (PTM) approach. PTM introduces the concept of pseudo document to implicitly aggregate short texts against data sparsity. By modeling the topic distributions of latent pseudo documents rather than short texts, PTM can gain excellent performance in both accuracy and efficiency.

Yet, two other algorithms mentioned in the literature, SVM (Support vector machine) and K-NN (k-nearest neighbors algorithm), were used for various purposes (Duwairi & El-Orfali, 2014; Javed et al., 2015; Zhang et al., 2019a). SVM works relatively well if the data is linearly separable; however, if it the data is not linearly separable - SVM will use Kernel Tricks to map the data in higher dimension and try to separate data in this dimension. Another frequently used algorithm is k-NN. One characteristic of this algorithm is that it does not learn any formulas or discriminative functions during the training period. However, it also has a major drawback - its setup forces the algorithm to store all the training data in order to make predictions. As a consequence, its performance (the speed of prediction) dramatically decreases with large or high-dimensional datasets.

In (Lee & Dernoncourt, 2016), the authors suggested performing text vectorization using CNN or RNN and feeding the result to another NN for class prediction. Using NN for text vectorization gives some degrees of freedom in choosing input's dimensionality with respect to the classifier. However, there is no universal solution for text vectorization, and in most cases researchers attempt to find an optimal solution empirically (via trials and errors on numerous experiments).

Researchers in (Tian & Fang, 2019), utilized another type of NN - Attention-based Autoencoder for short text topic modeling task. They also used this model for topic inference. The authors argued that attention mechanisms enhance topic coherence, by focusing on salient content of each document.

Another application of NN autoencoder architecture is described in (Yu et al., 2015). The authors of this study proposed the development of a deep neural network for hashing the semantics of an enriched short text.

In (Rashid et al., 2019), the authors suggested the application of a fuzzy c-means algorithm - called Fuzzy topic modeling (FTM). In particular, they used principal component analysis (PCA) to remove the high-dimensionality negative impact on global term weighting, and subsequently deployed their fuzzy c-means algorithm.

Another recently discovered method is described in (Yao et al., 2019), where the authors suggested using a Text Graph Convolutional Network (Text GCN) for text classification. The researchers claim that this method can achieve strong classification performance with a small proportion of labeled documents and can learn interpretable word and document node embeddings.

To conclude, there are a number of ways and approaches to text classification and clustering. Nowadays, NN can be considered a state-of-the-art approach being the one that is most widely used by researchers. The main advantage of NN is that in most of the cases it outperforms other approaches. Its major limitation is that it requires a lot of trials to succeed because none knows exactly, at the time of writing at least, the reason for this superior performance. This means that one needs to try a lot of different combinations of hyperparameters and different preprocessing techniques in order for the model to achieve optimal performance.

4.4. Quality assessment metrics

In the evaluation metrics, accuracy is the prevailing metrics considered (Fig. 7). It was used in (Banerjee et al., 2007; Bekkerman & Gavish, 2011; Camacho-Collados & Pilehvar, 2019, pp. 40–46; Chen et al., 2011; Duwairi & El-Orfali, 2014; Hu et al., 2018; Javed et al., 2015; Joulin et al., 2016, 2017, pp. 1–13; Lai et al., 2015; Lee & Dernoncourt, 2016; Liu et al., 2005; Ma et al., 2016; Sriram, 2010; Sun, 2012; Wang et al., 2014, 2015, 2016b, 2016c, 2017; Xu et al., 2017; Yao et al., 2019; Yi et al., 2020; Yu et al., 2012; Zhan & Dahal, 2017; Zhang et al., 2019a, 2019b). Its main disadvantage is the fact that it does not take into account how the data are actually distributed, hence it is completely unreliable if classes are imbalanced. Therefore, for this metric, the data should be carefully adapted.

Fortunately, in most of the reviewed papers we included in our SLR, this metric was used in combination with another, more representative metrics. One of those additional metrics often used in combination with accuracy is F_1-score, whose mentions score is almost equivalent to that of the accuracy's. This metric though, eliminates the main limitation of accuracy and leads to more precise and reliable results. In addition, there is a micro-averaging variation of this metric that may help to understand data in more precise ways.

Another metric mentioned in the literature is MI, in particular

Adjusted MI (AMI), which was used to compare different clustering results in (Wang et al., 2016b). In (Jin et al., 2011; Zhong, 2005) the authors used a Normalized MI (NMI), which allowed them to balance the quality of the clustering against the number of clusters individuated.

Based on the statistical analysis performed above, it seems reasonable and appropriate to summarize the information obtained and to conduct a critical review of our research questions, which is what we will do next.

5. Critical review of our research questions

The analysis conducted on this systematic literature review has revealed that there is only a limited number of papers related to user interests identification in our domain (Table 3). On the one hand, this outcome may indicate a promising research gap, hence it may highlight the novelty this study can bring about in the field. On the other hand though, this fact raises a number of pressing concerns. Why have the research attempts in this field been so rare? Was this research effort hard to implement or, maybe, there was no demand for this kind of system? At the time of writing, we do not have a definitive answer to these latter questions; nevertheless, the need for further research on this topic has been established, and the groundwork for future progress in the field has been laid down in this study.

5.1. Analysis of 1: What are the existing approaches for the identification of users' interests?

We identified a number of papers related to user interests identification. Of them, three papers can be said to represent the main direction of research in this area: (Kapanipathi et al., 2014; Qiu & Cho, 2006; White et al., 2009). In two of these papers the researchers used similar approaches to identify user's interests (more on these approaches below). Other papers were mostly focused on applying well-known classifications or clustering algorithms on different language families.

In (Qiu & Cho, 2006), the researchers attempted to automate user interests identification to make search results more relevant. Using past clicks history, the researchers created a topic preference vector as a representation of users' interests. Based on this vector and topic-sensitive page ranking, the researchers tried to find more relevant pages for users.

In (White et al., 2009), the authors relied on a similar idea and also utilized click history. However, they classified pages in this click history based on the page context. For page classification, the researchers used open directory project (ODP); as a result, users' interests were represented as a list of ODP category labels. These category labels were ranked in a descending order, based on each label's frequency in the given context. The main idea of this research was to predict future users' interests and to increase search results' relevance.

In (Kapanipathi et al., 2014), a different approach was employed. In particular, the researchers used tweets to create a hierarchical interest graph (HIG). They utilized wikipedia category graph (WCG) to create an HIG. From the HIG, the researchers were able not only to extract users' interests but also to suggest similar interests. The researchers report to have been able to retrieve up to 76% relevant results from the top-10 interests.

5.2. Analysis of 2: What technologies, tools and methods can be used to perform interests identification?

The number of methods described for interest identification is quite limited. Older methods (Li et al., 2007; Liu et al., 2010; Obendorf et al., 2007) are, as we have seen above, based on utilizing click history. In newer methods researchers also apply clustering methods (Duan et al., 2014; Liang et al., 2017, 2018; Qiu & Shen, 2017; Tang & Zeng, 2012). In particular, a number of different methods for text classification and clustering were identified in the literature, and the most popular are:

- K-nearest neighbors (k-NN) (Trstenjak et al., 2014),
- K-means (Singh et al., 2011),
- LDA (Pavlinek & Podgorelec, 2017),
- SVM (Wang et al., 2006),
- Naïve Bayes Classifier (Dai et al., 2007),
- NN (including CNN, RNN, Recursive NN) (Zhou et al., 1511; Lai et al., 2015; Liu et al., 1605).

Most of these methods have already been implemented in software, often as libraries of existing systems, such as (Pedregosa et al., 2011; Gulli & Pal, 2017; G é ron, 2019; Virtanen et al., 2020).

5.3. Analysis of 3: How effective and reliable are such methods?

In (Qiu & Cho, 2006), the researchers reported 0.3 relative error, which implies that their model can learn user's topic preferences with 70% accuracy.

In (White et al., 2009), the researchers used different models, and reported a number of measurements; in particular F_1-score varied from 0.4 to 0.75 for different models and topics.

In (Kapanipathi et al., 2014), the researchers achieved 0.88 (or 0.92, assuming "maybe" is a relevant answer) mean average precision (mAP) for one of their models.

As we can see quality of interests identification increased over the years as well as new and more accurate methods for evaluation become used.

5.4. Analysis of 4: Is the information that we gather from the browser tab name sufficient to perform interests identification?

Presently, a clear answer to this research question has not been found. There are, at least, three possible reasons for such a result: i. we selected a too narrow time interval, and, as a consequence some studies were left outside of our searches (this seems to be unlikely as our searches were comprehensive); ii. the topic is not of interest to researchers; iii. our research contributed to identify an important research gap in the literature, which we could contribute to fill in future works. In any case, whichever of these interpretations is correct, one could nevertheless probably claim that "short text clustering" is highly correlated to the topic of user interests identification and therefore that we could use information gathered on this specific topic (please see section 4) for the purpose of answering this research question.

5.5. A synoptic summary

Table 12 presents a synoptic summary of our results with respect to each of the research question tackled in this study.

There are, we found, three main types of approaches or methods to user interests identification 5.1. The first draws on click history to create a user profile (Qiu & Cho, 2006); this approach has long been an established practice in big companies, where it is used as a part of

Table 12
Synoptic Summary of our Results.

RQs	Summary
1	User interests identification has been tested and applied by researchers and with approaches similar to those for text classification.
2	There is a limited number of methods for interests identification; however, a lot of methods exist for text classification and clustering. The most popular ones are: K-NN, K-means, LDA, SVM, Naïve Bayes Classifier, NN (including CNN, RNN, Recursive NN).
3	It depends strongly on the data and on the algorithms used. Researchers in (Qiu & Cho, 2006) reported 70% accuracy; in (Kapanipathi et al., 2014), researchers achieved 0.88 mAP score, while researchers in (White et al., 2009) reported F_1-score from 0.4 to 0.75.
4	Presently, no convincing answer was identified in the literature with respect to our fourth research question.

browser history and used profiling. A more recent version of this approach (the second method) uses not only click history but also the context of users' environment as well as the contents of pages that were clicked in this click history (White et al., 2009). Finally, the third approach attempts to mine interests from social media networks such as Twitter, Facebook, and Instagram (Kapanipathi et al., 2014). All these approaches have respective strengths and weaknesses. Most importantly, all are related to specific key topics, i.e, identification from text, and moreover, some of these methods use the same techniques to identify users' interests.

With respect to methods based on 5.2 we identified in the literature two potentially applicable approaches for this purpose. The first uses an ontology (Kapanipathi et al., 2014) database to create a hierarchical category graph and further employs this graph to identify users' interests. The second (Banerjee et al., 2007; Yan et al., 2012) instead mainly relies on applying machine learning clustering techniques.

6. Limitations, threats to validity, and review assessment

In any scientific work, there is always room for improvement (Bird, 2007), (Robinson, 2000). In this section of our SLR, we would like to take a more critical stance of our findings and assess them objectively. We thus want to reflect more carefully about things that could have impacted on the impartiality, accuracy, and completeness of our study. The analysis of potential shortcomings, affecting our study is vital not only to identify potential problems with our work; but also -and perhaps more importantly-to better comprehend future research possibilities, as well as to encourage our readers to think more critically about the subject we investigated.

6.1. Limitations

We start by reviewing situations and elements that may restrict or affect, thus possibly alter, the validity of our research. The limited amount of data gathered and consequently the limited amount of data we analyzed, we feel, is the study's principal weakness in this regard. In particular, we can consider four factors as potentially, but not necessarily, problematic for our work:

1. the first potential limitation we mention has to do with the fact that we took into consideration only the first 150 results for each search. We are confident thought, that focusing on the first 150 results for each search was enough to guarantee a comprehensive and systematic review of the current literature, as the majority of search engines sort results by significance and credibility (h-index, number of citations, impact factor etc);. As a matter of fact, all the papers that make up our final log (61) were found within the top 100 results for each search query. No relevant paper was found in the interval between 100 and 150. This suggests that relevant and credible sources, at least for this topic, either appear in the top 100 results or do not appear at all.
2. the second potential limitation affecting our work lies in the fact that we searched a limited number of databases (six essentially). However, it can be argued that database searches usually overlap with each other. In addition, it is worth noting that one of the databases we searched (Google Scholar) combines all available databases together;
3. the third potential limitation affecting our study concerns the usage of grey literature. Sometimes in this field, grey literature does contain state-of-the-art methods and approaches. So, there may be some merit in occasionally including grey literature in a Systematic Literature Review (Mahood et al., 2014), (Garousi et al., 2016). However, this type of literature is hard to assess and manage, and even harder to validate. For this reason, in our study, we mostly focused on secondary sources (as customary) and on reviewing other Systematic Literature Reviews, disregarding grey literature;
4. One could perhaps argue that one of the inclusion criteria adopted in our work (selecting only papers written in English) is severely constraining the kind of research we were able to get. Issues in cross-cultural research are rightly emerging as vitally important in science (Henrich et al., 2010). We are fully aware of these issues and acknowledge this as a very serious point of contention. However, we also note that much of the literature in the field is in English, so the requirement we adopted is not an unusual one for the literature.

6.2. Threats to validity

We now consider potential threats (or biases) that might call into question the credibility or the overall validity of the conclusions presented in our systematic literature review. According to (Akl et al., 2019), there are, at least, seven types of different biases, which may affecting any piece of research: **a)** publication bias; **b)** time lag bias; **c)** multiple (duplicate) publication bias; **d)** location bias; **e)** citation bias; **f)** language bias; and **g)** outcome reporting bias.

After a critical analysis performed on our work, we believe that only one of such biases may have potentially affected our research. This is:

1. the **language bias** – As noted above, we considered only papers written in English. Thus, in our searches we might have neglected relevant studies published in other languages. We acknowledge this potential issue; however, we also note that this is a remote possibility because most of the literature is nowadays published in English.

6.3. Review assessment

As a final step in the critical assessment of our findings, we decided to answer four benchmark questions, which can be generally used to assess the overall quality of a systematic literature review (Kitchenham, 2004):

1. *Are the review's inclusion and exclusion criteria described and appropriate?* All criteria used for inclusion or exclusion were clearly mentioned upfront in our protocol. All the criteria are also reasonable and relevant; hence we believe that they are appropriate for the study.
2. *Is the literature search likely to have covered all relevant studies?* As we already mentioned in Section 6.1, there were certain limitations that prevented us from collecting all the possible data available. However, the process we establish to verify our protocol and the methodology we used to check our results was very through, sound, and comprehensive. We are thus confident that our work is scientifically sound and accurate.
3. *Did the reviewers assess the quality/validity of the included studies?* We count this condition as sufficiently met for two reasons: i. we included in our Systematic Literature Review only credible papers published in reputable journals or venues, and ii. we formulated a set of questions (see subsection 2.6 'quality assessment') to determine the quality of the papers we included.
4. *Were the basic data/studies adequately described?* This condition is certainly met because we built a comprehensive reading log to put all the relevant information extracted from the papers we selected. This allowed us to process our data systematically and comprehensively.

7. Conclusion

The aim of the present study was to investigate matters related to user interests identification, which is one of the crucial tasks for business nowadays. The study attempted to give an answer to the following research questions: i. (1) what are existing approaches for the identification of users' interests?, ii. (2) what technologies, tools, and methods can be used to perform interests identification?; iii. (3) how effective and reliable are such tools and methods? and iv. (4) is the information that we gather from the browser tab name sufficient to perform interests

identification?

With respect to **1**, we found that user interests identification has been studied by many researchers, often with approaches similar to those used in text classification. Concerning **2** our analysis demonstrated the existence of a rather limited number of methods for interests identification; however, we note that several methods exist for text classification and clustering. The most popular ones are: K-NN, K-means, LDA, SVM, Naïve Bayes Classifier, NN (including CNN, RNN, Recursive NN). With respect to **3** we can state that effectiveness in identification is strongly dependent on the data and on the algorithms used. Researchers in (Qiu & Cho, 2006) reported 70% accuracy; in (Kapanipathi et al., 2014), achieved 0.88 mAP score, while in (White et al., 2009) reported F_1-score between 0.4 and 0.75. Finally, the analysis of **4** paradigmatically illustrated the need of more research in the area, as we were unable to find any research evidence related to utilizing tab names for user interests identification purposes.This nevertheless highlights a potential positive contribution of this study for the literature; the discovery of a new research gap, which we hope we will contribute to fill in future works.

However, our results may also suggest that interests identification by short text might be closely related to text clustering; this is probable because the approaches and methods are similar. In any case, further work is needed to corroborate any of these speculations. We nevertheless hope that this systematic literature review will broaden interest in user interests identification and will provide new grounds for more detailed explorations into these extraordinarily rich and fascinating set of phenomena.

Authorship statement

The authors contributed equally to the writing of this paper.

Declaration of competing interest

The authors declare that there is no conflict of interest.

Acknowledgments

We thank the anonymous reviewers for helpful comments and extremely valuable feedback, which helped us significantly improving this manuscript.

Appendix

Table 13
PRISMA 2020 Checklist Template taken from: http://www.prisma-statement.org/documents/PRISMA_2020_checklist.docx.

Section and Topic	Item #	Location where item is reported
TITLE		
Title	1	1
ABSTRACT		
Abstract	2	2
INTRODUCTION		
Rationale	3	1
Objectives	4	2.1
METHODS		
Eligibility criteria	5	2.3
Information sources	6	2.2.1
Search strategy	7	2.5
Selection process	8	-
Data collection process	9	2.4
Data items	10a	3.2
	10b	-
Study risk of bias assessment	11	6.3
Effect measure	12	-
Synthesis methods	13a	2.5
	13b	-
	13c	-
	13d	2.5
	13e	-
	13f	-
Reporting bias assessment	14	6.2
Certainty assessment	15	6.3
RESULTS		
Study selection	16a	2.4
Study selection	16b	2.4
Study characteristics	17	-
Risk of bias in studies	18	6
Results of individual studies	19	4, 5
Results of syntheses	20a	5
	20b	-
	20c	-
	20d	-
Reporting biases	21	6.2
Certainty of evidence	22	6.3
DISCUSSION		
Discussion	23a	4
	23b	6.1
	23c	6.1
	23d	4, 5
OTHER INFORMATION		

(*continued on next page*)

Table 13 (*continued*)

Section and Topic	Item #	Location where item is reported
Registration and protocol	24a	-
	24b	2
	24c	-
Support	25	-
Competing interests	26	-
Availability of data, code and other materials	27	-

References

Akl, E., Altman, D., Aluko, P., Askie, L., Beaton, D., Berlin, J., Bhaumik, B., Bingham, C., Boers, M., Booth, A., Boutron, I., Brennan, S., Briel, M., Briscoe, S., Busse, J., Caldwell, D., Cargo, M., Carrasco-Labra, A., Chaimani, A., & Young, C. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Anger, I., & Kittl, C. (2011). Measuring influence on twitter. In *Proceedings of the 11th international conference on knowledge management and knowledge technologies* (pp. 1–4).

Antelmi, A. (2019). Towards an exhaustive framework for online social networks user behaviour modelling. In *Proceedings of the 27th ACM conference on user modeling, adaptation and personalization, UMAP '19* (pp. 349–352). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3320435.3323466.

Apaolaza, A. (2013). Identifying emergent behaviours from longitudinal web use. In *Proceedings of the adjunct publication of the 26th annual ACM symposium on user interface software and technology, UIST '13 adjunct* (pp. 53–56). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2508468.2508475.

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index-insights from facebook, twitter and instagram. *Journal of Retailing and Consumer Services, 49*, 86–101.

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'07* (pp. 787–788).

Bekkerman, R., & Gavish, M. (2011). High-precision phrase-based document classification on a modern scale. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 231–239).

Bird, A. (2007). What is scientific progress? *Noûs, 41*(1), 64–89.

Bishop, B., & McDaid, K. (2008). Spreadsheet debugging behaviour of expert and novice end-users. In *Proceedings of the 4th international workshop on end-user software engineering* (pp. 56–60). New York, NY, USA: WEUSE '08, Association for Computing Machinery. https://doi.org/10.1145/1370847.1370860.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software, 80*(4), 571–583.

Camacho-Collados, J., & Pilehvar, M. T. (2019). *On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis*. arXiv preprint arXiv:1707.01780.

Chen, M., Jin, X., & Shen, D. (2011). Short text classification improved by learning multi-granularity topics. In *IJCAI international joint conference on artificial intelligence* (pp. 1776–1781).

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management, 57*(2), Article 102034.

Dai, W., Xue, G.-R., Yang, Q., & Yu, Y. (2007). Transferring naive bayes classifiers for text classification. In *AAAI* (Vol. 7, pp. 540–545).

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 69–78).

Duan, J., Zeng, J., & Luo, B. (2014). Identification of opinion leaders based on user clustering and sentiment analysis. In *2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)* (Vol. 1, pp. 377–383). IEEE.

Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science, 40*(4), 501–513.

Ferdous, F. (2015). If you could know what users think: Urban design and preference of the visual attributes to design sustainable urban open spaces. *Journal of Advanced in Humanities, 3*(1), 143–151.

Fink, A. (2019). *Conducting research literature reviews: From the internet to paper*. Sage publications.

Garousi, V., Felderer, M., & Mäntylä, M. V. (2016). The need for multivocal literature reviews in software engineering: Complementing systematic literature reviews with grey literature. In *Proceedings of the 20th international conference on evaluation and assessment in software engineering* (pp. 1–6).

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gulli, A., & Pal, S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.

Hansen, W. J. (1972). User engineering principles for interactive systems. In *Proceedings of the November 16-18, 1971, Fall joint computer conference, AFIPS '71 (Fall)* (pp. 523–532). New York, NY, USA: Association for Computing Machinery.

Hassan, A., & Mahmood, A. (2017). Deep learning approach for sentiment analysis of short texts. In *2017 3rd international conference on control, automation and robotics (ICCAR)* (pp. 705–710). IEEE.

Hazrati, N. (2020). Recommender systems effect on user' choice behaviour. In *Proceedings of the 25th international conference on intelligent user interfaces companion, IUI '20* (pp. 21–22). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3379336.3381505.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80–88).

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning, 95*(3), 423–469.

Hu, J., Fang, L., Cao, Y., Zeng, H. J., Li, H., Yang, Q., & Chen, Z. (2008). In *Enhancing text clustering by leveraging wikipedia semantics, ACM SIGIR 2008 - 31st annual international ACM SIGIR conference on research and development in information retrieval, proceedings* (pp. 179–186).

Huntinghouse, J., Raj, A., Koduvayur, R., Falcon, S., Davis, A., Dawson, A., Burroughs, E., Hawke, N., Kushmaro, P., Malhotra, A., Sarkar-Basu, B., Parna, Selchau-Hansen, C., Bringé, A., Mendes-Roter, J.a., Cox, A., & Bender, E.. Council post: 16 critical things to get straight before creating a digital marketing strategy – panel. site visited on 1st October 2021. (Jul 2021). URL https://www.forbes.com/sites/forbescommunicationscouncil/2021/07/12/16-critical-things-to-get-straight-before-creating-a-digital-marketing-strategy/.

Hu, X., Wang, H., & Li, P. (2018). Online biterm topic model based short text stream classification using short text expansion and concept drifting detection. *Pattern Recognition Letters, 116*, 187–194.

Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., & Kang, T. S. (2015). Carotene: A job title classification system for the online recruitment domain. In *Proceedings - 2015 IEEE 1st international conference on big data computing service and applications* (pp. 286–293). BigDataService, 2015.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., & Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering, International Conference on Information and Knowledge Management. *Proceedings*, 775–784.

Johnson, J.. Daily time spent online by device 2021. site visited on the 21st April 2021. (Jan 2021) https://www.statista.com/statistics/319732/daily-time-spent-online-device/.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing text classification models*. arXiv preprint arXiv:1612.03651.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification, 15th conference of the European chapter of the association for computational linguistics. In *EACL 2017 - proceedings of conference* (Vol. 2, pp. 427–431).

Jung, H., Kim, H., & Ha, J.-W. (2020). Understanding differences between heavy users and light users in difficulties with voice user interfaces. In *Proceedings of the 2nd conference on conversational user interfaces, CUI '20*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3405755.3406170.

Kapanipathi, P., Jain, P., Venkataramani, C., & Sheth, A. (2014). User interests identification on twitter using a hierarchical knowledge base. In *European semantic web conference* (pp. 99–113). Springer.

Kim, D.-J., Chung, K.-W., & Hong, K.-S. (2010). Person authentication using face, teeth and voice modalities for mobile device security. *IEEE Transactions on Consumer Electronics, 56*(4), 2678–2685.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research, 50*, 723–762.

Kitchenham, B. (2004). *Procedures for performing systematic reviews* (Vol. 33, pp. 1–26). Keele, UK: Keele University, 2004.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and Software Technology, 51*(1), 7–15.

Koffler, R. P. (1986). Classifying users: A hard look at some controversial issues. *SIGCHI Bull, 18*(2), 75–76.

Kumar, V., & Reinartz, W. (2012). *Customer relationship management: Concept, strategy, and tools*. Springer Texts in Business and Economics, Springer. URL https://books.google.ru/books?id=wBLYtNotoE0C.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the National Conference on Artificial Intelligence, 3*, 2267–2273.

Lee, J. Y., & Dernoncourt, F. (2016). In *Sequential short-text classification with recurrent and convolutional neural networks, 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, NAACL HLT 2016 - proceedings of the conference* (pp. 515–520).

Liang, S., Ren, Z., Yilmaz, E., & Kanoulas, E. (2017). Collaborative user clustering for short text streams. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31).

Liang, S., Yilmaz, E., & Kanoulas, E. (2018). Collaboratively tracking interests for user clustering in streams of short texts. *IEEE Transactions on Knowledge and Data Engineering, 31*(2), 257–272.

Lilien, G. L., & Rangaswamy, A. (2004). *Marketing engineering: Computer-assisted marketing analysis and planning.* DecisionPro.

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 31–40).

Liu, L., Kang, J., Yu, J., & Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. In *Proceedings of 2005 IEEE international conference on natural language processing and knowledge engineering* (pp. 597–601). IEEE NLP-KE'05, 2005.

P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, arXiv preprint arXiv:1605.05101.

Li, L., Yang, Z., Wang, B., & Kitsuregawa, M. (2007). Dynamic adaptation strategies for long-term and short-term user profile to personalize search. In *Advances in data and web management* (pp. 228–240). Springer.

Li, B., & Zhang, J. (2007). Enhancing clustering blog documents by utilizing author/reader comments. In *Proceedings of the 45th annual southeast regional conference* (pp. 94–99).

Mac, M. (1982). An Airchinnigh, classifying the user. *SIGCHI Bull, 14*(2), 3–8.

Mahood, Q., Van Eerd, D., & Irvin, E. (2014). Searching for grey literature for systematic reviews: Challenges and benefits. *Research Synthesis Methods, 5*(3), 221–234.

Ma, C., Zhao, Q., Pan, J., & Yan, Y. (2016). Short text classification based on distributional representations of words. *IEICE - Transactions on Info and Systems, 99*(10), 2562–2565.

Mobasher, B. (2005). Web usage mining. In *Encyclopedia of data warehousing and mining* (pp. 1216–1220). IGI Global.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS Medicine, 6*(7), Article e1000097.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science, 23*(4), 579–595.

Obendorf, H., Weinreich, H., Herder, E., & Mayer, M. (2007). Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 597–606).

R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, arXiv preprint arXiv:1811.00770.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems, 32*(2).

Patino, C., & Ferreira, J. (2018). Inclusion and exclusion criteria in research studies: Definitions and why they matter. *Jornal Brasileiro de Pneumologia, 44*, 84.

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and lda topic models. *Expert Systems with Applications, 80*, 83–93.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology, 64*, 1–18.

Preoţiuc-Pietro, D., & Cohn, T. (2013). Mining user behaviours: A study of check-in patterns in location based social networks. In *Proceedings of the 5th annual ACM web science conference* (pp. 306–315). New York, NY, USA: WebSci '13, Association for Computing Machinery. https://doi.org/10.1145/2464464.2464479.

Pruett, M.. What marketers need to know about multi-tab mentality in 2020. site visited on 1st October 2021. (Sep 2020). URL https://www.criteo.com/blog/multi-tab-mentality/.

Qiu, F., & Cho, J. (2006). Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on world wide web* (pp. 727–736).

Qiu, Z., & Shen, H. (2017). User clustering in a dynamic social network topic model for short text streams. *Information Sciences, 414*, 102–116.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). *Pre-trained models for natural language processing: A survey*. Science China Technological Sciences.

Rafeeque, P., & Sendhilkumar, S. (2011). A survey on short text analysis in web. In *2011 third international conference on advanced computing* (pp. 365–371). IEEE.

Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management, 56*(6), Article 102060.

Robertson, M. J., & Jones, J. G. (2009). Exploring academic library users' preferences of delivery methods for library instruction: Webpage, digital game, and other modalities. *Reference and User Services Quarterly, 48*(3), 259–269.

Robinson, D. N. (2000). Paradigms and the myth of framework' how science progresses. *Theory & Psychology, 10*(1), 39–47.

Ruiz, N., Amatte, F., Kil, J., & Brandini, P. (2014). Opening the "private browsing" data - acquiring evidence of browsing activities. In *Proceedings of the international conference on information security and cyber Forensics*.

Shi, Z., Rui, H., & Whinston, A. B. (2014). Content sharing in a social broadcasting environment: Evidence from twitter. *MIS Quarterly, 38*(1), 123–142.

Singh, V. K., Tiwari, N., & Garg, S. (2011). Document clustering using k-means, heuristic k-means and fuzzy c-means. In *2011 international conference on computational intelligence and communication networks* (pp. 297–301). IEEE.

Sriram, B. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841–842).

Sun, A. (2012). Short text classification using very few words. In *SIGIR'12 - proceedings of the international ACM SIGIR conference on research and development in information retrieval* (pp. 1145–1146).

Tang, J., Wang, X., Gao, H., Hu, X., & Liu, H. (2012). Enriching short text representation in microblog for clustering. *Frontiers of Computer Science in China, 6*(1), 88–101.

Tang, X., & Zeng, Q. (2012). Keyword clustering for user interest profiling refinement within paper recommender systems. *Journal of Systems and Software, 85*(1), 87–101.

Tian, T., & Fang, Z. F. (2019). Attention-based autoencoder topic model for short texts. *Procedia Computer Science, 151*, 1134–1139.

Trstenjak, B., Mikac, S., & Donko, D. (2014). Knn with tf-idf based framework for text categorization. *Procedia Engineering, 69*, 1356–1364.

Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining influential users in internet social networks. *Journal of Marketing Research, 47*(4), 643–658.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods, 17*(3), 261–272.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977–984).

Wang, X., Jiang, W., & Luo, Z. (2016a). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2428–2437).

Wang, Z., Mi, H., & Ittycheriah, A. (2016b). In *Semi-supervised clustering for short text via deep representation learning, CoNLL 2016 - 20th SIGNLL conference on computational natural language learning, proceedings* (pp. 31–39).

Wang, Z.-Q., Sun, X., Zhang, D.-X., & Li, X. (2006). An optimal svm-based text classification algorithm. In *2006 international conference on machine learning and cybernetics* (pp. 1378–1381). IEEE.

Wang, F., Wang, Z., & Li, Z. (2014). Concept-based short text classification and ranking categories and subject descriptors. *Acm*, 1069–1078.

Wang, J., Wang, Z., Zhang, D., & Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI international joint conference on artificial intelligence* (Vol. 350, pp. 2915–2921).

Wang, P., Xu, J., Xu, B., Liu, C. L., Zhang, H., Wang, F., & Hao, H. (2015). In *Semantic clustering and convolutional neural network for short text categorization, ACL-IJCNLP 2015 - 53rd annual meeting of the association for computational linguistics and 7th international joint conference on natural language processing of the Asian Federation of natural language processing, proceedings of the conference - short pape* (pp. 352–357).

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.-L., & Hao, H. (2016c). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing, 174*, 806–814.

Wasim, M., Shahzadi, I., Ahmad, Q., & Mahmood, W. (2011). Extracting and modeling user interests based on social media. In *2011 IEEE 14th international multitopic conference* (pp. 284–289). IEEE.

Wen, J., Helton, W. S., & Billinghurst, M. (2013). Classifying users of mobile pedestrian navigation tools. In *Proceedings of the 25th Australian computer-human interaction conference: Augmentation, application, innovation, collaboration, OzCHI '13* (pp. 13–16). New York, NY, USA: Association for Computing Machinery.

White, R. W., Bailey, P., & Chen, L. (2009). Predicting user interests from contextual information. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 363–370).

Williams, Z.. How businesses can market in A third-party cookieless world. site visited on 1st October 2021. (Apr 2021). URL https://www.forbes.com/sites/forbesagencycouncil/2021/04/09/how-businesses-can-market-in-a-third-party-cookieless-world/.

Xiong, S., Wang, K., Ji, D., & Wang, B. (2018). A short text sentiment-topic model for product reviews. *Neurocomputing, 297*, 94–102.

Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., Zhao, J., & Xu, B. (2017). Self-Taught convolutional neural networks for short text clustering. *Neural Networks, 88*, 22–31.

Yang, S., Huang, G., & Cai, B. (2019). Discovering topic representative terms for short text clustering. *IEEE Access, 7*, 92037–92047.

Yan, X., Guo, J., Liu, S., Cheng, X. Q., & Wang, Y. (2012). Clustering short text using Ncut-weighted non-negative matrix factorization. In *ACM international conference proceeding series* (pp. 2259–2262).

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 7370–7377.

Yi, F., Jiang, B., & Wu, J. (2020). Topic modeling for short texts via word embedding and document correlation. *IEEE Access, 8*, 30692–30705.

Yin, J., & Wang, J. (2014). A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 233–242).

Yuan, S., Brüggemeier, B., Hillmann, S., & Michael, T. (2020). User preference and categories for error responses in conversational user interfaces. In *Proceedings of the 2nd conference on conversational user interfaces, CUI '20*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3405755.3406126.

Yu, H., Ho, C., Arunachalam, P., Somaiya, M., & Lin, C. (2012). Product title classification versus text classification. *Csie.Ntu.Edu.Tw*, 1–25.

Yu, Z., Wang, H., Lin, X., & Wang, M. (2015). Understanding short texts through semantic enrichment and hashing. *IEEE Transactions on Knowledge and Data Engineering, 28*(2), 566–579.

Zeng, J., Li, F., Liu, H., Wen, J., & Hirokawa, S. (2016). A restaurant recommender system based on user preference and location in mobile environment. In *2016 5th IIAI international congress on advanced applied informatics* (pp. 55–60). IIAI-AAI). https://doi.org/10.1109/IIAI-AAI.2016.126.

Zhan, J., & Dahal, B. (2017). Using deep learning for short text understanding. *Journal of Big Data, 4*(1), 1–15.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019a). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction, 99*(December 2018), 238–248.

Zhang, X., & LeCun, Y. (2015). *Text understanding from scratch*. arXiv preprint arXiv: 1502.01710.

Zhang, H., Ni, W., Zhao, M., & Lin, Z. (2019b). Cluster-gated convolutional neural network for short text classification. In *Proceedings of the 23rd conference on computational natural language learning* (pp. 1002–1011). CoNLL).

Zhang, F., Tang, J., Liu, X., Hou, Z., Dong, Y., Zhang, J., Liu, X., Xie, R., Zhuang, K., Zhang, X., Lin, L., & Yu, P. (2021). Understanding WeChat user preferences and "Wow" diffusion. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2021.3064233, 1–1.

Zhang, H., & Zhong, G. (2016). Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems, 102*, 76–86.

Zhong, S. (2005). Efficient streaming text clustering. *Neural Networks, 18*(5-6), 790–798.

C. Zhou, C. Sun, Z. Liu, F. Lau, A c-lstm neural network for text classification, arXiv preprint arXiv:1511.08630.

Zhuang, M. (2017). Modelling user behaviour based on process. In *Proceedings of the 25th conference on user modeling, adaptation and personalization, UMAP '17* (pp. 343–346). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3079628.3079705.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2105–2114).