University of Sindh Journal of Information and Communication Technology
(USJICT)
Volume 6, Issue 2, July 2022

ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

A Taxonomy of Text Mining

Huma Gul[1], Sadaqat Jan[1], Ibrar Ali Shah[1], Shams Ur Rehman[2]

Department of Computer Software Engineering[1]
Department of Computer Science[2], University of Engineering and Technology Mardan Khyber Pakhtunkhwa, Pakistan
E-mail:humawahab2012@gmail.com, sadaqat@uetmardan.edu.pk, ibrar@uetmardan.edu.pk

***Abstract:* With a rapid increase in the volume of textual data on the Internet, extracting useful information through innovative text mining techniques has become crucial. In this context, terminology jargon in the literature related to text-mining creates ambiguity and has made it very difficult for researchers to focus in a specific direction and bring innovation. For example, review mining and opinion mining may have different applications, however, from a technical perspective, they are very similar. In this paper, we propose a classification of the text mining terminologies from the perspectives of technical and text-mining processes. The classification is based on a comprehensive literature survey and analysis. This research study presents a clear classification of text mining terminologies based on technical and text mining processes to resolve the issue of terminology jargon. By utilizing the proposed classification, researchers will be able to easily choose a specific direction instead of diverging amongst similar research problems, thereby, driving innovation. Further, the proposed classification will help advance and improve the overall research progress in all text-mining related fields.**

Keywords: Text mining, Opinion Mining, Sentiment Analysis, Topic Modeling, Text Summarization, Review Mining

I. Introduction

In recent years, Artificial Intelligence (AI) in general, and Natural Language Processing (NLP) technique in particular, have seen tremendous advancement. Likewise, the volume of textual data on the Internet has grown exponentially and it has become challenging to process all types of text manually. Therefore, the field of text mining has emerged in NLP.

Text Mining is an automated technique of revealing and detecting patterns and inter-relationships in unstructured textual data. The main aim of text mining is to find un-discovered information in bulk amounts of data. Text mining methods are used for various purposes, such as opinion mining, sentiment analysis, review mining, text summarization, topic modeling etc.[1][2][3]. In opinion mining process, people's opinion are extracted from the web. It evaluates people's emotions, opinions and appraisals toward persons, actions, issues, topics, and its features[3]. Sentiment analysis involves the classification and determination of feelings or opinions conveyed in text or even the intensity of emotions[5]. Review mining is the process of extracting opinions from subjective contents and product features to evaluate user's feedbacks, sentiments, opinions etc. and summarizing them using a visual representation. Text summarization is a process used to generate a summary, i.e., brief and precise representation of the contents of a lengthy text document[5]. The summary is generated through an automatic text summarization system, i.e., a shortened form of the document, which covers a few important sentences [7][8]. Readers can get the main idea through a summary without reading the whole document or text by using this technology [8]. Topic modeling involves extracting concealed thematic ideas from a group of documents and represent them as few feasible topics. Topics are unlabeled in an unsupervised approach where words with the highest probability are chosen to represent in a topic.

All these applications of text mining follow similar basic text pre-processing[10]. Key tasks in pre-processing are removal of stop words - pronouns and prepositions, tokenization – splitting the sentences into small tokens/words, stemming – the process to convert different grammatical forms of words into its original form of word where suitable stop words are identified and removed. Review preprocessing tasks involved Part-of-speech (POS) tagging, POS filtering, stop-word elimination, and sentence selection[11].

However, when text mining is used in different application domains, different terminologies are often used for the same basic text mining technique, which result in a terminology jargon, thereby, causing confusion for new researchers and making it difficult for them to focus in a specific direction. For example, text summarization and text extraction techniques are very similar, but some researchers label them as text summarization while others as text extraction. Similarly, review and opinion mining are both used for similar term in the field of text data mining that classify opinion's polarity as being positive or negative and opinions are extracted from the texts. Furthermore, reviews and opinions can be considered two different terms based on their purpose and application in human language. However, text mining processes are mainly similar for review and opinion mining and technically these terms can be considered more similar as compared to the rest of terminologies like sentiment analysis or topic modeling. Such ambiguities in the literature create more distractions and stagnant the research progress, especially for the new researcher in the field.

To resolve these ambiguities, a clear taxonomy of the text mining is needed. Taxonomies help categorize and organize large volumes of research literature and techniques because the classification of objects helps researchers and practitioners to understand and analyze complex domains[12].

This paper aims to present a clear classification of all these terminologies and is based on technical and text-mining processes. To this end, we conduct a detailed literature survey and perform classification based on the surveyed literature. The results are then analyzed from different technical perspectives. Thus, this paper yields the following three-fold benefits:

i. It presents a clear classification of the text mining terminologies based on technical and text mining processes instead of applications.
ii. Based on these classifications, researchers can choose a specific direction instead of diverging amongst similar research problems.
iii. It will lead to the advancement and improvement of the overall research in all text-mining related fields.

The rest of this paper is organized as follow: Sections II is about the literature review, whereas Section III presents detail of classification methodology. In Section IV, experimental results and analysis are discussed in detail. Section V concludes the paper.

II. Literature Review

The role of taxonomies in management and research is vital since it helps practitioners and researchers to get a clear hint about the analysis of complex domains through objects' classification. In Information Systems (IS) research, it has been well recognized and established that taxonomies play a vital role by providing organization and structure to the knowledge base. This helps the practitioners and researchers to understand the relationships between different concepts. Various fields, e.g., information sciences, computational linguistics, computer science and statistics etc. have great influence on text mining. Due to its significance in various and diverse fields, text mining terminologies and features, when used in one field, has a totally different meaning in another field. Similarly, it is also possible that a term used in one field is not known in the other field. In [11], the authors have established the absence of a feature framework which can help practitioners and researchers in inter-discipline features' comparison and evaluation. Therefore, the focus of the current literature is mostly specific domain related features' comparison.

In this study, a comprehensive literature review was conducted by searching various computer science databases such as Association of Computing Machine Machinery (ACM)'s digital library, Institute of Electronics Engineering (IEEE)'s digital database, Science Direct etc. We selected only those research papers which were about these text mining techniques: opinion mining, sentiment analysis, review mining topic modeling and text summarization. Moreover, inclusion-exclusion criteria were used to filter out irrelevant publications. Further detail is given in the Section III.

III. Research Methodology

We have conducted Systematic Literature Review (SLR) which is a methodical way used for investigation of a specific research question through identification, assessing and analyzing published primary sources. SLR spans over three phases. The first phase is planning, which includes setting classification criteria, selection of data source, and selecting the search string(s). The second phase is execution which is conducting the review and the third phase is reporting the review by analyzing the results. The role of SLR is to investigate, evaluate, and interpret all potential research studies related to specific research questions, interests, or themes. We follow the SLR in Text Mining processes as suggested by Kitchenham which has three phases, i.e., planning, execution and results' analysis [13].

A. Planning

Classification Criteria: In this research, different text mining techniques will be classified to distinguish among the various available terms like Opinion mining, Sentiment Analysis, Review mining, Topic modeling, Text Summarization. In this study, these terms are distinguished from each other based on some parameters. Classification will be performed with respect to:

i. Preprocessing steps
ii. Domain
iii. Attributes along with Tools

Data Sources: Credible research database and sources have been used to select relevant literature, such as, IEEEXplore, ACM, ScienceDirect, Google Scholar and Springer.

Search String: The search queries single and in combinations, have been used to search the most relevant literature to answer the research question such as ("Opinion mining" OR "Sentiment Analysis" OR "Review Mining" OR "Topic Modeling" OR "Text Summarization").

In this study, different terminologies, as outlined in the search string, were searched one by one and the initial results were recorded it in the Table I according.

B. *Execution*

Literature was gathered through searches of major design, engineering, and computer science databases (e.g., ACM Digital Library, Technology Research Database) conducted between the years 2018 to 2020.

Inclusion criteria: For a research papers to be included in our study, it must:

i. Be published in 2018 and onwards.
ii. Be written in the English language.
iii. Report original research.
iv. Be Available in an online archive or database.
v. Constitute a full article (no editorials, book reviews, or calls for papers).
vi. Focus on one of the options (Text mining, Opinion mining, Sentiment analysis, Review mining, Topic Modeling, Text summarization).
vii. Contain at least one of the search word in the title of Paper (Text mining, Opinion mining, Sentiment analysis, Review mining, Topic Modeling, Text summarization).

Exclusion criteria: Exclude survey papers and those papers which are published before 2018. Results after applying the exclusion criteria are given in Table II.

TABLE I. INITIAL SEARCH RESULT FOR EVERY INDIVIDUAL TERM

Search Term/Technique	**Total results returned**
Text Mining	523000
Opinion Mining	169000
Sentiment analysis	284000
Review mining	812000
Text Summarization	408000
Topic Modeling	363000

TABLE II. SELECTED PAPERS AFTER APPLYING INCLUSION-EXCLUSION CRITERIA

Search Term/Technique	**Selected Papers**	**Nos. of Papers Published in 2018,2019,2020**
Text Mining	223	113
Opinion Mining	204	79
Sentiment Analysis	200	87
Review Mining	206	133
Text Summarization	220	117
Topic Modeling	209	105

IV. RESULTS AND ANALYSIS

As discussed in Section III, we performed categorization of the selected text mining techniques based on different criteria: pre-processing steps, domain, attributes along with various tools and techniques

A. *Categorization with respect to preprocessing steps*

Results of categorization on the basis of pre-processing steps are presented in Fig 1 to 6. As discussed earlier, the features extraction plays an important role in text mining process[15][16]. The difference in the number for every technique shows that all areas are different from one another when these techniques are applied in a particular application/use[17][18]. The use of Case Transformation is high in papers where topic modeling was focused and less in review mining related papers as shown in Fig.1. In topic modeling, the words from the whole text are considered and sometimes nouns and verbs are also identified, whereas, few techniques have been reported in the literature where keywords are directly collected for opinion and review mining.

Similarly, tokenization is mostly used in text summarization and topic modeling, while less in review mining. As other results of this study also suggest that text summarization and topic modeling are very similar in terms of many preprocessing techniques, in Fig.2 tokenization technique also advocate the same. In topic modeling papers, stop word removal is widely used as shown in Fig.3. The reason is that for assigning a topic to a paragraph, these words are useless and give no meaning. Stop word removal is mostly used in topic modeling papers while less used in review mining papers.

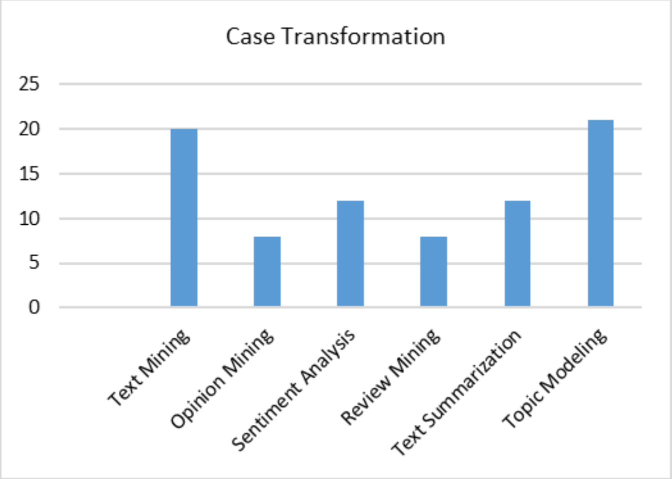

Fig.1 Number of Case Transformations in various selected fields

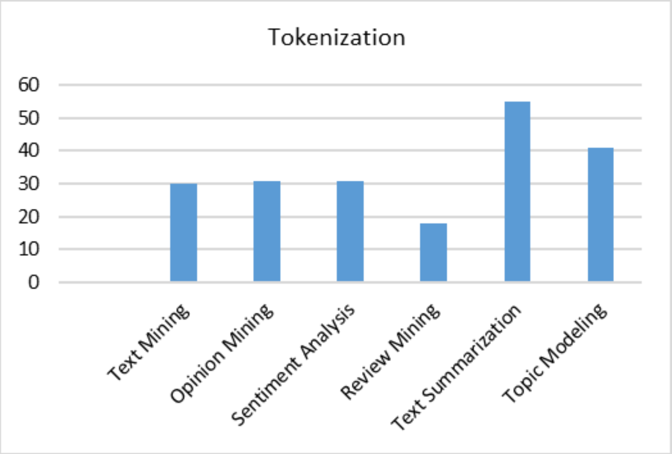

Fig.2 Number of Tokenization in various selected fields

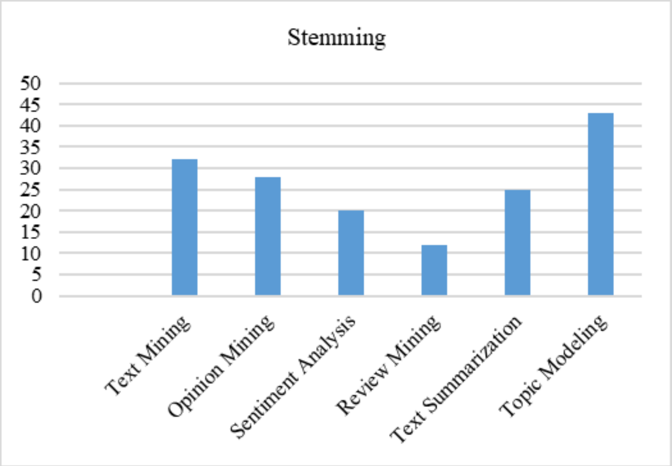

Fig.4 Number of stemming in various selected fields

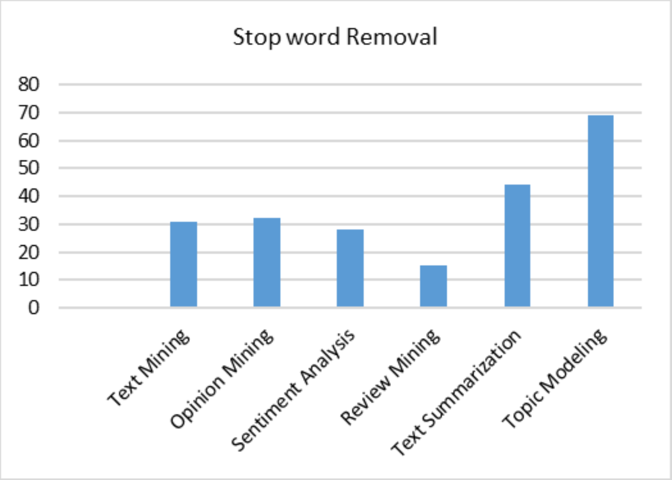

Fig.3 Number of Stopword Removal in various selected fields

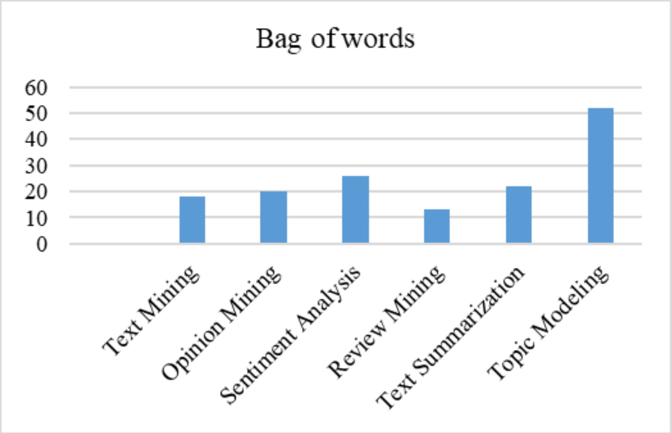

Fig.5 Number of Bag of words in various selected fields

Stemming is high in topic modeling while less in review mining as shown in Fig.4. It was observed that most of the articles which dealt with the problem of topic modeling used third party APIs. This use of APIs can be the reason because the stemming process is already implemented in many available tools. Bag of word is high in topic modeling and less in review mining as shown in Fig. 5. The reason is simply the requirement of words' collection. In opinion or review mining, the number of words is usually less as compared to the text for topic modeling or text summarization.

Similarly, POS is high in sentiment analysis papers while less in review mining papers as shown in Fig.6. Normally, the context would be required for further tagging in POS process and that's why research papers in the literature dealing with sentiment analysis have reported the use of text transformation POS more as compared to review mining.

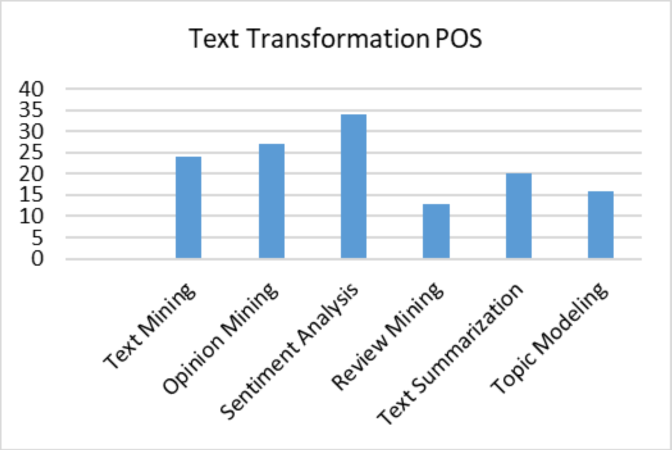

Fig.6 Number of Text Transformations POS in various selected fields

B. Categorization with respect to domain

From Table III, it can be observed that Text Summarization is the preferred text mining technique in the domain of education. One of the main reason is that text summarization is used in automatic SLRs as reported in many review articles [19 -23]. In many fields, Sentiment analysis is preferred on social networking platform and the reason is that on social network, emotion and feeling are expressed. Review mining is mainly used in medical and business domains. In the medical domain, people give reviews about treatments and medicine. Such reviews help other people in making informed decisions about buying medicine and choosing a treatment. In the domain of business, customers give reviews about products and services from which a business organization can learn about customers' preferences and issues and thus by incorporating such information in their decision-making process, a business can get competitive advantage over others.

From the above statistics, it is established that education is mostly focused on text summarization because in educational field, people mostly search the summary to understand the topic easily and quickly. Books' summary is also preferred instead of reading a whole book to understand idea and main crux in order to save time. Text summarization is also helpful in doing literature review and systematic review for educational purpose.

Table III. CLASSIFICATION WITH RESPECT TO DOMAIN

Technique	**Business**	**Education**	**Medical**	**Social Network**
Text Mining	29	25	30	19
Opinion Mining	16	37	5	19
Sentiment Analysis	9	34	3	39
Review Mining	36	42	36	15
Text Summarization	5	90	7	12
Topic Modeling	20	40	16	35

C. Categorization with respect to attributes along with tools

Table VI shows that for *tokenization*, Stanford Tokenizer is the most widely used tool. For *stop word removal*, Stop Word Filter is the most widely used tool. Similarly, for *stemming*, Porters Stemmer is the most widely used technique among all the available tools. As evident from the data in Table IV, most of the text mining fields use the majority of pre-processing techniques such as case conversion, stop-word removal and bag of words.

TABLE IV. CATEGORIZATION WITH RESPECT TO ATTRIBUTES ALONG WITH TOOLS

Pre-processing Technique	**Tools**	**No**
Case Transformation	Case converter	2
Tokenization	Standford Tokenizer	21
	Potts's tokenizer	1
	Open NLP Tokenizer	4
Stop word Removal	Stop word filter	11
	Stop-word handler called 'Rainbow'	3
Stemming	Snowball stemmer	1
	three stemming tools: Khoja Arabic Stemmer (Khoja & Garside, 1999), Information Science Research Institute (ISRI) Arabic Stemmer (Taghva, Elkhoury, & Coombs, 2005) and Tashaphyne Light Arabic Stemmer (Zerrouki,2012)	
	Porters Stemmer	23
Bag of words	Bag of word 3 Paragraph Vector Distributed Bag of Words (PV-DBOW) proposed by Mikolov.	15
	Continuous Bag of words CBOW	17
Text Transformation POS	Standford Tagger as POS Tagger.	31

V. CONCLUSION

To resolve the issues of the terminology jargon in text mining, a comprehensive literature survey was conducted and based on the analysis of surveyed literature, a clear classification was proposed. Categorization was performed with respect to the preprocessing steps, domains and attributes along with Tools. Results show that text summarization is mainly used in the domain of education, whereas review mining is mostly used in the medical and business domains. Further, results show that stop word removal is the most used preprocessing step.

Moreover, Porter Stemmer is the most widely tool used. Similarly, text summarization and topic modeling are very similar in terms of preprocessing techniques and fields of applications.

REFERENCES

[1] Younis, E.M., Sentiment analysis and text mining for social media microblogs using open source tools: an empirical study. International Journal of Computer Applications, 2015. 112(5).

[2] Dhanalakshmi, V., D. Bino, and A. Saravanan. Opinion mining from student feedback data using supervised learning algorithms. in 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC). 2016. IEEE.

[3] Bucur, C., *Using opinion mining techniques in tourism.* Procedia Economics and Finance, 2015. **23**: p. 1666-1673.

[4] Jeyapriya, A. and C.K. Selvi. Extracting aspects and mining opinions in product reviews using supervised learning algorithm. in 2015 2nd International Conference on Electronics and Communication Systems (ICECS). 2015. IEEE.

[5] Rani, M. and S, Jan. "Comparative Analysis and Enhancement of Sentiment Intensity Based Tools." 2020 14th International Conference on Open Source Systems and Technologies (ICOSST). IEEE, 2020

[6] Slamet, C., et al. Automated text summarization for indonesian article using vector space model. in IOP Conference Series: Materials Science and Engineering. 2018. IOP Publishing.

[7] Gambhir, M. and V. Gupta, *Recent automatic text summarization techniques: a survey.* Artificial Intelligence Review, 2017. **47**(1): p. 1-66.

[8] Sajid, Anamta, Sadaqat Jan, and Ibrar A. Shah. "Automatic topic modeling for single document short texts." 2017 International Conference on Frontiers of Information Technology (FIT). IEEE, 2017.

[9] Al-Radaideh, Q.A. and D.Q. Bataineh, A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms. Cognitive Computation, 2018. **10**(4): p. 651-669.

[10] Vijayarani, S., M.J. Ilamathi, and M. Nithya, Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 2015. 5(1): p. 7-16.

[11] Hu, Y.-H., Y.-L. Chen, and H.-L. Chou, *Opinion mining from online hotel reviews–A text summarization approach.* Information Processing & Management, 2017. **53**(2): p. 436-449

[12] Nickerson, R.C., U. Varshney, and J. Muntermann, *A method for taxonomy development and its application in information systems.* European Journal of Information Systems, 2013. **22**(3): p. 336-359.

[13] Park, H. and S. McKilligan. A systematic literature review for human-computer interaction and design thinking process integration. in International Conference of Design, User Experience, and Usability. 2018. Springer.

[14] Khan, M.T. and S. Khalid. Trends and Challenges in Lifelong Machine Learning Topic Models. in 2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube). 2018. IEEE.

[15] Kamal, A., Review mining for feature based opinion summarization and visualization. arXiv preprint arXiv:1504.03068, 2015.

[16] Fromm, H., T. Wambsganss, and M. Söllner, *Towards a taxonomy of text mining features.* 2019.

[17] Usai, A., et al., Knowledge discovery out of text data: a systematic review via text mining. Journal of knowledge management, 2018.

[18] Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420 (2021).

[19] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi, “Review of automatic text summarization techniques & methods” Journal of King Saud University - Computer and Information Sciences, 2020,

[20] Abilio, Ramon, et al. "Systematic literature review supported by information retrieval techniques: a case study." 2014 XL Latin American Computing Conference (CLEI). IEEE, 2014.

[21] Adeva, JJ García, et al. "Automatic text classification to support systematic reviews in medicine." Expert Systems with Applications 41.4 (2014): 1498-1508.

[22] Feng, Luyi, Yin Kia Chiam, and Sin Kuang Lo. "Text-mining techniques and tools for systematic literature reviews: A systematic literature review." 2017 24th Asia-Pacific Software Engineering Conference (APSEC). IEEE, 2017.

[23] Fernández-Sáez, Ana M., Marcela Genero Bocco, and Francisco P. Romero. "SLR-Tool: A Tool for Performing Systematic Literature Reviews." ICSOFT (2). 2010.