ISSN: 2755-0214

Journal of Economics & Management Research

SCIENTIFIC Research and Community

Research Article

Open Access

A Systematic Survey of Business Intelligence Literature Using Machine Learning Techniques

Georgios Fakas, Elias Houstis and Manolis Vavalis*

Department of Electrical and Computer Engineering University of Thessaly Sekeri and Heiden Pedion Areos, Zip 383 34, Volos Greece

ABSTRACT

Business intelligence is the field that develops methodologies and tools for analysis of business information to assist the management and decision process of a corporation. The principal aims of this study are a) to complement the existing literature surveys in the BI area by identifying publications for the period 2007 to 2020, b) to classify these publications according to research strategies and various well-defined research topic categories, and c) apply machine learning techniques to assess their 'Relevance' with the BI discipline. We have collected 332 papers using 'Google 'Scholar' using a set of related keywords associated with the BI literature. The results show that most papers appeared in 2015 and 2017. The classifications of the literature based on research strategies and topics indicate that most papers address 'formal theory and/or reviews' and belong to the 'benefits' topic category. For estimating the 'Relevance' of the surveyed publications, we extracted information from them using the natural language techniques 'term-document matrix (TDM)' and 'Topics' and apply machine learning techniques to the generated input feature spaces. The experiments indicate that the overall best individual classifier was the Random Forest with SMOTE sampling on 50% of the original data applied to the 'Topic' feature space, achieving 62.12% accuracy. The next best classifier is the Neural Networks with ROSE sampling on 50% of the original data with the 'TDM input feature space, giving 53% accuracy. The best ensemble type classifier was Neural Networks with ROSE sampling, polynomial SVM without oversampling, and Gradient Boosting without oversampling, which achieved 76.92% accuracy using the 'Topic' input feature space.

***Corresponding author**

Manolis Vavalis, Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece. Tel: ++30 6974738481; E-mail: mav@uth.gr

Received: March 23, 2022; **Accepted:** March 31, 2022; **Published:** April 02, 2022

Keywords: Business Intelligence, Business Intelligence Survey, Research Opportunities, Text Mining, Machine Learning

Introduction

Business intelligence has two different basic meanings related to the term intelligence [1]. The primary one is the human intelligence capacity applied in business activities. The second and most crucial one relates to intelligence as information. It can be viewed as the expert information, knowledge, and technologies utilized to manage a business. In recent years, Business Intelligence (BI) has been studied by many researchers whose resulting practices and technologies have been incorporated into the operations of many business organizations. The large number of publications in the BI area indicates its importance, and several surveys have been published that attempt to classify them and identify the research issues in BI analytics.

Martin Aruldoss et al [2]. proposed and defined seven categories for the classification of BI literature. These are (1) applications, (2) intelligent techniques, (3) information extraction, (4) integration of BI with other techniques and methods, (5) prototypes, design models, and frameworks for BI applications, (6) evaluation and performance assessment of BI systems, and (7) challenges and issues in BI implementation. They analyze and justify these categories and attempt to identify areas lacking in recent research, thereby offering potential opportunities for investigation.

R. Heang and R. Mohanh present a general overview of the factors used in a literature review of BI without considering specific data [3]. Their approach is based on existing surveys and data. Their proposal assumes that BI is integrated and implemented quite differently among organizations. Thus, they propose reviewing BI literature by adopting categories like BI application and its implementation, BI architectures, and enabling factors. Furthermore, they discuss how technological capabilities such as user access, data quality, the integration of BI with other systems in the firm, and organizational capabilities such as flexibility and risk management support are essential for BI success, regardless of the decision environment. They believe that their analysis could create value and input for enterprises that plan to implement a BI application in their organization.

Kowalczyk, Martin et al. conducted a structured and extensive BI literature review until 2007 from a managerial decision process point of view [4]. They develop a framework by integrating existing results into managerial decision processes. The results of their study concern the effects of decision support technologies on the distinct phases, characteristics, and outcomes of decision processes. Finally, they discuss the findings and implications for business intelligence and analytics systems from the perspective of the decision process.

In Jourdan Z. et al. surveyed 167 articles with topics closely related to Business Intelligence in the period 1997 to 2006 from ten leading Information Systems (IS) journals [5]. They selected the articles using the keywords' business analytics', 'business intelligence', 'data mining', and 'data warehousing'. They excluded book reviews and editorials. The categorical strategies and the results obtained are restated in the following sections. Our survey includes BI literature from 2007 to 2020 and relates it to the above systematic study utilizing similar methodologies [5].

For a more systematic literature review and estimating its 'Relevance' to a particular subject area, Stijn Jaspers et al [6]. In proposed applying machine learning techniques (MLT) for screening abstracts, data extraction, and critical appraisal. We have used their machine learning framework and tool for the survey of BI literature. The MLT techniques applied include support vector machines (SVM), Gradient boosting, neural networks, random forest, and ensemble methods [7].

The organization of the paper is as follows. Section 2 presents a brief review of the 1997 – 2006 survey of BI literature in [5]. Section 3 presents the BI literature survey for the period 2006 -2020 and discusses the findings of this study. Section 4 describes the AI methodologies and tools proposed in to estimate BI literature's Relevance. Section 5 elaborates on the performance of various statistical and machine learning algorithms applied to the pool of publications of this survey [6]. Finally, section 7 summarizes the results and observations of our survey.

Methodologies for Surveying BI literature

The survey of BI literature by Jourdan Z. et al. included 167 articles with topics closely related to Business Intelligence from 1997 to 2006 in ten leading Information Systems (IS) based on the following phases [5].

Phase 1: Accumulation of Article Pool using the ABI/INFORM database to search for the articles based on keywords such as 'business analytics,' 'business intelligence,' 'data mining,' and 'data warehousing.' They excluded book reviews and editorials.

Phase 2: Categorization of Articles by Research Strategy

In this phase, nine research strategies were adopted for the categorization of the articles, which are defined on the x-axis of the left graph in Figure 1. In addition, three reviewers provided their opinions, and their disagreements were recorded according to the following three strategy trade-offs:

- Generalizability to the population that supports the issue of external validity
- Precision in measurement and control of behavioral variables affecting internal and construct validity
- Realism of context

Figure 1 summarizes the categorization of the BI literature surveyed in according to the nine research strategies specified on the x-axis [5]. The left bar graph in Figure 1 displays the number of publications in these nine categories. The right bar graph depicts the distribution of the 167 articles during the period 1997-2006.

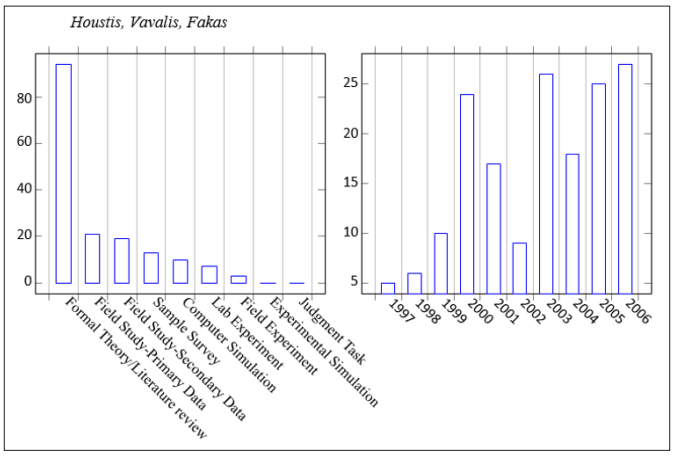

Figure 1: Number of BI publications according to nine categories of research strategies (graph on the left) and the publication year (graph on the right).

Phase 3: Categorization of the Articles by BI Categories

The following five categories were used in to classify the 167 articles in this survey [5].

AI - Artificial Intelligence category consists of algorithms and applications of AI. The applications of the AI category addressed classification, prediction, web mining, and machine learning.

BEN - Benefits category details how corporations have used data warehousing, data mining, and/or an enterprise-wide BI system to accomplish some tangible financial benefit.

DEC - Decisions category contains articles associated with improving overall decision-making and includes such subjects as data modeling, decision making, and modeling.

IMP - Implementation category covers project management issues in a variety of BI contexts, including data warehousing, data mining, customer relationship management (CRM), enterprise resource planning (ERP), knowledge management systems (KMS), and eBusiness projects.

STR - Strategies category focuses on how to apply BI tools and technologies in the modern business environment. This category covers improving internal performance (i.e., enterprise agility, marketing, and integrating business functions), working with external partners to enhance supply chain collaboration, and providing the customer a better experience through customization/ personalization relationship management (CRM).

Table 1 presents the distribution of BI categories versus research strategies used in articles on the various BI topics. The column labels are abbreviations of research strategies considered in [5] defined as FT/LR- Formal Theory/Literature Review, SS - Survey LE -Lab Exp, ES- Exp. Sim, FS/PD - Field – Pri, FS/SD - Field - Sec. Field Exp, JT - Judgment Task, and CS - Comp. Sim.

Table 1: Research Strategy by Research Category (1997-2006)

	CS	ES	FE	FS/PD	FS/SD	FT/LR	JT	LE	SS	Total
AI	7	-	2	-	10	17	-	-	1	37
Benefits	-	-	-	1	2	5	-	-	2	10
Decision	1	-	-	2	3	12	-	7	1	26
Implementation	-	-	-	10	2	17	-	-	6	35
Strategies	2	-	1	8	2	43	-	-	3	59
Total	10	0	3	21	19	94	0	7	13	167

BI Literature Survey and Analysis for Period 2007 To 2020

Our survey applies the BI literature review workflow framework proposed in [5] to 332 selected articles published during the period 2007 – 2020. These articles have been identified through 'Google Scholar' search tool using the keywords' 'business analytics,' 'business intelligence,' 'competitive intelligence,' 'data mining', and 'data warehousing.' We have selected only free access papers, included dissertations, and excluded book reviews and editorials. We also excluded several duplicate articles [8]. The final survey consists of 332 articles, with their per year distribution appearing in the second column of Table 2.

Classification of BI literature based on Research Strategies

For the classification of the identified articles in our survey, we adopted the nine categories of research strategies proposed in: 'formal theory/literature review,' ' sample survey,' ' laboratory experiment,' 'experimental simulation,' 'field study (primary data),' ' field study (secondary data)', 'field experiment', 'judgment task', and 'computer simulation [5].' The definition of these strategies is presented in [5].

The 3–11 columns of Table 2 indicate the categorization of the 332 articles of our pool according to the above nine research strategies [8]. It shows that almost half of the articles have been devoted to the 'formal theory/literature review' strategy. About 30% of the survey papers are associated with 'sample survey', 'field study-secondary data', and 'field study-primary data' research strategies.

Therefore, 'formal theory/literature reviews' is the most dominant category of BI research strategies, significantly different from all other research categories. It is almost four times bigger than the second 'sample survey' category. Table 2 presents the distribution of publications over the years 2007-2020 based on the above five research categories in the last five columns. The data indicate that the most dominant research category is 'Benefits', barely winning over the second research category 'Strategies'. Table 2 summarizes the distribution of our survey papers for all research strategies and categories considered in the period 2007-2020.

Table 2: The per Year Distribution of Papers in 2007 – 2020 Survey According to Identified Research Strategies and Categories

	Total	Research Strategy									Research Category				
		CS	ES	FE	FS PD	FS SD	FT LR	JT	LE	SS	AI	BEN	DEC	IMP	STR
2007	26	3	2	1	1	1	13	1	1	3	1	4	5	7	9
2008	16	1	4	1	-	2	6	-	-	2	3	2	3	3	5
2009	11	-	1	-	-	2	6	-	1	1	-	4	3	-	4
2010	15	1	1	-	1	2	9	-	-	1	3	2	2	2	6
2011	20	1	-	-	-	4	11	-	2	2	4	9	1	2	4
2012	32	5	1	2	1	5	14	2	-	2	5	11	1	8	7
2013	20	1	1	-	2	2	11	-	-	3	1	7	2	6	4
2014	30	2	2	1	2	3	15	-	-	5	3	8	3	7	9
2015	43	6	2	4	3	5	16	-	2	5	2	13	7	9	12
2016	29	1	1	-	4	2	16	-	1	4	2	8	5	8	6
2017	40	2	1	3	3	2	23	1	1	4	7	12	3	9	9
2018	23	2	-	1	6	3	6	-	1	4	3	7	2	4	7
2019	17	1	1	-	6	2	6	-	-	1	1	3	4	4	5
2020	10	-	1	-	2	-	7	-	-	-	-	2	3	2	3

Comparison of BI Research Category vs. Research Strategy

To compare Research Strategy and Research Category, we present in Table 3 the distribution of the survey papers among Research Strategies vs the Research Categories. We conclude that the 'Formal Theory/Literature strategy' category dominates across all Research Categories. Another observation worth noticing is that the 'Artificial Intelligence' category is often combined with the 'Computer Simulation strategy' because it is more technology focused.

Table 3: Research Strategy by Research Category

	CS	ES	FE	FS/PD	FS/SD	FT/LR	JT	LE	SS	Total
AI	11	5	1	2	3	9	2	1	1	35
Benefits	3	3	2	11	14	44	-	4	11	92
Decision	1	4	3	5	1	22	1	-	7	44
Implementation	6	3	3	7	7	38	-	3	4	71
Strategies	5	3	4	6	10	46	1	1	14	90
Total	26	18	13	31	35	159	4	9	37	332

Almost half of the articles associated with 'Benefits', 'Decision', 'Implementation', and 'Strategies' categories utilize the 'Formal Theory/Literature Review' strategy. The less technical categories such as 'Benefits', 'Implementation', and 'Strategies' include most of the survey papers. Comparing the research categories in 1997-2006 against those in 2007-2020, it is observed that the 'Artificial Intelligence' category remains at the same level. On the other hand, the 'Benefits' category skyrocketed due to the wide use of business intelligence systems. The 'Implementation' category also was doubled because of the various project implemented by corporations.

In conclusion, 'Formal Theory/Literature review' is the most dominant strategy in both survey periods.

Machine Learning Techniques for Literature Reviews

For a more systematic literature review and estimating its 'relevance' to a particular subject area, Stijn Jaspers et al. in proposed applying machine learning techniques (MLT) for screening abstracts, data extraction, and critical appraisal [6]. We have used their machine learning framework and tool for the survey of BI literature. The MLT techniques applied include support vector machines (SVM), Gradient boosting, neural networks, random forest, and ensemble methods and their description cab found in many references including [7].

In applying machine learning algorithms for literature classification, the data must be transformed into a proper format. Unlike structures data, features are not explicitly available in text data. Thus, we need to use a process to extract features from the text data. One way is to consider each word as a feature and find a measure to capture whether a word exists or does not exist in a sentence. This technique is called the bag-of-words (BoW) model. In this case, each sentence is treated as a bag of words. Each sentence is called a document, and the collection of all documents is called corpus. The first step in creating a BoW model is to create a dictionary of all the words used in the corpus. At this stage, we will not worry about grammar, and only the occurrence of the words is captured. Then, we will convert each document to a vector that represents words available in the documents. In this model, each word's occurrence (frequency of) is used as a feature for training a classifier and determines its 'Relevance.' This leads to the classical measure known as Term-Document Matrix (TDM). The key issue here is understanding that a TDM is an aggregated quantity and thus "hides" the raw information originally available in the text. A widely accepted way to "normalize" the term frequencies (TF) is to weight a term by the Inverse of Document Frequency (IDF). Another conversion model is the n-gram that predicts the occurrence of a word based on the occurrence of its n – 1 previous words. The bigrams model is a variation of the n-gram, which uses two consecutive words. The Latent Dirichlet Allocation (LDA) approach is a probabilistic modeling technique to identify topics presented in a corpus [9]. As a result, one might utilize the obtained topics instead of employing the term frequency as input parameters. In our case we denote that as 'Topics' input space. Another encoding of words in a text is the Word2Vec methodology, which helps establish the association of a word with another similar meaning word through vector representation of words.

Our study considered input space to machine learning algorithms, either the term-document matrix (TDM) or the topics approach ('Topics') for all the datasets utilized.

Class imbalance is a general concept related to classification. It is the unequal ratio of relevant and irrelevant abstracts in training sets. The solution to the above imbalance problem is the Synthetic Minority Oversampling Technique (SMOTE) and Random Over-Sampling Examples Technique (ROSE) sampling [10]. These techniques allow the creation of more balanced training datasets and used in our study.

We use different metrics for estimating the document classifier's performance. One of these metrics is the confusion matrix shown in Table 4. The elements of the matrix are defined as follows: True Positives (TP) is the number of relevant documents being classified as relevant, True Negatives (TN) is the number of irrelevant documents being classified as irrelevant, False Positives (FP) is the number of irrelevant documents being classified as relevant and False Negatives (FN) is the number of relevant documents that falsely classified as irrelevant.

Table 4: Confusion Matrix

Classified Category/ True	Relevant	Irrelevant
Relevant	TP	FP
Irrelevant	FN	TN

Other classifier performance metrics used in our study include the following:

Recall or Sensitivity: $\frac{TP}{TP+FN}$ indicates the proportion of correctly identified relevant documents positives among all the truly relevant documents.

Precision: $\frac{TP}{TP+FP}$ indicates the proportion of correctly identified relevant documents among all the documents that were classified relevant.

F-measure: $\frac{(\beta^2+1)TP}{(\beta^2+1)TP+FP+\beta^2FN}$ he parameter **β** blends precision and recall. **β** = 1 corresponds to the weighted harmonic mean, **β** < 1 indicates more weight is placed on precision and **β**>1 more weight on recall.

Specificity: $\frac{TN}{TN+FP}$ corresponds to the proportion of irrelevan

papers that were correctly classified.

Data Word Cloud

Usually, before we train the machine learning algorithms, we create a visual representation of a naive text summarization in the form of the four word-clouds depicted in Figure 3. These graphs follow the specific keywords mentioned in Section 5 and are used to select the articles considered in our study.

Figure 3: Four world clouds are depicted generated from all the article abstracts in our study and corresponding to input spaces TDM, TFIDF, bigram, and trigram with 50 minimum frequency and 50 words from left to right.

Performance of Machine Learning Classifiers for the Bi Literature Classification

We used 20%, 50%, and 80% of the constructed data from the survey data to train the machine learning algorithms considered in our study. Throughout, we denote these subsets as D20, D50, and D80.

The machine learning algorithms employed were **Support Vector Machine (SVM), Gradient Boosting (GBM), Neural Networks (NN), and Random Forest (RF)**. The ensemble methods considered consist of combinations of the above algorithms and data subsets. We implemented 18 different individual machine learning classifiers with ROSE and SMOTE sampling corresponding to the three different training subsets mentioned above. The classifiers were run 6 times, giving the same results. Table 5 presents the results obtained. Table 6 presents the overall results for the classifiers considered.

Table 5: The performance measures of ML algorithms with TDM (rows 4–15) and Topics (rows 16–27) input feature space for the datasets D20, D50, and D80 and the predictions obtained (% irrelevant vs. %relevant)

% of training			Test										
			Performance				Performance Validation				Prediction Validation		
20	F1	NN_rose	.85	.28	.49	.46	.79	.34	.46	.47	-1 1	16 4	31 15
	SE	RF_rose	.98	.03	.48	.33	1	0	.45	.29	-1 1	0 0	47 19
	SS	NN_rose	.85	.28	.49	.46	.79	.34	.46	.47	-1 1	16 4	31 15
	EN	GBM_rose, GBM_ orig	.03	1	.06	.7	0	1	0	.71	-1 1	47 19	0 0
50	F1	GBM_rose/ RF_rose	1	.02	.48	.33	1	0	.45	.29	-1 1	0 0	47 19
	SE	GBM_rose/ RF_rose	1	.02	.48	.33	1	0	.45	.29	-1 1	0 0	47 19
	SS	NN_rose	.49	.55	.39	.53	.37	.6	.31	.53	-1 1	28 12	19 7
	EN	NN_rose, svm_ Radial_smote	.27	.87	.34	.68	.26	.81	.3	.65	-1 1	38 14	9 5
80	F1	GBM_rose	.88	.17	.47	.38	1	.04	.46	.32	-1 1	2 0	45 19
	SE	GBM_rose	.88	.17	.47	.38	1	.04	.46	.32	-1 1	2 0	45 19
	SS	RF_rose	.69	.22	.4	.37	1	.13	.48	.38	-1 1	6 0	41 19
	EN	svm_R_O, svm_L_O, svm_R_R	.44	.92	.54	.77	.42	.66	.37	.59	-1 1	31 11	16 8
20	F1	RF_rose	.94	.14	.49	.39	.84	.06	.41	.29	-1 1	3 3	44 16
	SE	RF_rose	.94	.14	.49	.39	.84	.06	.41	.29	-1 1	3 3	44 16
	SS	svm_Poly_smote	.65	.25	.39	.38	.84	.26	.46	.42	-1 1	12 3	35 16
	EN	RF_rose, NN_	.03	.99	.06	.69	0	1	0	.71	-1 1	47 19	0 0
	orig	.03	.99	.06	.69	0	1	0	.71	-1 1	47 19	0 0	
50	F1	RF_smote	.49	.68	.44	.62	.53	.74	.49	.68	-1 1	35 9	12 10
	SE	svm_Poly_smote	.68	.35	.44	.45	.79	.26	.43	.41	-1 1	12 4	35 15
	SS	RF_smote	.49	.68	.44	.62	.53	.74	.49	.68	-1 1	35 9	12 10
	EN	svm_L_O, svm_R_R, svm_R_O	.27	.78	.31	.62	.11	.74	.12	.56	-1 1	35 17	12 2

80	F1	svm_Poly_rose	.56	.41	.39	.46	.58	.45	.39	.48	-1 1	21 8	26 11
	SE	svm_Poly_rose	.56	.41	.39	.46	.58	.45	.39	.48	-1 1	21 8	26 11
	SS	svm_Linear_smote	.38	.58	.32	.52	.42	.66	.37	.59	-1 1	31 11	16 8
	EN	NN_R, svm_P_O, GBM_O	.5	.89	.57	.77	.26	.68	.26	.56	-1 1	32 14	15 5

Table 6: The Performance of machine learning algorithms with TDM (rows 3–20), and Topic (rows 21-38) input feature spaces and the three datasets (D20, D50, D80) considered

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
	20 50 80	20 50 80	20 50 80	20 50 80	20 50 80	20 50 80	20 50 80	20 50 80	20 50 80
GBM_orig	.56 .58 .56	-.09 -.05 -.14	.22 .23 .15	.20 .20 .13	.72 .76 .75	.24 .27 .18	.66 .68 .66	.24 .27 .18	.20 .20 .13
GBM_rose	.69 .33 .38	.04 .01 .03	.11 .48 .47	.06 1 .88	.97 .02 .17	.50 .32 .32	.70 1 .75	.50 .32 .32	.06 1 .88
GBM_smote	.59 .56 .63	-.01 -.08 -.02	.26 .22 .17	.23 .20 .13	.76 .73 .86	.30 .24 .29	.69 .67 .69	.30 .24 .29	.23 .20 .13
NN_orig	.54 .55 .63	-.07 -.07 .10	.26 .25 .34	.26 .24 .31	.67 .69 .78	.26 .26 .38	.67 .67 .72	.26 .26 .38	26 .24 .31
NN_rose	.46 .53 .46	.09 .03 -.14	.49 .39 .26	.85 .49 .31	.28 .55 .53	.35 .33 .23	.80 .70 .63	.35 .33 .23	.85 .49 .31
NN_smote	.52 .55 .62	-.05 .02 .10	.31 .35 .38	.35 .39 .38	.60 .63 .72	.28 .32 .38	.67 .70 .72	.28 .32 .38	.35 .39 .38
RF_orig	.69 .69 .69	.00 .00 .00	– .00 –	.00 .00 .00	1 1 1	– .00 –	.69 .69 .69	– .00 –	.00 .00 .00
RF_rose	.33 .33 .37	.01 .01-.06	.48 .48 .40	.98 1 .69	.03 .02 .22	.31 .32 .28	.80 1 .62	.31 .32 .28	.98 1 .69
RF_smote	.46 .55 .54	-.13 -.15 -.12	.28 .14 .20	.33 .12 .19	.52 .74 .69	.24 .17 .21	.63 .65 .66	.24 .17 .21	.33 .12 .19
svm_Linear_orig	.69 .65 .69	.00 -.05 .00	– .04 –	.00 .02 .00	1 .93 1	– .14 –	.69 .68 .69	– .14 –	.00 .02 .00
svm_Linear_rose	.58 .57 .56	.09 -.02 -.02	.40 .30 .30	.44 .29 .31	.65 .69 .67	.36 .30 .29	.72 .68 .69	.36 .30 .29	.44 .29 .31
svm_Linear_ smote	.56 .48 .58	-.04 -.16 .04	.27 .23 .35	.26 .24 .38	.70 .59 .67	.28 .21 .33	.68 .64 .71	.28 .21 .33	.26 .24 .38
svm_Poly_orig	.67 .69 .69	-.03 .00 .00	.03 .00 –	.02 .00 .00	.97 1 1	.17 .00 –	.68 .69 .69	.17 .00 –	.02 .00 .00
svm_Poly_rose	.50 .61 .58	.01 -.10 -.03	.39 .07 .27	.52 .05 .25	.49 .87 .72	.31 .14 .29	.69 .67 .68	.31 .14 .29	.52 .05 .25
svm_Poly_smote	.52 .67 .56	-.02 -.04 -.02	.35 .00 .30	.41 .00 .31	.57 .97 .67	.30 .00 .29	.68 .68 .69	.30 .00 .29	.41 .00 .31
svm_Radial_orig	.68 .69 .69	.01 .00 .00	.08 .00 –	.05 .00 .00	.97 1 1	.38 .00 –	.69 .69 .69	.38 .00 –	.05 .00 .00
svm_Radial_rose	.68 .69 .69	.01 .00 .00	.08 .00 –	.05 .00 .00	.97 1 1	.38 .00 –	.69 .69 .69	.38 .00 –	.05 .00 .00
svm_Radial_ smote	.61 .70 .65	-.04 .10 .06	.19 .20 .25	.15 .12 .19	.82 .96 .86	.27 .56 .38	.68 .71 .70	.27 .56 .38	.15 .12 .19
GBM_orig	.61 .64 .65	-.04 .03 .06	.19 .23 .25	.15 .17 .19	.82 .86 .86	.27 .35 .38	.68 .70 .70	.27 .35 .38	.15 .17 .19
GBM_rose	.53 .38 .33	-.08 -.07 -.29	.27 .40 .22	.27 .66 .31	.65 .25 .33	.26 .28 .17	.66 .62 .52	.26 .28 .17	.27 .66 .31
GBM_smote	.57 .54 .58	.00 -.07 .01	.32 .27 .31	.33 .27 .31	.67 .66 .69	.31 .26 .31	.69 .67 .69	.31 .26 .31	.33 .27 .31
NN_orig	.69 .69 .69	.00 .00 .00	– – –	.00 .00 .00	1 1 1	– – –	.69 .69 .69	– – –	.00 .00 .00
NN_rose	.69 .53 .52	.00 –.08 –.07	– .26 .29	.00 .27 .31	1 .65 .61	– .26 .26	.69 .66 .67	– .26 .26	.00 .27 .31
NN_smote	.69 .52 .52	.00 -.06 -.04	– .30 .32	.00 .34 .38	1 .59 .58	– .27 .29	.69 .67 .68	– .27 .29	.00 .34 .38
RF_orig	.68 .67 .69	-.01 .01 .09	.03 .09 .20	.02 .05 .38	.98 .96 .94	.25 .33 .50	.69 .69 .71	.25 .33 .50	.02 .05 .13
RF_rose	.39 .44 .50	.05 -.02 -.13	.49 .40 .24	.94 .61 .25	.14 .36 .61	.33 .30 .22	.83 .67 .65	.33 .30 .22	.94 .61 .25
RF_smote	.57 .62 .58	.04 .16 -.07	.35 .44 .21	.38 .49 .19	.66 .68 .75	.33 .41 .25	.70 .75 .68	.33 .41 .25	.38 .49 .19
svm_Linear_orig	.69 .69 .69	.00 .00 .00	– – –	.00 .00 .00	1 1 1	– – –	.69 .69 .69	– – –	.00 .00 .00
svm_Linear_rose	.58 .45 .50	.09 -.05 -.06	.41 .37 .32	.45 .51 .38	.64 .43 .56	.37 .29 .27	.72 .66 .67	.37 .29 .27	.45 .51 .38
svm_Linear_ smote	.51 .50 .52	.04 -.05 -.04	.41 .33 .32	.55 .39 .38	.50 .55 .58	.33 .28 .29	.71 .67 .68	.33 .28 .29	.55 .39 .38
svm_Poly_orig	.68 .69 .69	.03 .00 .00	.11 – –	.06 .00 .00	.96 1 1	.40 – –	.69 .69 .69	.40 – –	.06.00.00
svm_Poly_rose	.61 .58 .46	.11 .03 -.02	.40 .34 .39	.42 .34 .56	.69 .69 .42	.38 .33 .30	.73 .70 .68	.38 .33 .30	.42.34.56
svm_Poly_smote	.38 .45 .46	-.07 .03 -.05	.39 .44 .36	.65 .68 .50	.25 .35 .44	.28 .32 .29	.62 .71 .67	.28 .32 .29	.65.68.50
svm_Radial_orig	.69 .69 .69	.00 .00 .00	– – –	.00 .00 .00	1 1 1	– – –	.69 .69 .69	– – –	00.00.00
svm_Radial_rose	.69 .69 .69	.00 .00 .00	– – –	.00 .00 .00	1 1 1	– – –	.69 .69 .69	– – –	00.00.00
svm_Radial_ smote	.58 .50 .60	.02 -.11 -.08	.32 .27 .16	.32 .29 .13	.71 .59 .81	.33 .24 .22	.70 .65 .67	.33 .24 .22	.32.29.13

Discussion of Results

This section discusses the performance of machine learning algorithms presented above with respect to the TDM and 'Topics' input feature spaces extracted from the pool of publications considered in this BI survey.

Discussion of BI Literature Classification Results for TDM input space

In this case, the best classifier was the ensemble of radial SVM, and linear SVM without oversampling, radial SVM with ROSE sampling trained in the D80 dataset produced a Sensitivity of 44%, Specificity of 92%, F1 score of 54%, and accuracy 76.92%. From the individual classifiers, the best one was the Neural Networks with ROSE sampling for the D50 training dataset. It achieved a Sensitivity of 49%, Specificity of 55%, F1 score of 31%, and accuracy of 53.03%. All the other classifiers achieved better Sensitivity, but their Specificity was relatively low. It is worth pointing out that sacrificing Sensitivity could increase the Specificity and lead to a more optimal classifier.

For D20, the results obtained have high Sensitivity, but the Specificity was below 28%. This implies that these classifiers incorrectly classify irrelevant articles as relevant. Only the ensemble classifier was able to raise Specificity, but it dropped the Sensitivity to 3%. We observed similar performance on D50 and D80. The only case where the results were satisfactory was the Neural Networks with ROSE sampling classifier on D50, giving Sensitivity and Specificity of almost 50%. The ensemble on D80 Specificity was raised enough without dropping Sensitivity below 30%.

Discussion of BI Literature Classification Results for Topics input space

For the 'Topic' input space case, the best classifier was the ensemble of Neural Networks with ROSE sampling, polynomial SVM without oversampling, Gradient Boosting without oversampling achieved better performance with Sensitivity of 50%, Specificity of 89%, F1 score of 57% and accuracy of 76.92%. From the individual classifiers, the best was the Random Forest with SMOTE sampling trained in D50. This classifier achieved a sensitivity of 49%, Specificity of 68%, F1 score of 44%, and accuracy of 62.12%.

The individual classifiers at D20 had high Sensitivity, but the Specificity was relatively low. The ensemble classifier on D20 achieved high Specificity but dropped the Sensitivity to 3%. On D50, the results were moderate for Sensitivity and Specificity, with Sensitivity fluctuating from 49% to 68% and Specificity from 35% to 68%, with the ensemble exception that had slightly lowered Sensitivity but higher Specificity. On D80, the Sensitivity and Specificity dropped, with the ensemble exception that had the same Sensitivity but achieved higher Specificity.

We can conclude that classifiers for the 'TDM' feature space achieved high Sensitivity, but the Specificity was low. This means that classifiers could classify relevant articles as relevant but could not classify irrelevant as irrelevant with high precision. The above ensembles increased Specificity but dropped the Sensitivity, especially with a low percentage of the training data, where the percentages between Sensitivity and Specificity were reversed. The same was observed for 'Topic' feature space with the difference that with 50% and 80% percentage of the training data, the Specificity was at moderate percentages. We observed that by increasing the amount of training data, the Specificity was improved too.

Epilogue

Business intelligence is the field that develops methodologies and tools for analysis of business information to assist the management and decision process of a corporation. The principal objectives of this paper are to:

- complement the existing literature surveys in the Business intelligence area by identifying publications for the period 2007 to October 2020
- classify the literature based on research strategies,
- classify the literature according to various well-defined research topic categories, and
- apply machine learning techniques to assess their "Relevance" in the Business intelligence field.

We have collected 332 papers using the 'Google Scholar' tool and a set of related keywords to the field of BI, and we have observed that the most articles were published in the years 2015 (43) and 2017 (40) and that there is an increase in BI articles through years [8]. For the BI literature classification, we first adopted nine research strategies: 'formal theory/literature reviews', 'sample survey', 'laboratory experiment', 'experimental simulation', 'field study (primary data)', 'field study (secondary data)', 'field experiment', 'judgment task,' and 'computer simulation'. The results show that the 'formal theory/literature review' is the most dominant BI research strategy, significantly different from other research categories. It is almost four times bigger than the second 'sample survey' category. The survey literature was classified based on five research topic categories: 'Artificial Intelligence', 'Benefits', 'Decision', 'Implementation', and 'Strategies'. The most dominant BI research topic category in this classification study is 'Benefits', slightly different from the second category of 'Strategies'. Finally, almost half of the articles associated with 'Benefits', 'Decision', 'Implementation', and 'Strategies' categories utilize the 'formal theory/literature review' research strategy.

The final part of this paper involves the classification of the 332 publications by 'Relevance.' We utilized machine learning techniques and the tool described in the EFSA report and dissertation [6, 8]. The overall best individual classifier was the Random Forest with SMOTE sampling on 50% of the original data using the 'Topic' feature space, followed by the Neural Networks with ROSE sampling on 50% of the original data using the 'TDM' feature space. The best ensemble of classifiers was the ensemble on 80% of the original data using the 'Topic' feature space. This ensemble of classifiers consisted of Neural Networks with ROSE sampling, polynomial SVM without oversampling, and Gradient Boosting without oversampling, achieved Sensitivity of 50%, Specificity of 89%, F1 score of 57%, and accuracy of 76.92%.

Acknowledgement

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under RESEARCH - CREATE - INNOVATE project code: T1EDK-02161.

References

1. Ranjan J (2009) Business intelligence: Concepts, components, techniques, and benefits 9: 60-70.
2. Aruldoss M, Lakshmi T, Venkatesan V (2014) A survey on recent research in business intelligence. Journal of Enterprise Information Management 10: 831-866.
3. Heang R (2017) Literature Review of Business Intelligence. Available from: http://www.diva-portal.org/smash/get/diva2:1080911/FULLTEXT01.pdf
4. Kowalczyk M, Buxmann P, Besier J (2013) Investigating Business Intelligence and Ana-lytics from a Decision Process Perspective: A Structured Literature Review https://econpapers.repec.org/paper/darwpaper/.
5. Jourdan Z, Rainer RK, Marshall TE (2008) Business

Intelligence: An Analysis of the Literature. Information Systems Management 25: 121-131.

6. Jaspers S, De Troyer E, Aerts M (2018) Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA. EFSA Supporting Publications 15:1427E.
7. James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning. Springer Texts in Statistics: with Applications in R. Springer 1-441-p.
8. Fakas G (2020) A Systematic Business Intelligence Literature Survey Using Machine Learning Techniques; University of Thessaly, Department of Electrical and Computer Engineering.
9. Blei D, Ng A, Jordan M (2003) Latent Dirichlet Allocation 3: 601-608.
10. Tantithamthavorn C, Hassan A, Matsumoto K (2020) the impact of class rebalancing techniques on the performance and interpretation of defect prediction models.
IEEE Transactions on Software Engineering 26: 1200-1219.