Trend Analysis in Machine Learning Research Using Text Mining

Deepak Sharma[1], Bijendra Kumar[1], Satish Chand[2]

[1]Department of Computer Engineering,
Netaji Subash Institute of Technology, New Delhi, India
[2]School of Computer & Systems Sciences,
Jawaharlal Nehru University, New Delhi, India
{deepak.btg,bizender,schand20}@gmail.com

***Abstract*—This paper aims to identify the trends in machine learning research using text mining. The researcharticles contain significant knowledge and research results. However, they are long and have many noisy results such that it takes a lot of human efforts to analyze them. Text mining can be used to analyze and extract useful information froma large number ofresearch articles quickly and automatically. Text mining is the method of defininginnovative, and unseenknowledge fromunstructured, semi-structured and structured textual data. This knowldege contributed to very important information that can derive from textual data. In this paper, text mining methodsareapplied to detect trends of termsthat occur in the research articles and how they varies over time. We collected21,906 scientific papers from six top journals in the field of machine learning published in period 1988-2017 and analyzed them usingtext mining. Our result analysis shows a changing trend ofvarious terms in Machine learning researchin three decades. The analysis of our study helps the upcoming researchers to explore the significant research area of machine learning.**

Keywords—text mining; machine learning; research trend analysis; data analysis

I. Introduction

Text mining denotes to a process of mining meaningful, non-trivial patterns or knowledge from a set of unstructured texts [1]. It is an essential task to uncover trends from large volume of textual data [2]. In particular, the advent of high-speed internet generates large amounts of textual data in a variety of forms [3]. As an aspect of this trend, research utilizing text mining technique is actively being carried out to find patterns and extract implicit data from the large volume of data in various fields such as academic article information and news article information [1,4,5].The goal of text mining is to determine hiddenknowledge which was not known ealier.[6]. In [7],text mining referred as agroup of techniquesemployed to identifytrends and produceknowledge from data.

Text mining techniquesare derived the frequenciesof important terms in thecontent of thetextual data such as internet chat rooms, articles, or web pages and classifyassociationsbetweenfeatures [8]. Text mining every so ofteninterpretsunorganized text into a effectivecollection of data suitable for data mining for thoroughinvestigation [9].Similar works on various research areas have been performed using text mining. In [10], the text mining has been applied to understand the trend analysis of consumer policy. In [11], the text mining has employed for identifying the primary trends on Big Data in marketing. The research outcome helped to progress more direct efforts in the direction of business for Big Data in the marketing arena. In [12], the text mining applied for knowledge discovery in academic research.

This paper proposes to identify the trends in research of Machine Learning. The research articles available in well-established mainstream journals overthe past three decades, i.e., 1988~2017. In this work, prominent journals included are IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE-PAMI), Journal of Machine Learning Research (JMLR), ScienceDirect Pattern Recognition (ScD-PR), IEEE Transactions on Neural Networks (IEEE-NN), Springer Machine Learning (Sp-ML), and ScienceDirect Neural Networks (ScD-NN) as primary data source.In this paper, text mining techniques employed in a framework for determining the trends of Machine learning research articles published in three decades. These articles include thetitle, abstract, and complete contents of the articles [13].This approach may be helpful to new researchers for further explorationof theirresearch area. The data used for processing in this study were only the title and abstracts of the research articles. Analyzing the title and abstract of a research article is relevant as it comprises the comprehensive objectiveof a research articleand prunedunneeded components of article i.e.figures and tables[13]. The remaining of the paperarranged as follows:Section 2 provides the data collection and preprocessing steps. Section 3 presents the result analysis. Finally, section 4 concludes the paper.

II. Methodology

In this section, we discuss the method of data preparation, description of the corpus, data preprocessing for corpus before applying text transformation in text mining. Fig. 1, shows the methodology for trend analysis in machine learning.

ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE

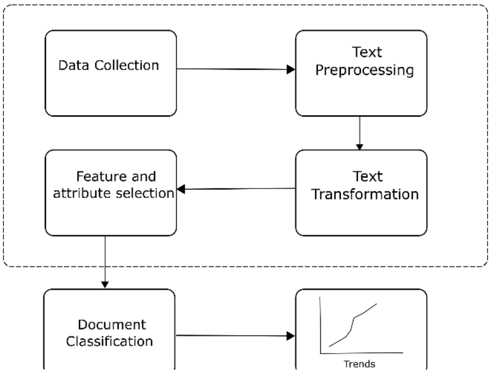


Fig. 1. Methodology for trend analysis

A. Data Preparation

The research data collected from various well-established journals published with high-quality research articles in machine learning. We include the established journals like IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE-PAMI), Journal of Machine Learning Research (JMLR), ScienceDirect Pattern Recognition (ScD-PR), IEEE Transactions on Neural Networks (IEEE-NN), Springer Machine Learning (Sp-ML), and ScienceDirect Neural Networks (ScD-NN). The data used for processing in this study were only the title and abstracts of the research articles from the mentioned journal. Recognizing significant contribution to research, we have included journal articles only for our work. The results are extracted from the time periodfrom 1988 to 2017. Table Ishows the count of journal articles included in this studyas perthe selected journals.Each dataset has considered a separate corpus.

TABLE I. THE NUMBER OF ARTICLES INCLUDED IN THIS STUDY

S.No.	Journal Name	Duration	#Numbers of Years	#Articles Published
1	IEEE-PAMI	1988~2017	30	4,630
2	IEEE-NN	1990~2017	28	4,349
3	JMLR	2000~2017	18	1,755
4	ScD-PR	1988~2017	30	6,567
5	ScD-NN	1988~2017	30	3,294
6	Springer-Machine Learning	1988~2017	30	1,311
Total				21,906

TABLE II. DISTRIBUTION OF NUMBER OF ARTICLES IN EACH DECADE

Journals/Year	1988~1997	1998~2007	2008~2017
IEEE-PAMI	1,245	1,509	1,876
IEEE-NN	878	1,476	1,995
JMLR	0	474	1,281
ScD-PR	1,267	2,105	3,195
ScD-NN	797	1,077	1,420
Springer-ML	293	443	575
#ArticlesCount =21906	4,480	7084	10,342

B. Description of Corpus

Data is gathered by preparing a list of relevantarticles from well-established journals of machine learning. Table II shows the dispersion ofa count of research articles in our study in each decade. The corpus has prepared by collecting articles. It has divided into three datasets. The first dataset (a.k.a. Decade1), second dataset (a.k.a. Decade2), and third dataset (a.k.a. Decade3) consist data for the periodsfrom 1988~1997, 1998~2007, 2008~2017 respectively.The titles and abstracts of the research articles have extracted from the above-mentioned journals.

C. Text Preprocessing

The preprocessing stagecomprises the removal of unwanted words/characters from the corpus, and it performed by executing the following steps. Initially, the titles and abstracts of the articlesare converted into tokens. The producedtokenstransformed into lowercase wordsfor each document. The removal of commas,exclamation points, quotation marks,punctuation characters, apostrophe, question marks, and hyphen performed. Further, the numerical values are eliminated to keep only the textual tokens. Then, the standard English words as specified in *nltk* python package [14] and customized stop-word list [15] with the phrases are removed from the literature dataset. Afterwards, for preparing a useful literature dataset, the wordforms are stemmed to their original root form by using the Porter Stemmer algorithm [16]. Stemming is performed for tokens for each document and converts them into their root term. Finally, we are transforming documents into sparse vectors.The text files in a corpus contain titles and abstracts of articles. The bag-of-words document representations used for converting the documents into vectors. In this representation, each articlerepresented by one vector, where each vector element depicts a pair of *word-wordcount*. The mapping between the words and their word count is called a dictionary. The sparse vectors are created by counting merely the number of occurrences of each distinct word and convert each word to its integer *word_id*.The above steps are used to transform a corpus into vector representation for text mining.

D. Text Transformation

In this phase, the sparse vectors are transformed to TfIdf (term*frequency–inverse document frequency*) vector. The transformation of the articles from one vector representation into another serves two purposes; first, it brings out the hidden structure in the corpus to discover the relationships between the words and describe the documents more semantically. Second, it makes the document representation more compact. The terms are filtered by using two parameters such as word frequency and inverse document frequency. The terms having low occurences in the corpus and lowoccurences in the each document removed. In Bags of words representation, each wordrepresented as a separate variable having numeric weight.In this step, the sparse vectors of the corpus are converted to *TfIdf* vectors using Eq. (1) as a formula for *TfIdf* weights of term i, in document j, in a corpus of D documents.

$$weight\{i,j\} = frequency_{\{i,j\}} * \log_2\left(\frac{\mathrm{D}}{document_{freq\ \{i\}}}\right) \qquad (1)$$

$frequency_{\{i,j\}}$ is the number of word occurrences in a document (term frequency); d$ocument_{freq\ \{i\}}$ represents the number of documents containing the word (document frequency); Drepresents the count of all documents; $weight\{i,j\}$ is the relative significance of the word in the document.

E. Feature and Attribute Selection

In this phase, a subgroup of the features was picked to depict a text document. The selected featuresproduced an enhanced textual description,ascompared to several features that have very few information regarding data. The number of indicator variables reducedby eliminating a list of stop words. Stemming is performed on the terms, which are converted into root form. The terms are filtered by using two parameters such as word frequency and inverse document frequency. The terms having low occurences in the corpus and lowoccurences in the each document removed.Features were elected based on classification,andeliminate the fewinsignificant attributes.

III. Result and discussion

In this section, we discuss the research article classification of the dataset used for study and articles contribution in each decade. Finally, we have presented the trend analysis in machine learning research.

A. Research Article Classification

The Gensim package is used to perform text mining on the titles and abstract of the collected articles. It based on the idea of handling on substantial unstructured text corpora, document after document, in a memory-independent fashion. Also, it implements the Vector Space Model (VSM) algorithms [17] and includes corpus transformations such as TfIdf, LSI, Random projection, etc. For experimental purpose, Gensim as a python library used for implementing the trend analysis and document streaming [18]. The articles classified into 14 machine learning areas by applying the steps discussed in Section 2. Table III shows the classification produced by the term frequencies and weights. The descriptive terms (1-14) represented as a classification labels with their corresponding terms. Also, the table shows the count of articles retrieved in each decade as per the terms identified using TfIdf model.

B. Research Contribution In Each Decade

In this subsection, we discuss the contribution of descriptive terms in each decade. Fig. 2, shows the percentage of research contribution of descriptive terms (1-14)in each decade from 1988~2017. The donut shape for each of the descriptiveterms represents the percentage contribution for decade1 (i.e., from 1988~1997), decade2 (i.e., from 1998~2007) and decade3 (i.e., 2008~2017). The research contribution in each decade is calculated using Eq. (2).

$$CED_{(t,d)} = \frac{ac_t}{\sum_{t=1}^{n} ac_t} \qquad (2)$$

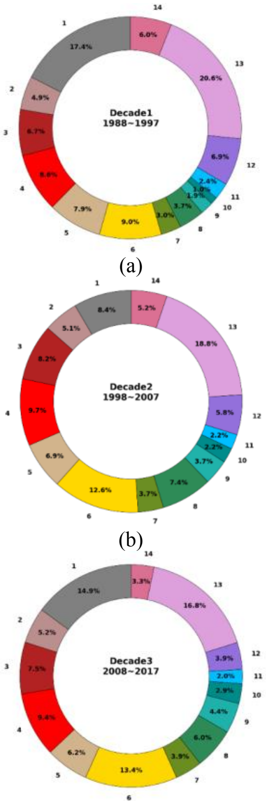
where $CED_{(t,d)}$ is the research contribution for each descriptive term t in each decade d, ac_t is the article count for each descriptive term t, n is the total number of descriptive terms and $\sum_{t=1}^{n} ac_t$ is sum of article count for all descriptive terms in each decade.

TABLE III. ARTICLE CLASSIFICATION WITH TERMS FROM 1988~2017

S. No.	Descriptive Terms	Terms	1988~ 1997	1998~ 2007	2008~ 2017
1	Artificial Neural Network and Deep learning	belief, boltzmann, convolutional, deep, forward, learning, logic, network, propagation, recurrent	1459	1052	2703
2	Bayesian statistics	base, knowledge, average, bayesian, dependence, estimator, gaussian, multinomial, naive, network	410	635	945
3	Classifiers	binary, classifier, discriminant, hierarchical, linear, machine, multi, naive, probability, support	560	1020	1363
4	Cluster analysis	birch, dbscan, fuzzy, hierarchical, mean, algorithm, cluster, group, optics, expectation	721	1214	1711
5	Decision tree algorithm	c4.5, c5.0, decision, detect, id3, iterative, random, sliq, stump, tree	659	860	1135
6	Dimensionality reduction	component, correlation, discriminant, extraction, factor, feature, least, mapping, principal, stochastic	754	1572	2432
7	Ensemble learning	ada, aggregate, average, boost, ensemble, forest, gradient, machine, random, tree	248	464	714
8	Instance-based learning	algorithm, base, learn, map, near, object, organize, quant, self, vector	309	926	1082
9	Regression Analysis	adapt, least, linear, logistic, multi, ordinary, regression, spline, step, variable	158	457	800
10	Regularization algorithm	absolute, angle, elastic, least, net, operator, regression, ridge, select, square	88	271	526
11	Reinforcement learning	action, advance, algorithm, automata, difference, learn, prior, reward, state, temporal	201	269	372
12	Semi-supervised learning	active, density, generate, graph, learn, method, model, separate, train, trans	577	723	713
13	Supervised	algorithm, annova, boost,	1720	2347	3052

	learning	classify, hidden, learn, model, near, support, target			
14	Unsupervised learning	expect, maximize, algorithm, generate, map, method, text, mine, group, vector	503	649	600

The percentage contribution of each descriptive term's result shown in the donut. The top five research areas in decade1 were the supervised learning, artificial neural network, dimensionality reduction, cluster analysis, and decision tree algorithm. Similarly, in deacde2 the top five areas were supervised learning, dimensionality reduction, cluster analysis, artificial neural network, and classifiers. Finally, in the decade3top, five areas were supervised learning, artificial neural network and deep learning, dimensionality reduction, cluster analysis, and classifiers.



(a)

(b)

(c)

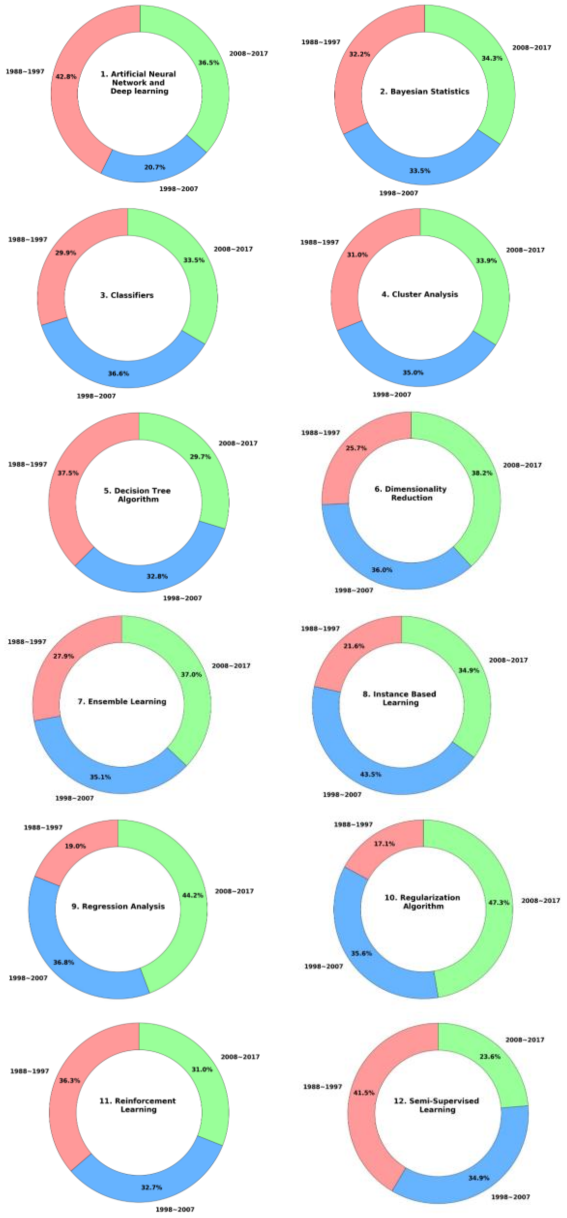
Fig. 2. (a-c) Percentage of research contributed for each descriptive terms in each decade

C. Research Contribution Across Decades

In this subsection, we discuss the contribution of descriptive terms across decades. Fig. 3, shows the percentage of research contribution of descriptive terms (1-14) across decade from 1988~2017. The donut shape for each descriptive termsrepresents the percentage contribution in red color for decade1 (i.e., from 1988~1997), blue color for decade2 (i.e., from 1998~2007) and green color for decade3 (i.e.,2008~2017). The research contribution across decades calculated using Eq. (3).

$$CAD_{(t,d)} = \frac{CED_{(t,d)}}{\sum_{k=1}^{d} CED_{(t,k)}} \quad (3)$$

where $CAD_{(t,d)}$ is the research contribution for each descriptive term t across decade d, $CED_{(t,d)}$ is the research contribution for each descriptive term t in each decade d, and $\sum_{k=1}^{d} CED_{(t,k)}$ is sum of each decade contribution for descriptive term t.



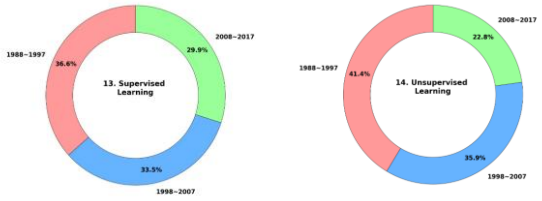


Fig. 3. Percentage of research contributed for each descriptive terms across decades

The percentage contribution of each descriptive terms results shown as the top five research areas across decade1 was the artificial neural network, semi-supervised learning, unsupervised learning, decision tree algorithm, and supervised learning. Similarly, across deacde2 the top five areas were instance-based learning, regression analysis, classifiers, dimensionality reduction, and unsupervised learning. Finally, across the decade3top, five areas are regularization algorithm, regression analysis, dimensionality reduction, ensemble learning, and deep learning.

D. Trend Analysis

Trend analysis employed for detecting trends in research articlesaccumulated over a period [13].It is an essential task to uncover trends from large volume of textual data in several fields [2,19]. The appearances of specific terms across the two decades are used to understand the trends and research patterns of research areas under study. The research trends representedby two figures for excellent visibility of each research areas. Fig. 4, and Fig. 5, show the research trends of descriptive terms (1-7) and (8-14) respectively. The research trends of artificial intelligence were very high in decade1 due to the advent of a backpropagation algorithm for training the neural network which went down in decade2 due to deficiency of computational resources and further rose in decade3 due to the evolution of deep learning and availability of high computation graphical processing unit (GPU).

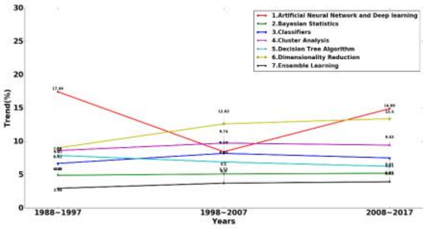


Fig. 4. Trend analysis of descriptive terms (1-7) in each decade

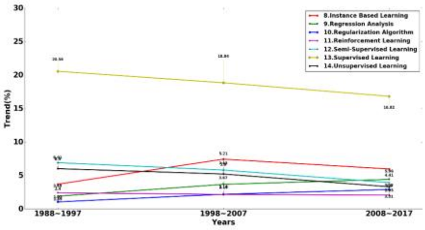


Fig. 5. Trend analysis of descriptive terms (8-14) in each decade

The following research areas showed the consistent increase in their research trends since decade1 is Bayesian statistics, cluster analysis, dimensionalityreduction, ensemble learning, regression analysis, and regularization algorithm.Also, the research areas that showed the consistent decrease in their research trends since decade1 are decision tree algorithm, reinforcement learning, semi-supervised learning, supervised learning and unsupervised learning. Finally, the research areas which showed increasing trend from decade1 to decade 2 and decreasing trends in next decade are classifiers and instance-based learning.

Fig. 6, shows the percentage increase of each descriptive term between decade1 (1988~1997) and decade2 (1998~2007). During this period, the research in artificial neural network area decreased significantly. The research in regularization algorithm,instance-based learning, and regression analysis increased substantially in decade2 as compared to decade1.

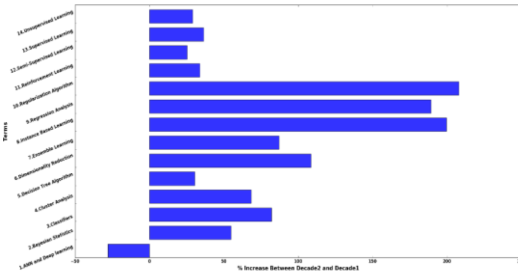


Fig. 6. Percentage increase of each descriptive term between Decade2 and Decade1

Fig. 7, shows the percentage increase of each descriptive term between decade2 (1998~2007) and decade3 (2008~2017).

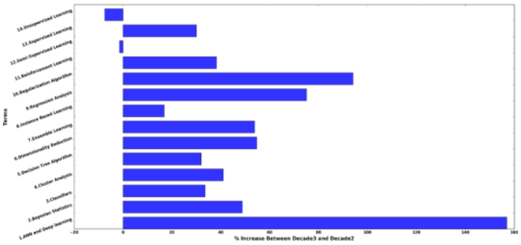


Fig. 7. Percentage increase of each descriptive term between Decade3 and Decade2

During this period, the research in unsupervised learning, and semi-supervised learning decreased significantly. The studyof artificial neural network showed the highestincrease amongst rest of the descriptive terms during this period. Fig. 8, showed the percentage increase of each descriptive term over three decades. Each descriptive term increased significantly from earlier decades, but the regularization algorithm and regression analysis showed the highestrise in decade3.

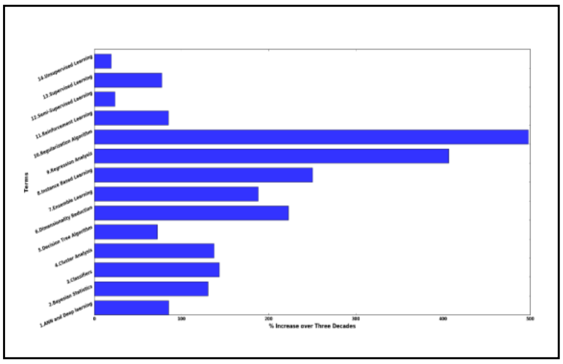


Fig. 8. Percentage increase of each descriptive term over three decades

Thus, this section concludes the result analysis of machine learning research trends using text mining.

IV. CONCLUSION

In this paper, text mining technique is utilized to perform the trend analysis in the research area of machine learning in three decades. The content collection is prepared from the published research articles in sixwell-established journals.Torealize the scholar'sresearch interest in descriptive termsover previous30 years, the datasetsplit into three sets for the time span of 1988~1997, 1998~2007, 2008~2017.This study can be useful to the upcoming researchers in the area of machine learning to get an intuition of trends to their area of interest. This framework can also be applied to identifytrends in research areasassociated to other field of study.

REFERENCES

[1] J.-L. Hung and K. Zhang, “Examining mobile learning trends 2003–2008: a categorical meta-trend analysis using text mining techniques,” *Journal of Computing in Higher Education*, vol. 24, no. 1, pp. 1–17, Oct. 2011.

[2] A. Kao and S. R. Poteet, *Natural language processing and text mining*. London: Springer, 2010.

[3] I. O. R. Patterns and T. T. Mining, “Identification of Research Patterns and Trends through Text Mining,” *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 233–235, 2012.

[4] S. Lee *et al*., "Using Patent Information for New Product Development: Keyword-Based Technology Roadmapping Approach," *2006 Technology Management for the Global Future - PICMET 2006 Conference*, Istanbul, 2006, pp. 1496-1502.

[5] A. Balahur and R. Steinberger, “Rethinking Sentiment Analysis in the News: from Theory to Practice and back,” *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis,* University of Sevilla, pp. 1-12, 2009.

[6] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications,” *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60-76, 2009.

[7] Louise Francis, and Matt Flynn, “Text Mining Handbook,” *Casualty Actuarial Society E-Forum*, Spring, pp. 1-61, 2006.

[8] Louise Francis, “Taming Text: An Introduction to Text Mining,” *Casualty Actuarial Society Forum*, Winter, pp. 51-88, 2010.

[9] P. Cerrito, “Inside text mining. Text mining provides a powerful diagnosis of hospital quality rankings,” *Health management technology*, vol. 25, no. 3, pp. 28-31,2004.

[10] M.-J. Kim, K. Ohk, and C.-S. Moon, “Trend Analysis by Using Text Mining of Journal Articles Regarding Consumer Policy,” *New Physics: Sae Mulli*, vol. 67, no. 5, pp. 555–561, 2017.

[11] A. Amado, P. Cortez, P. Rita, and S. Moro, “Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis,” *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1–7, 2018.

[12] A. K.Ojo and A. B. Adeyemo, “Knowledge Discovery In Academic Electronic Resources Using Text Mining,” *International Journal of Computer Science and Information Security*, vol. 11, no. 2, pp. 1-10, 2013.

[13] Z. Shaik, S. Garia and G. Chakraborty, “SAS® Since 1976: An Application of Text Mining to Reveal Trends,” *Proceedings of the SAS.Global Forum 2012 Conference*, Data Mining and Text Analytics, SAS Institute Inc., Cary.

[14] S. Bird, “NLTK: the natural language toolkit,” *In Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, pp. 69-72, 2006.

[15] Available for download from ftp://ftp.cs.cornell.edu/pub/smart/english.stop.

[16] M. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp.130-137, 1980.

[17] G. Salton, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.

[18] R. Řehůřek and P. Sojka, “Software Framework for Topic Modeling with Large Corpora,” *In Proceedings of LREC workshop New Challenges for NLP Frameworks*, Valletta, Malta: University of Malta, pp. 46-50, 2010.