Topic Modeling and Classification of Cyberspace Papers Using Text Mining

Sohrabi, Babak; Vanani, Iman Raeesi; Shineh, Mohsen Baranizade

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Sohrabi, B., Vanani, I. R., & Shineh, M. B. (2018). Topic Modeling and Classification of Cyberspace Papers Using Text Mining. *Journal of Cyberspace Studies*, *2*(1), 103-125. https://doi.org/10.22059/jcss.2017.239847.1009

Topic Modeling and Classification of Cyberspace Papers Using Text Mining

Babak Sohrabi *
Iman Raeesi Vanani
Mohsen Baranizade Shineh

(Received 15 August 2017; accepted 18 November 2017)

Abstract
The global cyberspace networks provide individuals with platforms to can interact, exchange ideas, share information, provide social support, conduct business, create artistic media, play games, engage in political discussions, and many more. The term *cyberspace* has become a conventional means to describe anything associated with the Internet and the diverse Internet culture. In fact, cyberspace is an umbrella term that covers all issues occurring through the interaction of information systems and humans over these networks. Deep evaluation of the scientific articles on the cyberspace domain provides concentrated knowledge and insights about major trends of the field. Text mining tools and techniques enable the practitioners and scholars to discover significant trends in a large set of internationally validated papers. This study utilizes text mining algorithms to extract, validate, and analyze 1860 scientific articles on the cyberspace domain and provides insight over the future scientific directions or cyberspace studies.

Keywords: cyberspace, text Mining, trend discovery, topic modeling.

Babak Sohrabi: (corresponding author) Professor, Department of IT Management, Faculty of Management, University of Tehran (UT), Tehran, Iran- Email: bsohrabi@ut.ac.ir
Iman Raeesi Vanani: Assistant Professor of Industrial Management, Allameh Tabataba'i University (ATU), Tehran, Iran- Email: imanraeesi@atu.ac.ir
Mohsen Baranizade Shineh: Master of IT Management, Faculty of Management, University of Tehran (UT), Tehran, Iran- Email: baranizade.mohsen@gmail.com

Introduction

In recent years, the term "cyber" is used to describe relating concepts regarding the computers and networks especially in the business and security fields. This area of knowledge investigates on many different subjects of cyberspace such as business opportunities, social media, state-on-state cyber policies and warfare, cyber terrorism, cyber militias, cyber communications, etc. Every once in a while, a new term comes along or an old term gets a novel meaning and suddenly it is everywhere. The word "cyber" has become popular creating a long list of new words. Examples of terms that have surfaced in academic papers include cyber society, cyber-attack, offensive cyber capability, cyber defense, cyber warfare, cyber-crime, cyber terrorism, etc. They are all related to the concept of *cyberspace*, which is often the presumed context or environment for the cyber concept in question. Cyber (cyberspace in general) can refer to a variety of subjects. Evaluating the scholarly articles can lead to finding out these issues (Ottis & Lorents, 2010).

For the analysis of numerical and textual data, there has been many attempts in the recent literature (Sohrabi et al., 2017; Raeesi Vanani & Jalali, 2017). For the purpose of integrating the extracted knowledge and intellectual findings from various numerical and textual sources in order to create new knowledge and insight, there has also been many research endeavors which has yielded fruitful results (Sohrabi et al., 2012; Sohrabi et al., 2010; Shafia et al., 2009; Akhgar et al., 2012).

In this study a variety of text mining techniques are investigated in order to extract useful knowledge from cyber articles. The dataset was collected from the Scopus repository in order to evaluate the major trends of cyber articles. Scopus is the world's largest abstract and citation database of peer-reviewed research literature, with over 22,000 titles from more than 5,000 international publishers[1]. This dataset contains abstracts, titles and keywords of cyberspace articles. The articles were limited to the last five years and the county that was more active in this area of knowledge. Then, different text mining techniques were applied to reach the objectives of the research. This study investigates on abstract and key words of cyberspace articles to evaluate the cyber articles to recognize the patterns and the major trends of this filed. After pre-processing and feature extraction, Latent Dirichlet Allocation (LDA) algorithm was used to distinguish the different topics in the cyberspace domain. Ultimately, different classifications of algorithms such as SVM were employed to predict the area of knowledge of a given article.

[1]. https://www.scopus.com/freelookup/form/author.uri

Literature Review

The Cyberspace Definition

As the overview of definitions showed, there is no common definition for cyberspace and the ones that are used are often vague or missing key components. The definitions do not properly address the dynamic nature of cyberspace. In order to correct this issue, the following definition is proposed: "*cyberspace is a time-dependent set of interconnected information systems and the human users that interact with these systems*" (Ottis & Lorents, 2010). Interconnected information systems include the information, the hardware, the software and the media that connects them. A convenient way to model such systems is to use graphs, where nodes represent computers, networking devices, sensors, user interfaces etc. and edges represent connections between the nodes (cables, radio links, etc.). In this definition, human users have also been included. Cyberspace is an artificial space, created by humans for human purposes. Without human users, cyberspace would become stagnate, fall into disrepair and eventually cease to exist. Unless something else can take over the maintenance and development of the cyber infrastructure and content, the human remains an important part of cyberspace (Ottis & Lorents, 2010). Considering the number of nodes in the global network, it becomes clear why cyberspace is considered "unthinkably complex". The International Telecommunication Union (ITU)[1] estimates that nearly a quarter of the world's population is using the Internet, while over 60% are using smart phones (ITU, 2009). Supporting the user side is the core infrastructure of these networks as well as the countless service providers that allow people to communicate, shop, play, work and to live online.

However, the complexity increases even further if you consider that this network is not static. Both the elements and the relations between elements can change (or remain unchanged) in time-dependent sets and systems as the time progresses. In cyberspace, this means that users, nodes and connections can appear and disappear, and information is transformed over time. Compared to other time-dependent systems, dramatic changes can take place in extremely short periods of time in cyberspace. For example, a piece of malicious code can replicate, infect and effectively disable large parts of a global network in a matter of seconds or minutes (Ottis & Lorents, 2010).

Text Mining Tools and Techniques

Text mining (TM) is a very recent and increasingly interesting area of research that seeks to use automated techniques to investigate a high-

[1]. http://handle.itu.int/11.1002/pub/80b7079c-ar?locatt=id:0

level of information from huge amounts of textual data and present it in a useful form to its potential users (Choudhary et al., 2009). TM or text analytics refers to the application of a variety of techniques to extract useful information from document collections. In general, TM tasks can be categorized to text classification, text clustering, text summarization, topic identification and association of rules (Kumar & Ravi, 2016). More specifically, TM tasks include assigning texts to one or more categories (text categorization), grouping similar texts together (text clustering), finding the subject of discussion (concept/entity extraction), finding the tone of a text (sentiment analysis), summarizing documents, and learning the relations between entities described in a text (Entity Relationship Modelling) (Truyens & Van Eecke, 2014).

Dealing with textual data requires a lot of effort. Most statistical analysis and machine learning focus on numerical data types which mean these methods are not appropriate for textual documents. Therefore, documents need to be converted to structured data that are suitable for analytical models (Jun et al., 2014).

There are various ways to convert document-term matrices based on weights. Currently researchers can use this type of text analysis through different approaches such as Binary Terms (BT), Terms Frequency (TF) and, Term Frequency-Inverse Document Frequency (TF-IDF) (Kumar & Ravi, 2016). Before applying the underlying feature extraction, it is important to utilize text normalization techniques. Text normalization is defined as a process that consists of a series of steps to wrangle, clean, and standardize textual data into a form that could be consumed by other Natural Language Processing (NLP) and analytic systems and applications as inputs. Often, tokenization is also a part of text normalization. Besides tokenization, various other techniques including cleaning text, case conversion, correcting spellings, removing stop words and other unnecessary terms, stemming, and lemmatization are being used for text mining and pre-processing (Sarkar, 2016).

Topic Modeling

Topic modeling is a powerful approach to analyzing a massive amount of unclassified text. A topic contains a batch of words that frequently occurs together. A topic modeling connects similar words in terms of their meanings. To have a better management approach to the explosion of electronic document archives, it requires using new techniques and analytical approaches that deals with automatically organizing, searching, indexing, and browsing large collections (Hwang et al., 2017).

The idea of topic models is to work with documents when these

documents have a variety of topics, where a topic is a probability of distribution over words. In other words, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents are generated from (Alghamdi et al., 2015). There are huge repositories of online documents, scientifically interesting blogs, news articles and literature that can be used for textual analysis. On the major trends in modeling textual data is the latent topic modeling that has become very popular in the domain of unsupervised techniques (Nallapati et al., 2008).

Text Classification

The problem of classification, concentrates on predicting various sets of records (observed data set) in a way that each record is labeled based on the learned target values of the previous records. A wide variety of classification techniques exist for quantitative or categorical data. Since text may be modeled as quantitative data with frequencies on the word attributes, it is possible to employ most classification methods for text classification Invalid source specified. After labeling the dataset based on k-means clustering, four algorithms of classification (Multinomial Naïve Bayes, Logistic Regression, Boost tree regression, and, linear Support Vector Machine (SVM)) have been utilized to classify the dataset and predict the domain of knowledge for new articles.

Research Method

This study intends to identify the different subjects (areas of knowledge) of similar articles based on their abstracts using LDA topic modeling algorithms. Each topic is considered in a specific area of knowledge and assigned an appropriate label based on content and term frequency. After labeling each topic, labeled articles are fed to classification algorithms to recognize the domain of knowledge for each article. Objectives of the research have been reached through five steps explained in the following sections.

Figure 1 shows the research framework of this study. The research contains data collection, Pre-processing, feature extraction, clustering and classification steps. Following, each step and its sub steps are explained in more details.

Data Collection

The dataset collected from the Scopus repository, using the advance search toolbar and launching cyber keywords. All articles in this area for the last five years were collected. In the next step, the dataset is

narrowed down to countries which had at least 50 articles on the cyber domain of knowledge. Figure 2 shows the number of articles in each country.

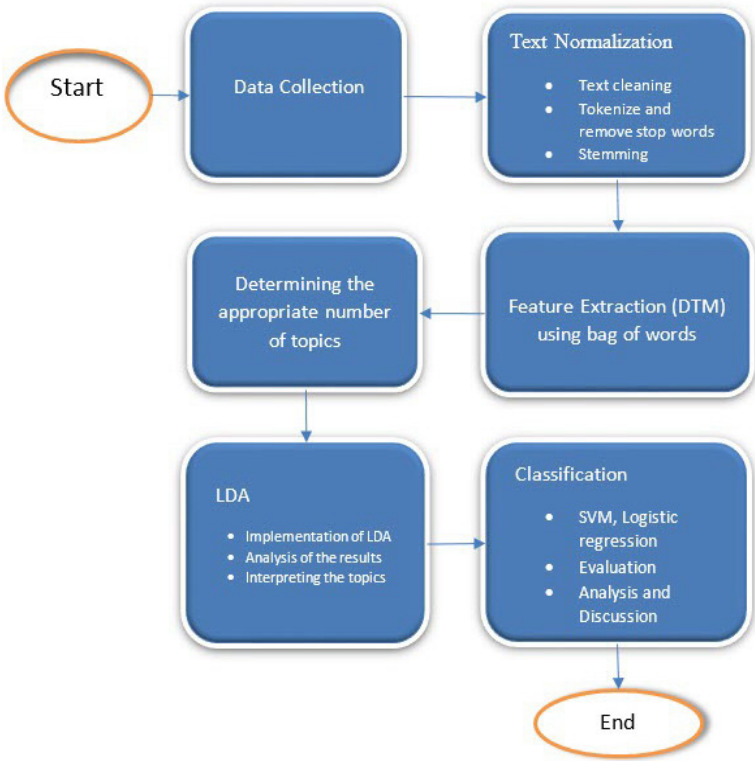

Figure 1. Framework of research

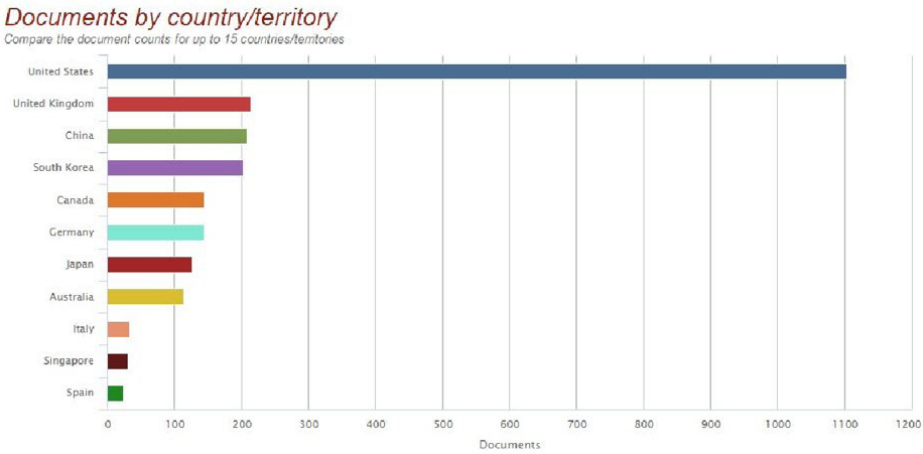

Figure 2. Number of articles for most active countries

It is observed that the United State with approximately 1100 articles has been the most active country in this area of knowledge. Figure 3 shows the number of articles by year from the most popular journals.

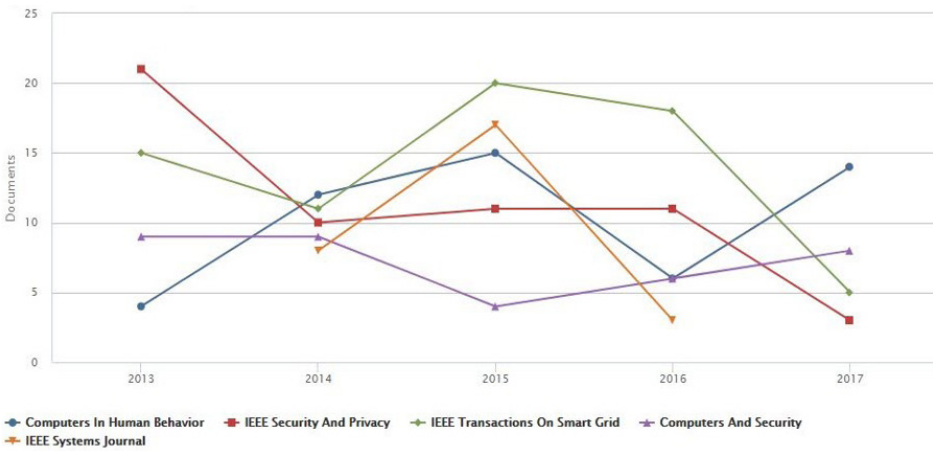

Figure 3. Most active journals in five years

After the data collection, the dataset investigated the missing values. It observed that 21 articles had no abstracts. Before the data preparation step, these articles were eliminated. The total number of articles after elimination the missing values was reduced to 1839 articles from the original 1860.

Data Preparation

Data Preparation steps were applied in the following order.

Text Normalization. One of the important parts of any NLP and text analytics solution is the text pre-processing stage, since the characters, words, and sentences identified at this stage are the fundamental units passed to all further processing stages (Kannan & Gurusamy, 2014). Text pre-processing includes a variety of techniques such as stop words elimination, stemming, part of speech tagging and so on (Vijayarani et al., 2015). At this stage, the more relevant techniques to the structure of documents in the dataset were employed.

Text cleansing. Characters such as digits, HTML tags, "@" tags and all undesirable characters were deleted. All punctuations such as commas, question marks, and semicolons were also eliminated.

Tokenization and removing stop words. After tokenizing words, all the stop words were removed inform the dataset. Stop words are defined as words which usually refer to the most common words in a language. There is no universal list of stop words as stop words in different

languages are unique and specific to that language (Saini & Rakholia, 2016). Stop words listed by NLTK package in python, were used in this study as the source for stop words removal.

Stemming. Stemming is the process of conflating the variant forms of a word into a common representation, the stem. Stemming is widely used in text processing to match similar words in a text document in order to improve the efficiency and effectiveness of information retrieval (IR) systems (Kannan & Gurusamy, 2014). In this case, after using stem algorithms, for example words such as "analytics", "analytical", "analysis" will all turn into "analytic".

Feature Extraction

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier (Sivic, 2009). In this model, a text (such as a sentence or a document) is the bag (multiset) of words, disregarding grammar and word order, keeping the multiplicity. The bag-of-words model is also used for computer vision.

Determining Optimized Number of Topics

To select the optimum number of topics for a corpus is the most challenging part of the topic modeling algorithm. In fact, there is no valid method to determine the exact number of topics for a collection of documents. A number of heuristic methods were developed by the researcher for the purpose of this study. Metrics used to estimate the number of topics. Figure 4 shows the metrics for a range of 2 to 100 topics. This study used four heuristic methods which is developed by different researchers (Arun et al., 2010; Cao et al., 2009; Deveaud, 2014; Griffiths et al., 2004) to detect the number of topic in advance.

It is obvious that all the metrics cannot be used at the same time. In this case, the metrics were not informative enough because researchers offer an extended range of the optimum number of topics. For example, Griffiths calculated the optimum number of topics in the range of 60 to 80. After the implementing LDA with 60 topics, the number of topic were too much and meaningful results were not produced. In fact, the topics are so close together that many of them could have been considered as one topic.

Deveaud et al. (2014) estimated the number of topics appropriately. This metric offered the optimum number of topics between 4 to 8 topics. The maximum value of this research assigned to topics is 7. The LDA algorithm was implemented for 7 topics. Table 1 shows the most

important terms and the number of articles which have been assigned to each topic through the categorization of topics.

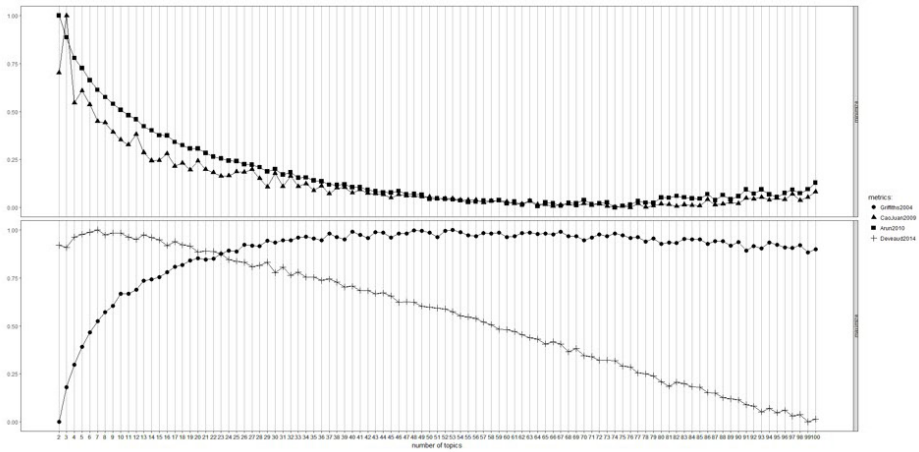

Figure 4. Determining the optimum number of topics

Table 1. Imortan terms for each topic

Topic number	Top terms	Number of articles
1	Cyber, systems, security, data, physical, based, paper, information, control, design, approach, model, software, network, attacks, smart, cps, new, research, analysis	931
2	Cyber, power, network, control, physical, data, systems, proposed, time, attacks, base, attack, grid, model, communication, problem, energy, results, networks, algorithm, performance, using, sensor, smart, real, nodes, state, information	464
3	Cyber, social, security, information, online, study, learning, internet, users, data, use, student, cybersecurity, technology, new, media, results, policy, analysis, using, education, international, findings, public, cyberspace, important, law	333
4	Detection, traffic, malware, based, approach, time, accuracy, proposed, classification, data, using, results, detect, malicious, botnet, techniques, real, darknet, detecting, used, image, activities, attacks, false, manufacturing, approaches, road, botnets, method	61
5	Cyberbullying, victimization, bullying, students, sns, adolescent, participants, online, victims, sex, traditional, results, behavior, college, self-behaviors, cyber, negative	41
6	Nist, underground, governance, Friday, systemic, terrorist, warnings, Monday, money events, insider, band, ads, governments, holiday, surveillance, consumer, small, warning, radar	3
7	Cell, air, fuel, on serve, cybercide, flood, turbine, SaaS, fps, valve, dust, cold, projection, elderly, cathode, task, speed, pressure, aneurysm,	6

Figure 5 shows the inter-topic distance map and the marginal topic distribution with other topics. Topic 2 and 3 were the major trends of the articles with a high subscription rate and other topics have an independent nature and a coherent subject in the cyber domain. The most salient terms are shown in Figure 5. Terms such as "cyber", "security", "cyberbullying", "data", "malware", and "detections" provide great insight about the dataset. In the next stage each topic is examined regarding its content in order to choose an appropriate name for them.

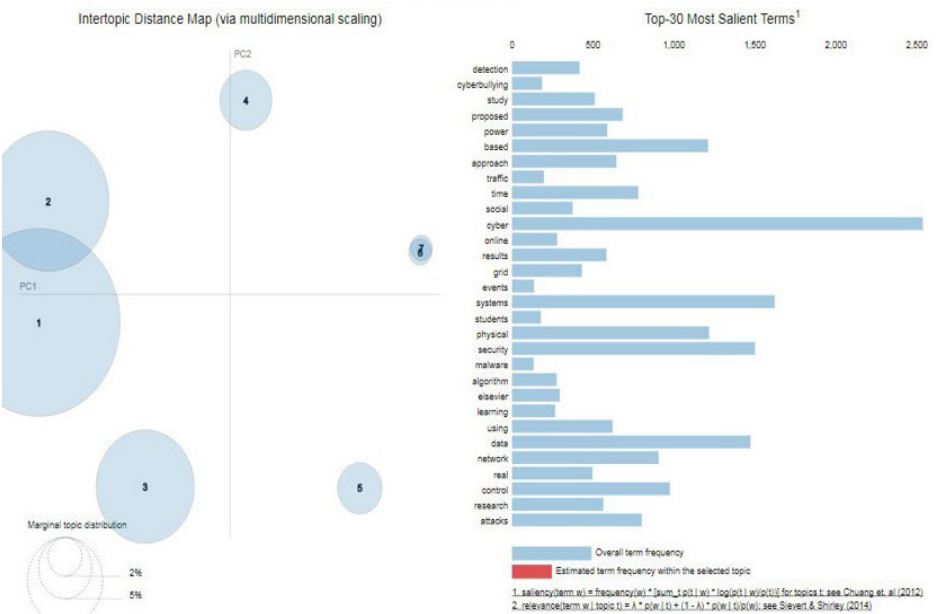

Figure 5. Inter-topic distance map for 7 topics

Major topics and their trends are analyzed and evaluated in the following section.

Topic One: Cyber, IoT, Cloud and Industry

The Internet of Things (IoT) becomes an attractive research topic, in which a real entity of the physical world becomes a virtual entity in the cyber world, and both physical and virtual entities are enhanced with sensing, processing, and self-adapting capabilities to perform interactions through a special addressing scheme. Along with the combination of Internet and modern sensor technologies such as Radio Frequency Identification (RFID), Near Field Communication (NFC), and Wireless Sensors and Actuator Networks (WSAN), IoT itself is facing rigorous security challenges. Several issues in terms of system architecture,

standards, and human involvement are raised. The following security problems seem to be intense speculations, on the design appropriate for the security framework of intelligent applications. What is advanced security technology applied into mass data processing? How to maintain a balance between high security requirements and supporting infrastructure? And how human society securely participates in both cyber and physical worlds that are interconnected? (Ning, 2012).

Cloud platforms are distance-based service provision infrastructures that ensure 24×7 "On Demand" private and public strategic alignment with regulatory and compliance priorities towards governance objectives as well as providing many detailed services to the lower levels of business operations and reducing the risk of infrastructure failure (Bhagat, 2011).

This topic as a major topic of the dataset has discussed state-of-art issues such as IoT and cloud computing in their relation with cyber subjects. All subjects investigated business concerns on this area. Terms such as "cloud", "computing", "industrial", "platform", "IoT", "software", and "integration", which all refer to business concerns, are observed only in this topic. Because this is the major trend in cyberspace, Figure 6 with different metric adjustments is represented. For example, when λ equal to 0.2 the destitution of terms is illustrated in Figure 6 (λ is calculated as shown in the bottom of Figure 6).

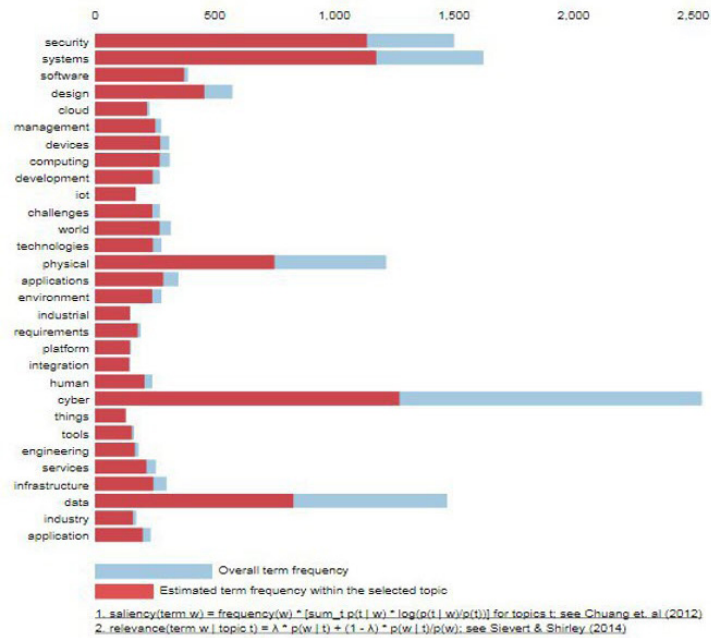

Figure 6. Terms of topic 1 for λ equal 0.2

Figures 7 and 8 illustrate the terms of topic one when λ equals to 0 and 1 respectively.

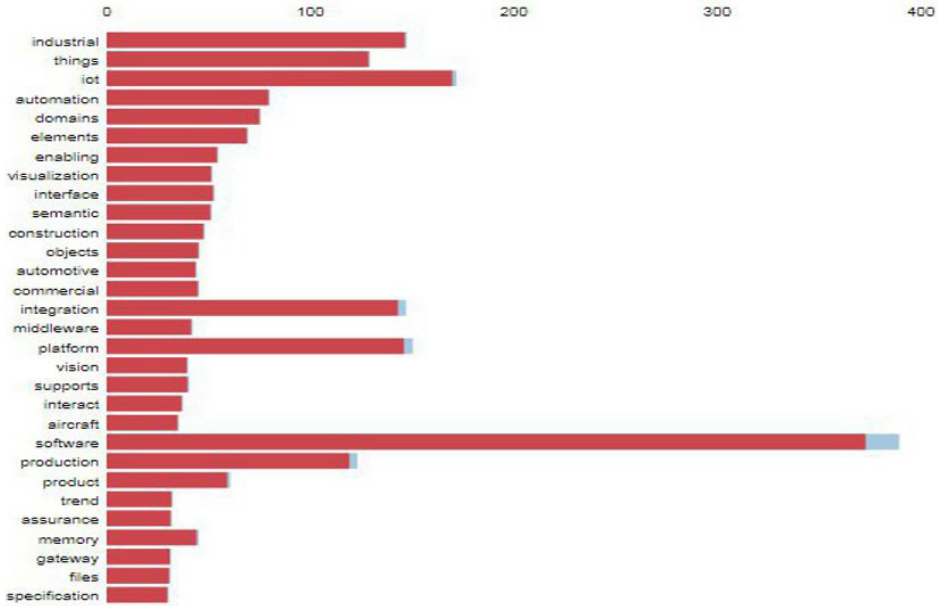

Figure 7. Most relevant terms in topic 1

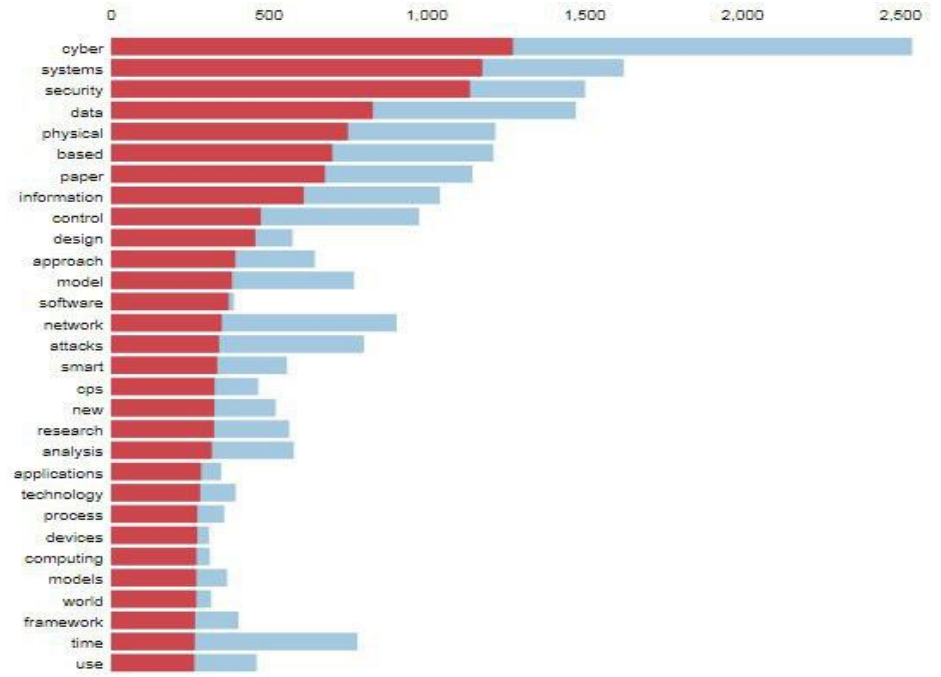

Figure 8. Terms of topic 1 with shared terms with other topic

Topic Two: Cyber-Physical Systems

The cyber-physical systems (CPSs) are the combination of computational elements and software and physical entities that can interact with humans via various techniques. The security risks includes the attempts by outside and inside adversaries that disrupts or destructs the functions of the physical systems. They are distributed and intelligent and act in real-time networking (wired or wireless).

A typical design of CPS consists of a network with many interacting elements using physical inputs and outputs. CPS provides the links among computational and physical elements to improve efficiency, usability, reliability, safety, adaptability, and full functionality (Shi et al., 2011). Topic two has been named as CPS because it has the all characteristics of CPSs. Through this topic, terms such as "attacker", "delay", "nodes", "defender", "WSNs" appeared regularly. As mentioned before topic two has too many words shared with topic one. This topic has shared terms such as "real", "time", "network", "sensor", "smart" and "data" with topic one (Figures 9 and 10).

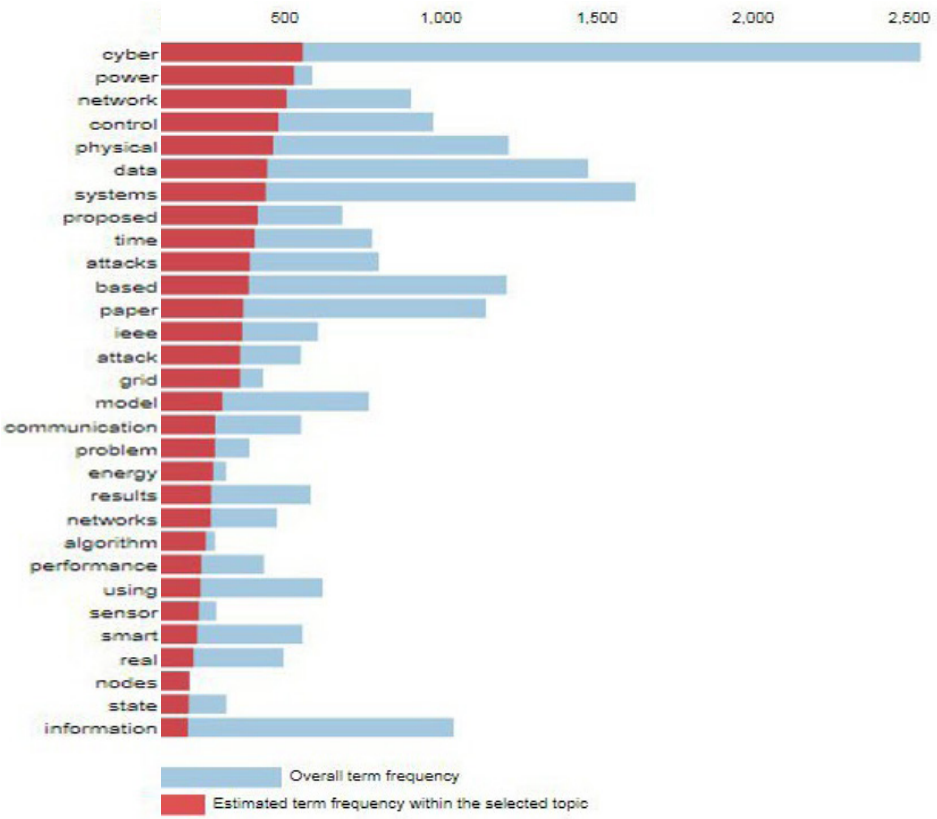

Figure 9. Terms of topic 2

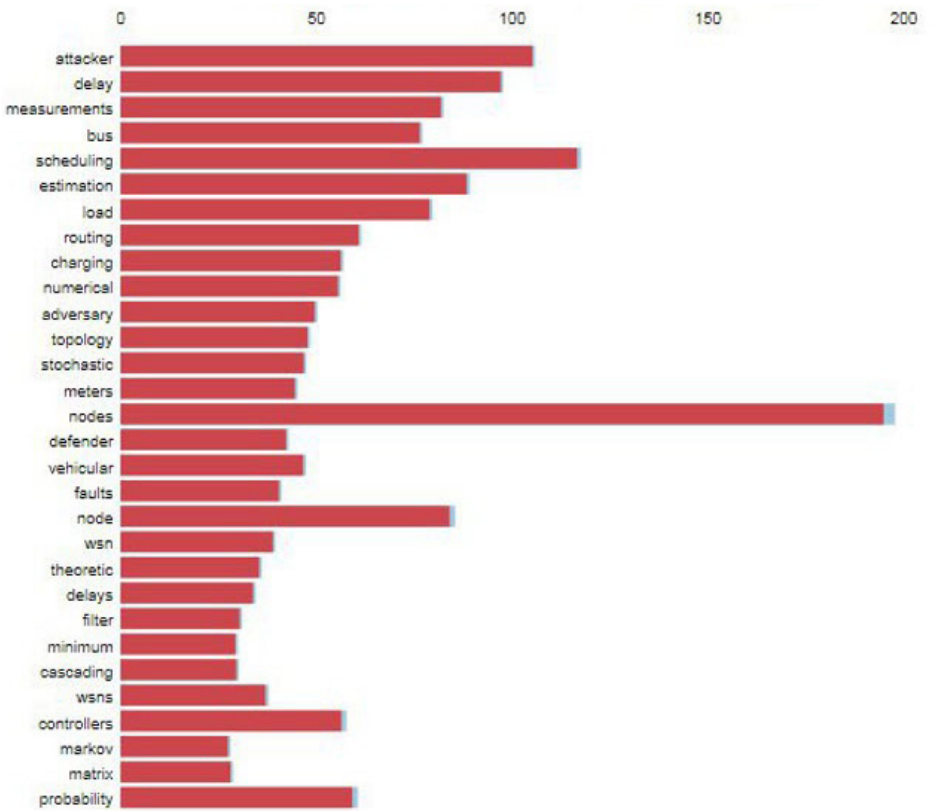

Figure 10. Most relevant terms of topic 2

Topic Three: Cyber Law and Crime

The expression *cybercrime law* is applied to signify the "legal regulation of digital crime and digital evidence." Crime and evidence are different phenomena governed by different sets of legal rules.Where as provisions of *substantive criminal law* describe the criminal offenses, rules sconcerning the collection and use of evidence in criminal investigation and prosecution are laid down in *rules of procedural (criminal) law*. There is a close factual relation between criminal offenses and evidence. The law defines criminal offenses by description of the conditions that the perpetrator must fulfill. Conviction of a crime requires proof beyond any reasonable doubt. The prosecutor has the burden of proof and must prove each condition of the crime(s) included in the criminal charge (Sunde, 2017). The terms of this topic were convinecing enough to name this topic as cyber law and crime. Figure 11 shows the most relevent terms on this topic. "Legal", "crime", "legislation", 'cybercrime", and "passwords" requlary appeared in these articles.

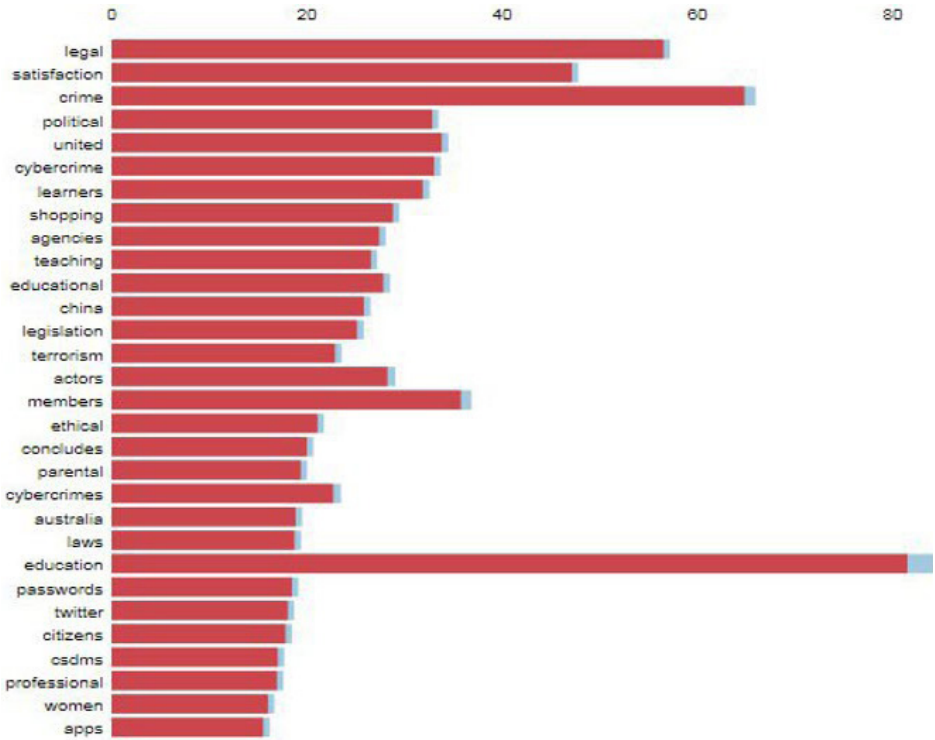

Figure 11. Most relevant terms of topic 3

This topic has shared terms such as "social", "security", "information", "finding" and "cyberspace" with topics 1, 2 and 5.

Topic Four: Cyber Malware

Malicious pieces of code (viruses, Trojans, rootkits, worms, bots, and spyware) and weaponized zero days can be deployed not only to perpetrate common cybercrime but also to engage in cyber warfare. Identity theft, online scams and fraud, and theft of intellectual property or classified information usually fall under the first category that of "common cybercrimes." Other cyber activities, depending on their scale, effect, originators, and targets, are sometimes characterized as a "cyber act of war". The truth is, however, that there is no litmus test for the distinction between the two groups of malicious activities. This topic is called cyber malware since words such as "botnet", "bot", "classifier", "detection", "malware", "dark net" have been used in these articles. This topic is the most technical topic that aims to use programing languages such as FPGA and ELM to detect any malicious devices. Figures 12 and 13 show the most relevant terms and shared terms with other topics respectively.

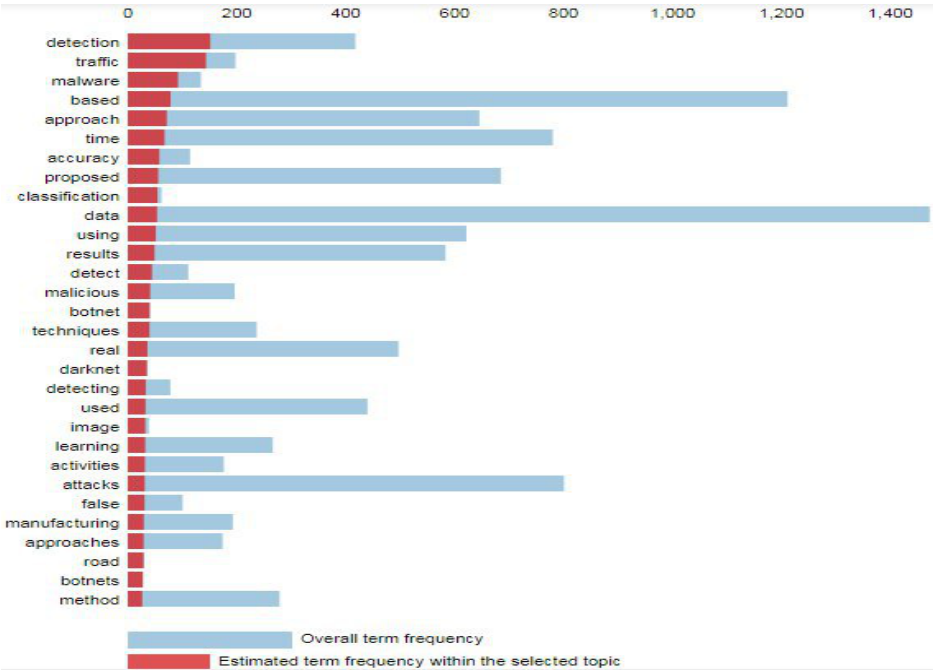

Figure 12. Terms of topic 4

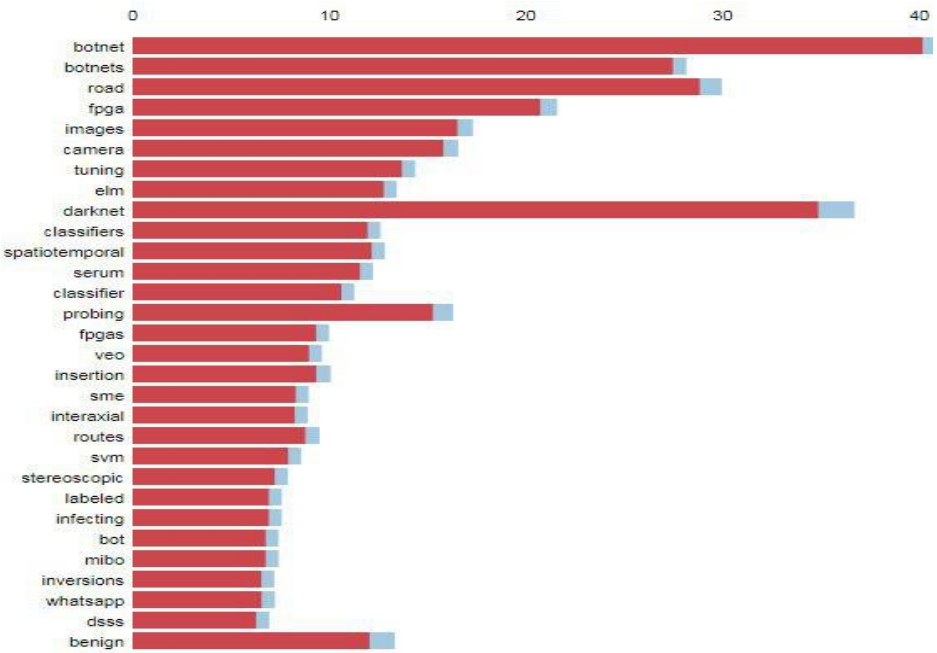

Figure 13. Most relevant terms of topic 4

Topic five: Cyberbullying

Cyberbullying has been defined as willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices (Hinduja et al., 2010). The Internet has created a whole new world of social communications, particularly for young people whose use of e-mail, websites, instant messaging, web cams, chat rooms and text messaging is exploding worldwide. While these new tools of electronic communication are widely used for positive purposes, they can also be a means through which children and adolescents bully and are bullied by their peers (Frias et al., 2017). Topic number 5 is named as cyberbullying since terms such as "cyberbullying", "bullying", "adolescents", "victimization", and "harassment", were observed frequently. Figure 14 shows the terms that appeared on this topic.

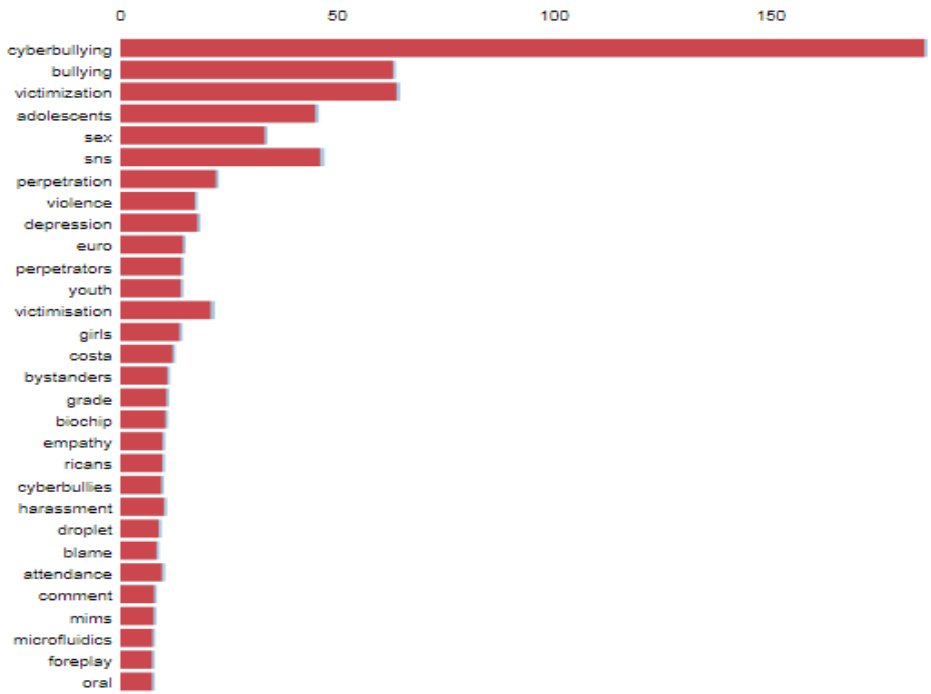

Figure 14. Most relevant terms of topic 5

Topics six and seven: General Cyberspace Concerns

The total number of articles that were assigned to these topics are nine articles (3, 6 articles respectively). These topics discussed general subjects and did not form a coherent subject or domain in cyberspace. In order to improve the result of the classification algorithms these topics were deleted.

Predictive Classification

Labeling the articles based on probabilistic LDA is the most challenging part of this study. Since probabilistic LDA allocated (percent of possibility) to a document over different topics, therefore, a specific document can belong to a variety of subjects. For example, articles with the index 1166 relate to different topics however, the most related topic for this article is topic number 3. Figure 15 shows the probabilistic distribution for some articles over the five major topics.

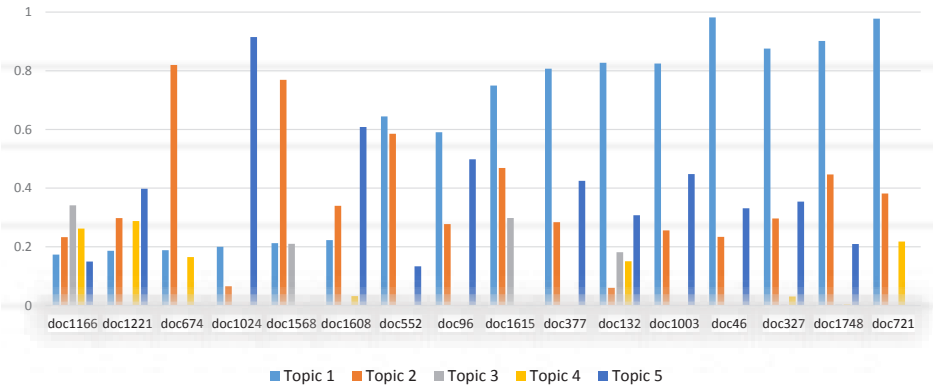

Figure 15. Possibility distribution

This study labeled articles based on the most value of topic possibility for a specific article. For example, articles with index number of 1166 are labeled as cyber law and crime (topic 3).

In this section, classification algorithms were implemented for learning the results of topic and then, the results of two algorithms are examined and validated. The results of using algorithms were examined with different metrics such as precision. Boost tree algorithms have predicted the target label more accurately. Table 2 has summarized the results of the two algorithms (namely boost tree and logistic regression) with precision, recall and accuracy.

Table 2. Result summery of classification algorithms

Classification Algorithms	Accuracy	Precision	Recall
Boos Tree	0.82	0.85	0.81
Logistic Regression	0.81	0.79	0.78

To detect the main accruing errors in confusion matrix for both algorithms are evaluated in order to find out why and how the algorithms are filed to predict the targeted label. The figure shows the confusion matrix for the boost tree classifier. 13 articles that have IoT

& cloud labels are predicted as physical-systems. On the other hand, 34 articles with physical-system labels predicted as IoT & cloud label. As mentioned topic one and two (IoT, Cloud and Industry and Cyber physical-system) have a lot of terms and key phrases in common. IoT & cloud, and law & crime labels were predicted exchangeable by the boost tree algorithms. After examining the terms on these two topics we can see that terms such as "security", "information", "online", "internet", and "cybersecurity" are common within these topics however they may refer to different concepts (Table 3).

Table 3. Confusion matrix for boost tree algorithm

Confusion Matrix

		Predicted Label					
		cyberbullying	general	iot&cloud	law&crime	malware	physical-system
Target Label	cyberbullying	0	0	4	0	0	1
	general	0	0	2	1	0	1
	iot&cloud	2	0	185	9	0	13
	law&crime	0	0	13	31	3	1
	malware	0	0	1	4	4	0
	physical-system	0	0	34	2	0	62

The figure shows how "cybersecurity" spread in the two topics. Cybersecurity plays an important role in law cases for both individuals and organization security needs (Figure 16).

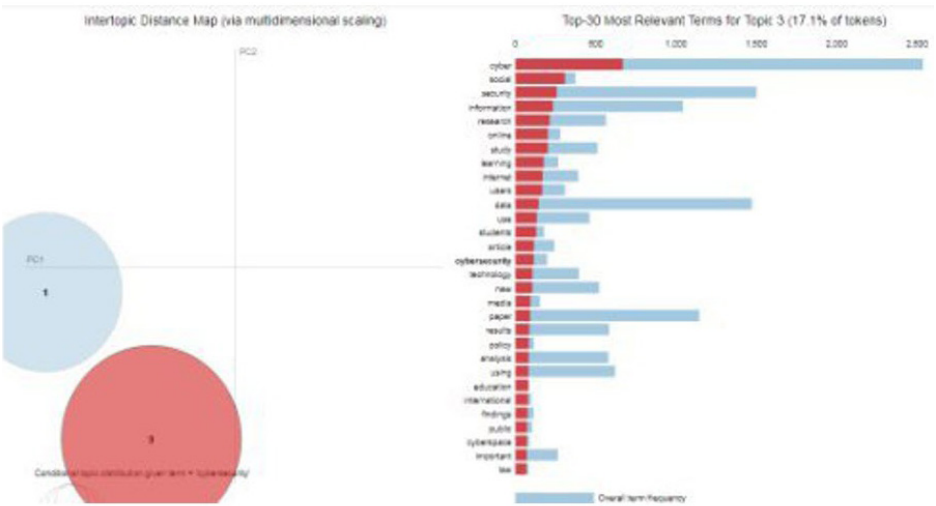

Figure 16. Distribution of cyberspace word over topic 1 and 3

Reasons mentioned above can influence the results of algorithm classifications. The same issues can be used for logistic regression. The validation results are shown in Table 4.

Table 4. Confusion matrix for logistic regression algorithms

Accuracy : 0.809651474531

Confusion Matrix :

target_label	predicted_label	count
physical-system	iot&cloud	11
iot&cloud	law&crime	1
physical-system	law&crime	1
malware	malware	8
law&crime	physical-system	2
malware	iot&cloud	1
law&crime	law&crime	23
general	law&crime	1
law&crime	iot&cloud	17
law&crime	malware	6

The errors occurred in IoT & cloud, physical-systems and law & crime categories but did not affect the final predictive results. The learning process makes the algorithm capable of extracting and predicting the usefulness of the content of the articles based on the patterns learnt from previous article evaluations. Thus, it can find the best suitable articles for new research and developments.

Common words of each cluster and their frequencies are shown in Table 5.

Table 5. Term frequency for each topic

Topic Number	topic Name	Common Terms (Frequency)
one	Cyber, IoT, Cloud and Industry	Cyber (1300), cloud (361), computing (373), iot (211), things (125), smart (356), data (700), industrial (149), middleware (45), automotive (51), platform (50), integration (51), infrastructure (67)
Two	Physical systems	Attacker (108), nodes (156), topology (52), defender (45), WSNS (48), controller (49), algorithms (250), sensor (198), physical (398), based (289)
Three	Law and crime	Legal (58), crime (65), cybercrime (38), legislation (21), passwords (19), social (301), law (69), hackers (21), user (190), information (250), internet (233)
Four	Cyber malware	Botnet(s) (65), dark net (33), classifier (10), tuning (12), detection (120), classification (59), malware (98), malicious (51)
Five	Cyberbullying	Cyberbullying (186), victimization (59), bullying (58), sex (33), gender (37), depression (28), adolescent (39), violence (15), aggression (16), negative (28), harassment (14)

Conclusion

Investigating the words and key phrases for each area of knowledge, identify five major concerns in cyberspace domains of research. Cyber, IoT, Cloud and Industry papers mostly invest on platforms and business systems. Cloud based and IoT solutions in industry provide clues on why words and key phrases such as "platform", "IoT", "cloud", and "innovation" "Cyber", "systems", "security", "data", "software", "network" were frequently used in articles. While topic one invests on IoT and cloud computing solutions in industry, topic two (Cyber-Physical Systems) concentrates on cyber-physical systems (CPSs) that are interconnected physical and computational systems used for gathering intelligence throughout a fully functional system so as to secure the processing environment and enforcing the regulations.

Cyber law and crime (topic 3) are also important in cyberspace. The legal regulations developed to deal with cyberspace law cases. Terms such as "legislation", "crime", "law", and "cybercrime" were repeatedly used in these articles. However, topic 5 (cyberbullying) explains how willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices can influence the life of individuals, particularly the youth and students.

Cyber Malware (topic 4) evaluate the safekeeping of business systems from malware and botnets and detecting them using classification algorithms in real-time. The terms "botnet", "dark net", "malicious", "attack", "detection", and "classification" are used in this topic area.

References

Akhgar, B., Rasouli, H., & Raeesi Vanani, I. (2012). Evaluation of knowledge-based competency in Iranian universities: a practical model. *International Journal of Knowledge and Learning*, *8*(3-4), 282-297.

Alghamdi, R., Alfalqi, Kh. (2015). A survey of topic modeling in text mining. *I. J. ACSA,* 1, 147-153.

Arun, R., Suresh, V., Madhavan, C.V., & Murthy, M.N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Bhagat, B.C. (2011). Cloud computing governance, cyber security, risk, and compliance business rules system and method. *Google Patents*, 2-5.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7), 1775-1781.

Choudhary, A.K., Oluikpe, P.I., Harding, J.A., & Carrillo, P.M. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. *Computers in Industry, 60*, 728-740.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61-84.

Frias, S. M., & Finkelhor, D. (2017). Victimizations of Mexican youth (12–17 years old): A 2014 national survey. *Child Abuse & Neglect*, *67*, 86-97.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228-5235.

Hinduja, S., & Patchin, J.W. (2010). Bullying, cyberbullying, and suicide. *Archives of suicide research, 14*, 206-221.

Hwang, S.Y., Wei, C.P., Lee, C.H., & Chen, Y.S. (2017). Coauthorship networkbased literature recommendation with topic model. *Online Information Review*, *41*(3), 318-336.

International Telecommunication Union (ITU) (2009). Measuring the information society. *The ICT Development Index*, ISBN 92-61-12831-9.

Jun, S., Park, S.S., & Jang, D.S. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications, 41*, 3204-3212.

Kannan, S., & Gurusamy, V. (2014). Preprocessing Techniques for Text Mining. *RTRICS.* Podi: researchgate. Retrieved from https://www.researchgate.net/profile/Vairaprakash_Gurusamy/publication/273127322_Preprocessing_Techniques_for_Text_Mining/links/54f8319e0cf210398e949292.pdf.

Kumar, B., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114*, 128-147.

Nallapati, R.M., Ahmed, A.X., Eric, P., & Cohen, W.W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (542-550). New York: ACM New York, NY, USA.

Ning, H.A. (2012). Cyber-physical-social based security architecture for future internet of things. *Advances in Internet of Things, 2*, 1-2.

Ottis, R., & Lorents, P. (2010). Cyberspace: Definition and Implications. In Proceedings of the *5th International Conference on Information Warfare and Security*, Dayton, OH, US, 8-9 April. Reading: Academic Publishing Limited, pp. 267-270.

Ottis, R.A. (2010) International Conference on Cyber Warfare and Security (p. 267). Sonning Common: Academic Conferences International Limited.

Raeesi Vanani, I., & Jalali, S.M.J. (2017). Analytical evaluation of emerging scientific trends in business intelligence through the utilisation of burst detection algorithm. *International Journal of Bibliometrics in Business and Management*, *1*(1), 70-79.

Saini, J.R., & Rakholia, R.M. (2016). On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. *Procedia Computer Science, 89*, 313-319.

Sarkar, D. (2016). Processing and Understanding Text. In D. Sarkar, *Text Analytics with Python* (167-215). New York, United States: Apress. doi:10.1007/978-1-4842-2388-8.

Shafia, M.A., Sohrabi, B., Raeesi Vanani, I., & Faghih Mirzaii, S. (2009). A model to evaluation components of intellectual capital. *Proceedings of the 2nd International Conference on Intellectual Capital*, 7-8.

Shi, J., Wan, J., Yan, H., & Suo, H. (2011). A survey of cyber-physical systems. In Wireless Communications and Signal Processing (WCSP). *International Conference on IEEE*, 1-6.

Sivic, J.A. (2009). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 591-606.

Sohrabi, B., Raeesi Vanani, I., & Baranizadeh Shineh, M. (2017). Designing a Predictive Analytics Solution for Evaluating the Scientific Trends in Information Systems Domain. *Webology*, *14*(1), 32-52.

Sohrabi, B., Raeesi Vanani, I.R., Qorbani, D., & Forte, P., (2012). An integrative view of knowledge sharing impact on e-learning quality: A model for higher education institutes. *International Journal of Enterprise Information Systems*, *8*(2), 14-29.

Sohrabi, B., Raeesi Vanani, I., & Shafia, S. (2010). An applied model for measuring the knowledge sharing capability. *Iranian Journal of Information Processing & Management*, *26*(1), 5-28.

Sunde, I.M. (2017). Cybercrime Law. *Digital Forensics*, 51-116.

Truyens, M., & Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law \& Security Review, 30*, 153-170.

Vijayarani, S., Ilamathi, M.J., & Nithya, M. (2015). Preprocessing techniques for text mining- An overview. *International Journal of Computer Science & Communication Networks, 5*, 7-16.