ISSN No:-2456-2165

Automate Data Classification in an Unstructured Data Flow using Self-Organizing Maps

Dilushinie Narmada Fernando, Dr. Lakmal Rupasinghe
Dept. of Information Systems Engineering
SLIIT Colombo, Sri Lanka

Abstract:- Nowadays, when protecting the information of an organization, professionals would consider the level of confidentiality and sensitivity of the data as a major concern. This is reflected in a manual process where ideas, decisions, and expectations of the data owners and other professionals classify data according to their perspectives. The classification of data will depend on the decisions made by humans and expose sensitive data to many users who are unauthorized to access and alter it. This research was developed to reduce the involvement of humans in making decisions on data classification and divided them into different clusters according to the level of confidentiality. The system divides documents into 3 major categories, such as confidential, sensitive, and public data, using the unsupervised self-organizing map method, which is an artificial neural network originally designed for the clustering of high-dimensional data samples onto a low-dimensional map.

Keywords:- *Information Technology, Intellectual Property, Self-Organizing Map, Information retrieval, Statistical Natural Language Processing.*

I. INTRODUCTION

The knowledge with regard to the nature of information has launched the second technological revolution, in which mental "labor" can be reduced through the use of technological devices, such as data processing systems. The "information" is the key importance for informatics theories and communication techniques. Every human being requires knowledge that is critical to the functioning of any living or inanimate entity. Information is defined as a source of data that may be obtained as values attributed to parameters, as well as knowledge, which refers to the comprehension of actual or abstract concepts, [1]. It's also a crucial quantity in fields as diverse as cybernetics, linguistics, biology, history, and theology, [2]. The information approaches people in different ways and presents itself over a wide range of appearances; such as messages, telephone conversations, computer-controlled processes, technical drawings, etc. Some information may available physically such as books, magazines, newspapers, journals, etc. But some may available in the digital format such as websites, e-books, personal blogs, news websites, organizational sites, e-journals, formal documentations etc.

Due to the significant growth in the size and complexity of document data (e.g., news, blogs, and web pages), an adequate implementation for delivering eloquent information to consumers has evolved. Meanwhile it is a necessity to semantically understand the content of the documents and the importance of the information specified. In order to understand document data it is important to categorize them based on related information. With the explosive growth of the information, discovering relevant information about anything is of the highest importance, especially in electronic documents. As a result, finding and understanding relevant information, as well as delivering meaningful information to consumers, is extremely difficult. This is impossible to perform manually, hence some automated approaches are required, [5].

II. LITERATURE REVIEW

People were constantly exposed to a large amount of data as IT technology advanced. Data in the corporate world, such as paper files, carbon copies, and filing cabinets, has already been digitized. This has resulted in an enormous proliferation of various text documents on the Internet, such as Web News, electronic books, and e-mail communications, and the challenge of organizing and navigating these papers has become increasingly crucial and urgent. Today every organization faces the problem of various information risks as they rely on computer networks, digital information, remote cloud-based storage, electronic commerce, social media, electronic mail, instant messaging, and Internet use in general. As a result, this information could be harmful to the individual or used by malevolent people as a security hazard.

In the modern environment, looking for precise information has become an extremely time-consuming task. As a result, it is vital to organize documents into categories to make document retrieval easier and more efficient. It is impossible to see what is going on without mechanical aid when classifying News articles from ten years ago. Therefore to overcome this problem the documents are organized into a set of related topics which enable the user query process, [8]. Content filtering is a key technology in Big Data analytics for examining meaningful data or excluding data including security risk. Furthermore, conventional content filtering algorithms for structured data processing are ineffective when dealing with large amounts of unstructured text data. As a result, the efficiency of filtering must be increased to extract useful data and filter content in real-time from a large data set, [9].

Security concerns on computer networks and the problems of someone, or something, getting unauthorized access to sensitive data put this information in danger in any industry that uses technology. Many organizations have been highlighted in the news for purportedly failing to secure non-public or private information in their possession, custody, or control. Annually, a large sum of money is spent to repair, reprogram, patch, re-enter, and restore information and/or systems that have been harmed by external (and sometimes

internal) threats or attacks. As per the definition, Information Risk means, the probability that nonpublic or confidential electronically stored information could be accessed and/or exploited by unauthorized parties, [11]. The requirement to protect the information in all formats, for example, customer information, financial information, medical information, etc. of organizations is increasing today more than ever.

A. Primary Reasons to protect Data

- Information Technology (IT) has recently been selected as a weapon of choice for terrorists.
- The Internet is increasingly being used for important commercial transactions. It is a general awareness among business professionals that conducting business over the Internet without taking sufficient security precautions exposes consumer and company information to fraud and theft.
- Government policies such as Gramm Leach Bliley and the Health Insurance Portability and Accountability Act (HIPAA) hold businesses accountable for establishing data privacy, access, storage, and exchange protection procedures.
- Companies spend a lot of money every year to repair, reprogram, patch, re-enter, and restore data and/or systems that have been harmed by external attacks.
- Reputational injury that occurs due to a security attack often reflects into loss of millions of dollars and there is a potential impact on the stock price.

B. Identifying Risk of Non-classified Data

When establishing proper controls, the nature of the business, the sensitivity of the information held, and the method used to access the information must all be taken into account. Many businesses perceive risk analysis and information classification, which link security measures to business needs, to be excessively costly and ineffective. Instead, they turn to information technology support organizations to determine the data that needs to be safeguarded, the level of security that has to be offered, and the technical solution that needs to be implemented [12].

Risk managers have various levels of confidence in the security of their electronic data from prying eyes, and it's often unclear how much protection is needed or how to assess the efficiency of their solutions. Any company could implement a security plan for the entire firm that would be effective for a few years. Policies for Information security, classification of information, and risk analysis are all tried and true methods for safeguarding firm data. A solution applied to secure electronic resources or information has a fairly short half-life in this setting. The ever-changing nature of digital data will determine how long any security plan can be successful, [11]. The senior management team is in charge of defining, communicating, and enforcing information security policies.

Information classification is an important aspect of a company's information assurance plan, and it delivers the most benefit to the majority of businesses. A security policy is a high-level plan outlining how security should be applied inside an organization, what activities are permissible, and what level of risk the organization is ready to take. Information security policies should be seen as the bare minimum for meeting a company's information protection responsibilities due to its non-specific structure. Risk analysis compares the value of a company's assets to the likelihood of loss threats in order to develop countermeasures that reduce risk to acceptable levels. Protection measures to reduce risk, according to this measured methodology. Because determining the value of information when it does not create revenue is difficult, many organizations find this strategy unfeasible. Classification of information is the most cost-effective technique of data security when compared to all these mechanisms since each category has its own set of requirements for information confidentiality, integrity, and availability based on qualifying information value and risk acceptance. The ability to design and explain precise information protection procedures based on inferred business values and goals is provided by information classification, [12].

The most compelling reason to classify data is to comply with regulatory requirements. For example, the Gramm-Leach-Bliley Act and the HIPAA, both require information security controls for financial and medical firms, respectively. Although information classification is not necessary as a security precaution, medical and financial data indicate additional handling requirements for their sensitivity. Information classification can provide opportunities for work and cost savings in addition to meeting legal duties and industry and consumer expectations. Establishing an information classification system demonstrates a company's commitment to safeguarding consumer data. And, when presented strategically, this could provide you with a competitive edge against organizations that don't take data security as seriously. Formalizing your company's information security requirements through information classification can enhance audit results from two perspectives: it gives auditors a benchmark against which to measure company compliance (rather than industry best practices), and it gives employees more concrete objectives to strive for, [12].

Document classification is described as the assignment of one or more predetermined categories or subjects to documents based on their content, i.e., a collection of words determines the best-suited category for this collection of words, [12]. All document classifiers have the same goal: to assign documents to one or more content categories, such as technology, entertainment, sports, politics, and so on. Any sort of text document can be classified, including traditional documents like memos and reports and also e-mails, web pages, and so on. But this must be effective, easy to understand, use and maintain. Although classification can be made according to other criteria, in this research it studies in terms of confidentiality, because this is the most common type of information classification.

C. Managing Information Classification

The following *Fig. 1*, provides good practice that classification should be done, [13].

Fig. 1: The four-step process for managing classified information

The asset owner is in charge of classifying the information, which is typically done based on the outcome of the risk assessment: the higher the value of the information (the greater the risk of a breach of confidentiality), the higher the classification level should be. For example, in a medium-scale organization, you might employ three sensitive levels and one public level of information classification, [12-14].

Classification	*Confidential levels*	*Description*
Confidential	Top	Unauthorized disclosure or dissemination could result in severe financial or reputational damage, [14].
Restricted	Medium	This is subject to controls on access, such as only allowing valid logons from a small group of staff. Must be held in such a manner that prevents unauthorized access i.e. on a system that requires a valid and appropriate user to log in before access is granted, [14].
Internal use	Lowest	Can be disclosed or disseminated by its owner to appropriate members, [14].
Public	Everyone can see	Can be disclosed or disseminated without any restrictions on content, audience or time of publication. Disclosure or dissemination of the information must not violate any applicable laws or regulations, such as privacy rules, [14].

Table 1: Information Classification levels for a mid-size organization

The current information classification mechanism executes as a manual process which involves human collaboration aids with standards, policies and procedures such as ISO 27001. This is assigned by humans who read the documents and are knowledgeable in the subject matter. But the major problem arise is inherently imprecise, since experienced people can differ on their decisions and expectations with respect to the same data, [15]. Due to the potential of improper disclosure, anyone involved in this procedure may divulge sensitive or nonpublic personal information, increasing the number of threats.

According to FBI records, Insiders are a key cause of competitors' attempts to steal sensitive data. Insider intellectual property (IP) thieves are frequently in technical roles, and they may already have a new job lined up. Insiders that are malicious usually take information that they have been granted access to. In reality, in 52% of cases, trade secrets were stolen. Business data like billing information, price lists, and other administrative data were stolen in 30% of cases, followed by source code (20%), proprietary software (14%), customer data (12%), and business plans (6%). Insiders steal IP through technical ways, but the majority of theft is uncovered by non-technical staff, [11].

D. Research Gap

Currently there are no major systems that are capable of classifying information automatically based on the confidentiality level, which is developed using unsupervised learning approaches in machine learning and neural network technologies. At the moment organizations and individuals are more interest in information risk and risk management processes, but conducting a risk analysis is an expensive and time consuming. Therefore, this system is capable to provide better solution to classify their important information securely. As mention in the previous section this is independent from different human decisions.

Data mining, information retrieval, and machine learning are all areas that deal with these issues, [6]. Machine learning has lot of algorithms for classification, clustering and regression, but according to the literature review unsupervised machine learning algorithms highly supports document classification, such as k-means clustering and Kohonen maps or Self-organizing maps. The usefulness of statistical natural language processing (SNLP) and machine learning in determining the sensitivity of unstructured text is investigated in this system. A number of tests were carried out to see how best to use existing SNLP and machine learning approaches to the task of assigning a security classification to a document while optimizing the system's parameters, [16].

III. METHODOLOGY

The existing information classification process handles manually based on human collaboration, discloses confidential or nonpublic data to unauthorized users. The research focuses on securing data from unauthorized access when classifying data. So it helps to apply necessary security controls to protect data in different levels.

The system accepts a collection of documents with an unknown content. The collection of data will be manipulate using conventional preprocessing, to transform words into their word stems and discard useless little words with less information value. Then the frequency of words will be considered to ignore repetition of words. In order to classify data Kohonen maps also known as Self-Organizing map (SOM) will use and come up with the output of classified data based on their level of confidentiality.

A. Preprocessing

Initially, the SNOWBALL stemmer was used to convert words to their stems; for example, the word 'continuing' was transformed into 'continu.' Then stopwords, or useless "small" words like "a," "about," and "are," were eliminated. As a result, words with fewer than three letters, numerals, and special characters were excluded, [16-17]. Because the indexing file size is reduced, the procedure aids in enhancing the effectiveness of text processing, [14]. After that, using the vector space model (*tf•idf* weighting for all remaining word stems), calculate the frequency of occurrence and compute document vectors for all documents or text.

Every document now has its own document vector as a result of this operation. Then the length of each document vector was reduced to 1000 of middle frequency (around the median) word stems from the complete word frequency distribution sorted in ascending order, [16-17]. Only a learning set was used to compute the document vectors in cross-validation. In order to create a realistic situation, the system was not using information about its corresponding test set, where the system learns an existing learning set and its words in advance, [15].

B. Document Classification with Self-Organizing Map

This research was focused on SOM to be use in data classification as mention in the previous section. To begin, the map nodes were labelled in some meaningful way with the training data set's class labels. The map can then classify new documents (test set samples) by mapping the labelled nodes, which represent document classes. To label the SOM with class labels, the following basic procedure was used, [16]:

- Using a training data set, built a SOM.
- Each sample from the training set should be mapped to the map.
- Determine a class for each map node based on the number of training documents of various classes that are mapped on that node. The node's class is determined by the most common document class. If more than one class has the same maximum, label the node with the document class (from the maximum classes) that is closest to the node's model vector.
- After assigning labels to the map nodes, the test set was classified by mapping each test sample and comparing the classification result provided by the map with the sample's known class label, [16].

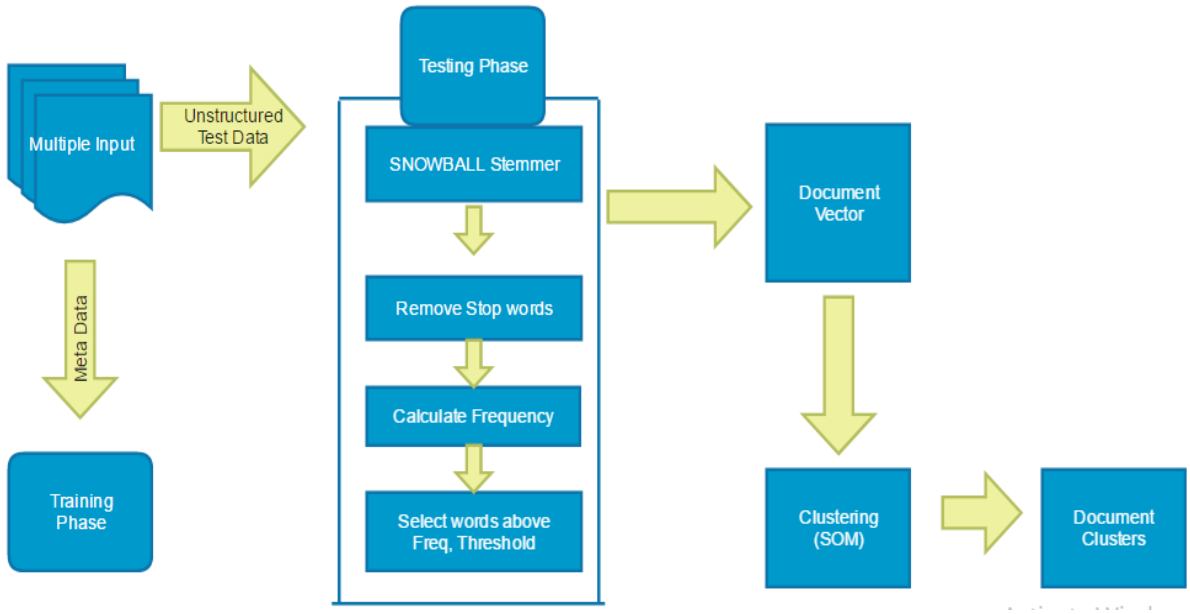

Fig. 2: System Architecture for Document Clustering

C. Data Sources and Data Collection Techniques

The existing researches has used large amount of data for experiment in order to build accurate and effective solution. According to C. Apte, the researchers used a collection of 294809 documents from CLEF 2003 from the years 1994 and 1995, [15]. The articles were from newspapers and a total of 60 test topics were included in the collection, [15].

Similarly, the system needs only 4 main topics, such as confidential, restricted, internal use and public, but for the outcome, we need to use a large data set for training the system as well as testing it. It is supposed to use a database management system to get company information, but then it raises the question of exposing the nonpublic information of a company. Therefore here we need to use some sample data that we can categorize as confidential and restricted data.

D. Graded Relevance Scale

Sormunen (2002) uses a four-level graded scale (0-3) that was originally designed for a Finnish document collection at the University of Tampere (Sormunen, 2000), [18]:

- Irrelevant: There is no information regarding the topic in the document.
- Marginally relevant: The document merely refers to the subject. It does not include any further or additional information than the topic description. Typical extent: one sentence or fact.
- Fairly relevant: The document contains more details than the topic outline, but it is not comprehensive. Only some of the sub-themes or opinions are presented in the case of a multi-faceted topic. Typical extent: one text paragraph, 2-3 sentences or facts.
- Highly relevant: The topic's concepts are thoroughly discussed in this document. All or most sub-themes or opinions are covered in the case of a multi-faceted issue. Typical extent: several text paragraphs, at least 4 sentences or facts.

A. Latent Dirichlet Allocation (LDA)

A generative probabilistic model for discrete data collections like text corpora. LDA is a three-level hierarchical Bayesian model in which each collection item is represented as a finite mixture over an underlying set of topics. Each topic is thus represented as an infinite mixture over a collection of topic probabilities. The topic probabilities give an explicit representation of a document in the context of text modelling, [25].

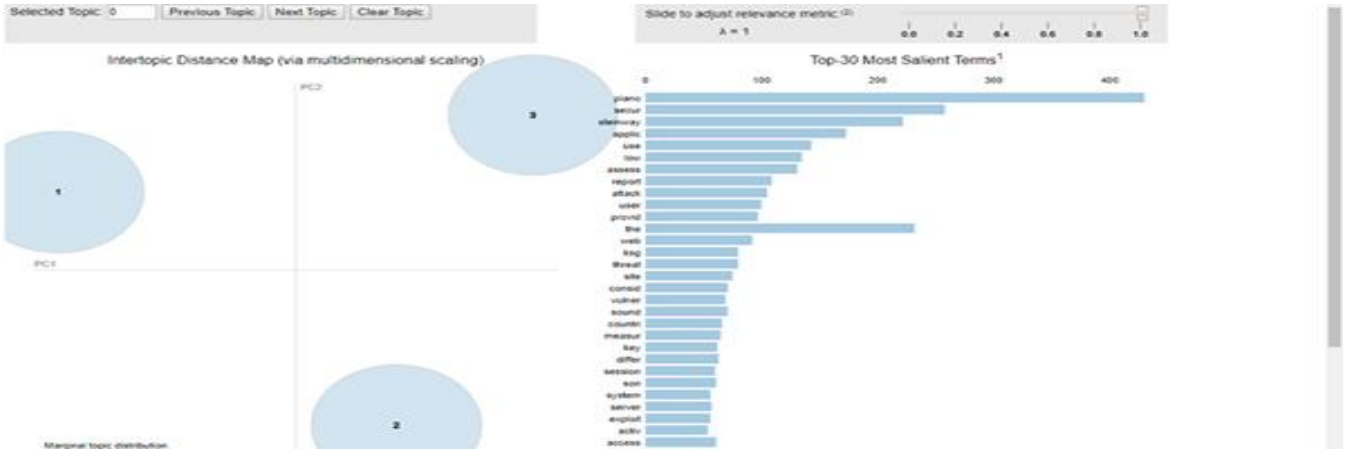

Fig. 3: How the data divided into clusters

IV. RESULTS AND DISCUSSION

Fundamental approaches like document clustering, classification, and summarization with the aid of data mining, information retrieval, and machine learning techniques are used to discover important and relevant information from documents. It is possible to navigate through papers easily once they have been sorted and pre-classified, [5-6].

A collection of documents is partitioned into separate groups called clusters in document clustering so that comparable texts are grouped together. Document clustering has been studied for use in a variety of text mining and information retrieval applications. Document clustering has been researched as a way to improve the precision or recall of information retrieval systems and as a quick approach to discover a document's closest neighbors. Clustering has been proposed for use in viewing a collection of papers or organizing search engine results returned in response to a user's query. Document clustering is often used to create hierarchical document groupings automatically, [6-7]. It produces a summary by lowering the size of documents while keeping the major properties of the source texts in a document summarizing. Existing summarization algorithms typically rank sentences in documents based on scores generated using a set of predetermined criteria such as term frequency-inverse sentence frequency (TF-ISF) sentence or term position, and quantity of keywords, [6].

The K-means approach is basic and straightforward to use; the structure of SOM is more complex, but the clustering results are more visible and accessible. The disadvantage of k-means is that the value of K must be specified in advance, and the initial document seeds must be chosen at random, which has an impact on the clustering outcomes, [17].

REFERENCES

[1.] Wikipedia.Information.[Online].Available:https://en.wikipedia.org/wiki/Information
[2.] Dr. Werner Gitt. "Chapter 6 Information in Living Organisms" in In the Beginning Was Information, Answers in Genesis, 2009. [Online]. Available: https://answersingenesis.org/genetics/information-theory/information-in-living-organisms/
[3.] VirginiaTech. Types of information sources. [Online]. Available : http://www.lib.vt.edu/help/research/info-sources.html
[4.] University of Nottingham. Types of information resources in Studying Effectively.[Online].Available:https://www.nottingham.ac.uk/studyingeffectively/reading/infotypes.aspx
[5.] J. Saarikoski et al, "Self-Organising Maps in Document Classification: A Comparison with Six Machine Learning Methods" in Adaptive and Natural Computing Algorithms: Proc. of the 10th Int. Conf., ICANNGA 2011, Ljubljana, Slovenia, April 14-16, 2011, A. Dobnikar, U. Lotrič, and B. Šter Eds., Springer-Verlag Berlin Heidelberg, 2011. pp. 260–269.
[6.] Dingding Wang, "Document Understanding Using Data Mining and Machine Learning Techniques," Ph.D. dissertation, College of Eng. and Computing, Florida Int. Univ., Miami, Florida, 2010.
[7.] M.Steinbach et al., "A Comparison of Document Clustering Techniques," Dept. of Comput. Sci. and Eng., Univ. of Minnesota, Rep. #00-034.
[8.] B.H.ChandraShekar and Dr.G.Shoba, "Classification Of Documents Using Kohonen's Self-Organizing Map," International Journal of Computer Theory and Engineering, vol. 1, no. 5, pp. 610-613, December, 2009.
[9.] Jong-Yeol Yoo and Dongmin Yang, "Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier," in Advanced Science and Technology Letters, vol.111, (COMCOMS 2015), pp.263-266, 2015 © SERSC. doi: http://dx.doi.org/10.14257/astl.2015.111.50

[10.] Yiheng Chen et al., "The Comparison of SOM and K-means for Text Clustering," Computer and Information Science, vol. 3, no. 2, pp. 268-274, May, 2010.
[11.] John Wurzler, "Information Risks & Risk Management", SANS Institute InfoSec Reading Room, Apr. 23, 2013. [Online]. Available: https://www.sans.org/reading-room/whitepapers/awareness/information-risks-risk-management-34210. [Accessed: May. 2, 2018].
[12.] Susan Fowler, "Information Classification – Who, Why and How", SANS Institute InfoSec Reading Room, Feb. 28, 2003. [Online]. Available: https://www.sans.org/reading-room/whitepapers/auditing/information-classification-who-846. [Accessed: May 2, 2018].
[13.] Dejan Kosutic, "Information classification according to ISO 27001," The ISO 27001 & ISO 22301 Blog, May 12, 2014. [Online]. Available: http://advisera.com/27001academy/blog/2014/05/12/information-classification-according-to-iso-27001/. [Accessed: Mar. 3, 2018].
[14.] Jethro Perkins, "Information Security - Information Classification", LSE Governance, LSE Community, Mar. 12, 2013. [Online]. Available: http://www.lse.ac.uk/intranet/LSEServices/policies/pdfs/school/infSecStaIT.pdf. [Accessed: Mar. 3, 2018].
[15.] Apte et al., "Automated Learning of Decision Rules for Text Categorization," IBM Res. Div., Yorktown Heights, NY, Rep. RC 18879, 1994.
[16.] J. David Brown and Daniel Charlebois, "Security classification using automated learning (SCALE) Optimizing statistical natural language processing techniques to assign security labels to unstructured text," Defence R&D Canada, Ottawa, Tech. Memo. DRDC Ottawa TM 2010-215, Dec. 2010.
[17.] Jyri Saarikoski, "On Text Document Classification and Retrieval Using Self-Organising Maps," Academic Dissertation, Sch. of Info. Sci., Tampere Univ., Tampere, 2014.
[18.] Jyri Saarikoski et al., "A study of the use of self-organising maps in information retrieval," Emerald Group Publishing Limited, vol. 65, no. 2, pp. 304-322, 2009.
[19.] Fabrizio Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys (CSUR), vol. 34, pp. 1-47, March 2002.
[20.] J. Saarikoski et al., "On Document Classification with Self-Organising Maps" in Adaptive and Natural Computing Algorithms: Proc. of the 9th Int. Conf., ICANNGA 2009, Kuopio, Finland, April 23-25, 2009, M. Kolehmainen et al., Eds., Springer-Verlag Berlin Heidelberg, 2009. pp. 140–149.
[21.] Jyri Saarikoski et al., "On the influence of training data quality on text document classification using machine learning methods," International Journal of Knowledge Engineering and Data Mining, vol. 3, pp. 143-169, August 2015.
[22.] J.Y. Yoo and D. Yang, "Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier," Advanced Science and Technology Letters, vol. 111, pp. 263-266, 2015.
[23.] "Introducing Machine Learning", MathWorks, 2016. [Online]. [Available: http://www.mathworks.com/tagteam/89703_92991v00_machine_learning_section1_ebook_v12.pdf?s_tid=solmain_mlr_cta2]. [Accessed: Mar. 5, 2018].
[24.] James Poole and Alok Ojha, "Assisted Learning for Document Classification," U.S. Patent 9,342,795B1, May 17, 2016.
[25.] David M. Blei et al, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, pp. 993-1022, 2003.