© Universiti Tun Hussein Onn Malaysia Publisher's Office

EmAIT

Journal homepage: http://publisher.uthm.edu.my/ojs/index.php/emait
e-ISSN : 2773-5540

Emerging Advances in Integrated Technology

Detection of topic on Health News in Twitter Data

Shum Chen Yau[1], Juhaida Abu Bakar[1*], Azian Azamimi Abdullah[2,3*], Nor Hazlyna Harun[1], Ruziana Mohamad Rasli[4], Lim Zheng Yang[1], Evon Thum Yi Mun[1]

[1]Data Science Research Lab, School of Computing, UUM College of Arts and Sciences,
Universiti Utara Malaysia, 06010 Sintok, Kedah, MALAYSIA

[2]Medical Devices and Life Sciences Cluster, Sport Engineering Research Centre, Centre of Excellence (SERC),
Universiti Malaysia Perlis (UniMAP), 02600 Arau, Perlis, MALAYSIA

[3]Faculty of Electronic Engineering Technology,
Universiti Malaysia Perlis (UniMAP), 02600 Arau, Perlis, MALAYSIA

[4]Department of Information Technology and Communication,
Tuanku Syed Sirajuddin Polytechnic, Pauh Putra, 02600 Arau, Perlis, MALAYSIA

DOI: https://doi.org/10.30880/emait.2021.02.02.004
Received 27 September 2021; Accepted 18 October 2021; Available online 15 November 2021

Abstract: The development and rapid popularization of the internet has led to an exponential growth of data in the network, thus, the text mining becomes more important. Users search for the information from the immense information available online. The ways to obtain valuable information, and to classify, organize and manage vast text data automatically make the text processing even more difficult. Therefore, in order to solve those problems and requirements, intelligent information processing has been extensively studied. Topic modelling has been widely employed in the field of natural language processing. Current research directions are more focused on ways to improve the classification speed and accuracy of text classification and topic detection as well as selecting feature methods in achieving better dimension reduction operations. Latent Dirichlet Allocation (LDA) topic model works well on data noise reduction. The LDA is widely used as a feature model combined with the classifier design in order to achieve a good classification effect. This study aims to conduct data mining and save load from the huge database. Thus, three supervised learning algorithms are run, which are Naïve Bayes, Decision Tree and Random Forest. Random Forest classifier outperforms the other two classifiers with 99.99% accuracy. Seven clusters for topic modelling have been revealed using Random Forest classifier. Each output has been set to four highest word and shows the highest term and its weight. The highest term used in the dataset is term 'Ebola'. Based on the finding of this study, it shows that the combination of the LDA and supervised learning algorithm effectively solve the problem of data sparseness in short text sets. The method of selecting microblogs that are most likely to discuss news topics will significantly reduce the size of data objects of concern, and to a certain extent eliminate the interference of non-news blogs.

Keywords: Topic modelling, text classification, feature selection, Latent Dirichlet Allocation, machine learning algorithm, random forest classifier

1. Introduction

Large amounts of data are collected every day. As more information becomes available, it is difficult to assess for the required data. Due to the difficult nature of text, analyzing, understanding, organizing, and sorting through text data is complex and time-consuming, therefore it is rather challenging to extract value from that the data. Therefore, tools and techniques are required to organize, search and understand large quantities of information.

**Corresponding author: juhaida.ab@uum.edu.my, azamimi@unimap.edu.my*

Manually detecting topics is time-consuming, and costly. Topic Detection with machine learning helps to search data extensively, opening up completely new avenues for gathering useful insights. Furthermore, by analyzing data sources such as reviews, polls, social media messages, emails, and customer service tickets, relevant insights are gained using subject detection. The most commonly debated topics about a product or service are discovered within seconds.

Topic detection, also known as topic modelling, is a machine learning technique for organizing and understanding large collections of text data by assigning tags or categories based on the topic or theme of each individual text [1]. Topic detection assist to discover patterns and semantic constructs within each individual texts. Topic modelling and topic classification are the two most popular machine learning methods for topic detection. Topic modelling is a form of unsupervised machine learning. This means it can infer patterns and group related expressions without the need to identify topic tags or train data. Topic classification, unlike topic modelling, involves knowledge of a collection of topics across texts before analyzing them. Data is manually tagged with these topics so that a topic classifier can learn and make predictions on its own.

When there is too much information to be categorized, the process becomes time-consuming, costly, and unreliable. Topic detection is useful for analyzing large volumes of data immediately and saves cost. Topic detection has proven to be a great alternative for quickly and effectively analysing data because it saves time, effort, and money.

Every day, companies produce and collect huge quantity of data. Analyzing and processing this data using automated topic detection can assist businesses in making better decisions, optimizing internal processes, identifying patterns, and various other benefits that will help businesses to become more effective and profitable. Machine learning models are critical when it comes to sorting through large quantities of data. Topic identification enables users to quickly search documents and determine the ongoing discussion.

There are multiple methods to conduct topic detection. This study proposes a Latent Dirichlet Allocation (LDA) based in labelling topics with health tweets. This algorithm is based on the generative property of LDA. The LDA is an unsupervised statistical machine learning technique that allows to classify text in a document to a particular topic [2]. The LDA automatically finds a likely set of topics over a large document collection and represents each document in the collection in the form of topic proportions. In LDA, a topic can be interpreted by its most probable words. The LDA applied in this study to a set of documents and divided into topics. This study shows how this approach can be augmented by incorporating class labels assigned to few words in the collection.

2. Related Work

This section describes the related work of topic modelling. Research by [3] described fuzzy latent semantic analysis (FLSA), a singular method in subject matter modelling that use fuzzy angle. The FLSA can handle fitness and clinical corpora redundancy problem and offers a brand-new approach to estimate the quantity of subjects. The quantitative reviews show that FLSA produces advanced overall performance and functions to latent Dirichlet allocation (LDA), the most popular subject matter version.

Most clinical reports and electronic health records (EHRs) are in text format, which makes information processing and identifying appropriate documents difficult. Finding ways to automatically retrieve a large amount of fitness and science knowledge has always been an interesting subject. In recent years, powerful strategies for automating text processing have been created. Subject matter modelling is a common method for retrieving data based on discovering topics in health and scientific corpora. However, new perspectives are needed for this technique.

Traditionally, large amount of health and medical text data have been produced and stored. This vast volume of text data and EHRs creates a clear opportunity for businesses to save billions per year by employing advanced data analytics. The centre of big health and medical data science research is developing efficient techniques for finding hidden structure in broad, complicated health and medical data sets to address questions about those data. New data analytic methods and techniques were developed with significant resources.

However, retrieving large amount of health and medical text data is currently a major challenge. Bag-of-words (BOW) is a common method for representing medical text data. Topic modelling is a common technique for dealing with issues of sparsity and high dimensionality. Two matrices are the product of topic modelling. The probability of words for each subject or P is the first, and the probability of topics for each document or P is the second.

Topics Documents matrix reduces the number of dimensions from the number of words in BOW approach to the number of topics. Topics Documents matrix comprises of 100 rows and 5000 columns. In this study, Fuzzy Latent Semantic is proposed. This model outperforms other models in both redundant and non-redundant documents, and aids topic models in estimating the number of topics in a corpus.

Research by [4] progressed with the description of a topic model, with a focal point at the knowledge of topic modelling. A widespread outline is supplied on a way to construct a software in a subject version and a way to increase a subject version. Meanwhile, the literature on utility of topic fashions to organic information was searched and analyzed. According to the sorts of fashions and the analogy among the concept of file-subject matter-phrase and a biological item (in addition to the tasks of a subject model), the associated studies are categorised and an outlook on the use of topic models for the improvement of bioinformatics applications is provided.

The rapid accumulation of biological datasets necessitates the designing of system gaining knowledge of strategies to automate information evaluation. In recent years, so-called subject matter fashions originated from the field of natural language processing were receiving attention in bioinformatics because of their interpretability. Therefore, this study reviews the software and development of subject matter fashions for bioinformatics.

Topic modelling is a useful method (in comparison to the traditional means of data discount in bioinformatics) and enhances researchers' ability to interpret biological statistics. Nevertheless, due to the lack of topic models optimized for specific biological statistics, the research on subject matter modelling in organic records have a long way forward. Nonetheless, subject matter fashions are a promising approach for numerous programs in bioinformatics research.

3. Methodology

Natural Language Processing allows the computer to analyze, understand and generate new text. Technologies such as information extraction, information, categorization, clustering, visualization and summarization can be used in text mining process. The following sections discuss each technology and the role in the respective process.

3.1 Information Extraction

Information extraction is the initial step when computer analyses the unstructured text by identifying their keyword and relationship within the text [5]. This text mining technologies specializes in computing the extraction of entities, attributes, and their relationships from semi-based or unstructured texts. The extracted records are then stored in a database for destiny access and retrieval. The efficacy and relevancy of the outcomes are checked and evaluated based on the use of precision and process. The process is show in Fig. 1.

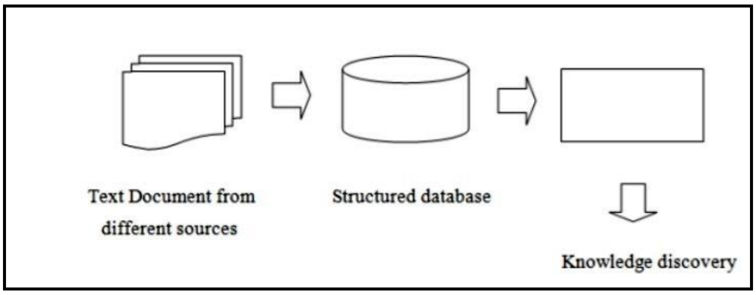


Fig. 1 - Information extraction process

3.2 Categorization

Categorization automatically assigns one or more category to loosen textual content record. Categorization is supervised by studying method due to the fact it is based totally on input-output examples to classify new files. Predefined classes are assigned to the textual content documents primarily based on their content material. The typical text categorization process consists of pre-processing, indexing, dimensionally reduction, and classification [6][7]. The goal of categorization is to educate classifier on the premise of recognized examples after which unknown examples are labelled routinely. Naïve Bayes and Random Forest are employed to rent the text categorization in this project.

3.3 Clustering

In clustering part, Latent Dirichlet Allocation (LDA) is used to cluster the text. LDA is a statistical and graphical version, which might be used to reap relationships between more than one documents in a corpus. It evolves the use of Variational Exception Maximization (VEM) algorithm for obtaining the most likelihood estimate from the entire corpus of text [8]. Traditionally, this could be solved by way of picking out the top few words within the bag of phrases. However, this lacks in the semantics of the sentence. This model follows the concept that each document may be defined by way of the probabilistic distribution of subjects and every subject matter may be defined by way of the probabilistic distribution of phrases. Thus, a clearer understanding is obtained on the ways the topics are linked.

3.4 Visualization

Text mining visualization strategies can improve and simplify the discovery of applicable facts. To represent individual files or agencies of documents, text flags are used to show record class and to expose density colours that are used. Visual text mining places big textual sources in a visible hierarchy [9]. In this phase, the user can interact by indicating which row is related with the topic with the aid of zooming and scaling. The process is shown in Fig. 2.

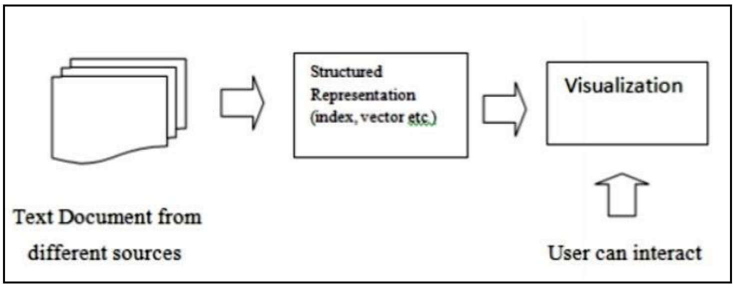

Structured Representation (index, vector etc.)
Visualization
Text Document from different sources
User can interact

Fig. 2 - Visualization process

3.5 Summarization

Text summarization refers to the technique of robotically producing a compressed model of a particular text that holds valuable statistics for the user [10]. The aim of mining technique of the newsletter is to browse via a couple of textual content sources to craft summaries of texts containing a widespread percentage of statistics in a concise format, keeping the general meaning and the unique files are identical.

The complete setting of the experiment uses KNIME Analytics Platform. The experiment includes reading file, pre-processing (transformation of strings to document), topic extractor analysis using LDA, Partitioning (Training and Testing datasets), and Learning model with three Machine Learning classifier (Decision Tree, Naïve Bayes, and Random Forest). The File Reader node is used to read files in the first phase. Then, in the second phase, the text will go through the pre-processing phase which transforms text to a specific form suitable for the analysis. Several pre-processing steps are applied including index column to string values (Column Rename Node), conversion of strings to documents (Strings to Document Node), and filter all columns except the document column (Column Filter Node). Then, the document column will go through specific process such as Punctuation Erasure, Number Filter, N Chars Filter, Stop Word Filter, and Case Converter. All this specific process is the standard pre-processing of text document. Then, in the third phase, the documents will go through the LDA analysis to extract the highest word frequency and its weight. Fig. 3 shows the process of the topic extractor using LDA. Certain analysis also implemented in the experiments such as concatenate terms for topics, count the occurrence of each term and creating tag cloud for each topic.

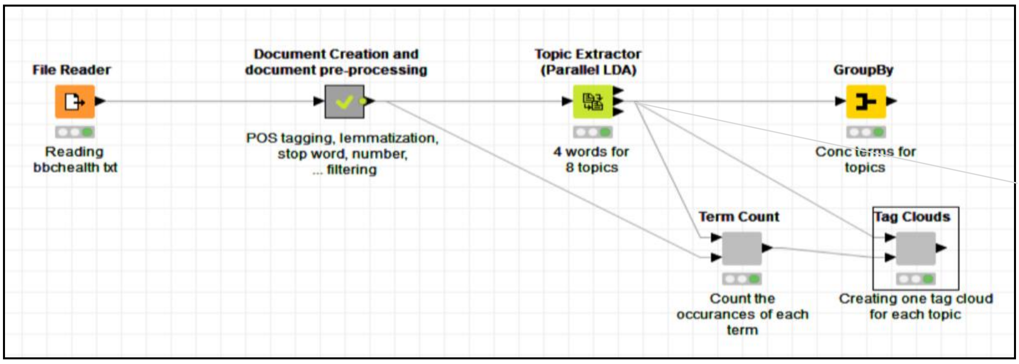

File Reader
Reading bbchealth txt
Document Creation and document pre-processing
POS tagging, lemmatization, stop word, number, ... filtering
Topic Extractor (Parallel LDA)
4 words for 8 topics
GroupBy
Conc terms for topics
Term Count
Count the occurances of each term
Tag Clouds
Creating one tag cloud for each topic

Fig. 3 - First part of the experiment (File reader, Pre-processing, and LDA analysis)

In the fourth phase, the document is divided to two partition using K-Fold cross-validation. A 10-fold cross validation has been applied in this study. Cross validation is used to validate the performance of the learning model [11]. Then, in the fifth phase, learning model is performed. Three state-of-the art classifier model namely Decision Tree, Naïve Bayes and Random Forest are chosen to perform the classification model. A second part of the experiment is shown in Fig. 4, which is an example of the experiment perform on Decision Tree classifier. Another similar experiment is also performed to another two classifiers; Naïve Bayes and Random Forest.

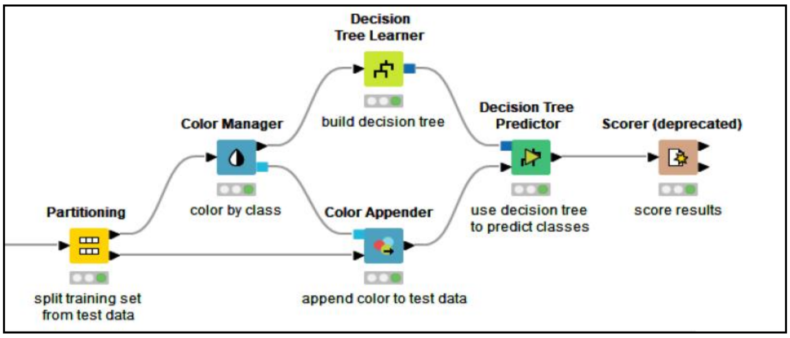


Fig. 4 - Second part of the experiment (Partitioning, and Learning model)

4. Findings and Analysis

In this study, for large-scale Twitter data sets, a method based on hidden topic mining and topic detection is proposed to implement the detection of current health news topics. Three (3) supervised learning algorithms are run which are Naïve Bayes, Decision Tree algorithm and Random Forest with the aim of mining the data and save load from the huge database. The dataset was downloaded from a UCI machine learning from the link *https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter*. Only *bbchealth.txt* was used in this study and the data included the id, date, time, and tweet from [3]. This dataset is suitable for topic modelling specifically and clustering task in general. The date of the data was for the duration from 30 September 2013 to 9 April 2015.

Table 1 - Results for Twitter Health News classification trained by decision Tree Algorithm

Split data (Training: Testing)	Decision Tree Algorithm				
	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	ROC (%)
90:10	99.75	97.8	98	98.9	99.93
80:20	98.6	97.2	95.8	96.8	99.91
70:30	99.3	97.6	98.4	98.8	99.93
60:40	99.1	98	98.5	98.6	99.87
50:50	99.3	97.1	97.2	98.5	99.88
40:60	98.6	95	96.3	97	99.92
30:70	96.5	93.4	92.5	93.1	99.91
20:80	96.3	92.3	92.5	92.9	99.92
10:90	94.9	84.9	92.6	90.8	99.90

As shown in Table 1, the study defines that decision tree achieved the highest accuracy which is 99.75% in this data, with 90% of training set and 10% of testing set. The 90:10 split data ratio have high percentage as well. Meanwhile, the split data ratio 10:90 achieved the lowest accuracy among decision tree algorithm which is 94.9%.

Table 2 - Results for Twitter Health News classification trained by Naïve Bayes algorithm

Split data (Training: Testing)	Naïve Bayes Algorithm				
	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	ROC (%)
90:10	97.50	89.30	93.70	94.3	99.93
80:20	98.3	93.2	96.2	96.5	99.91
70:30	98.6	95.5	97.2	97.4	99.93
60:40	97.8	93.2	95.2	96.5	99.87
50:50	98.9	96.7	97.6	98.3	99.88
40:60	98.1	93.8	96.1	96.5	99.92
30:70	98.3	96.2	95.8	97.6	99.91
20:80	96.7	89.2	91.9	94.2	99.92
10:90	95.2	84.6	90.3	90.3	99.90

Table 2 indicates that Naive Bayes obtained the highest accuracy which is 98.9% with 50% of training set and 50% of testing set. In the 50:50 split data ratio, the precision, recall, F-measure and ROC show the highest percentage among the data which is 96.7%, 97.6%, 98.3 and 99.88%, respectively. The lowest accuracy achieved in Naive Bayes is 95.2% with the split data of 10% of training set and 90% of testing set.

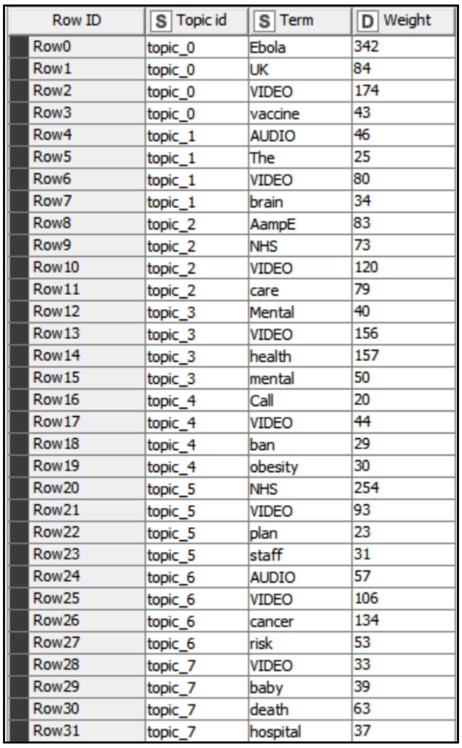
Table 3 - Results for Twitter Health News classification trained by Random Forest algorithm

Split data (Training: Testing)	Random Forest Algorithm				
	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	ROC (%)
90:10	99.7	98	98	99	99.93
80:20	99.9	98.6	97.9	98.9	99.91
70:30	99.3	96.5	97	98.2	99.93
60:40	99.4	98.1	98.5	99	99.87
50:50	99.4	96.3	98.1	97.2	99.88
40:60	99	96.8	97.6	97.5	99.92
30:70	98.4	96.2	97	97.4	99.91
20:80	97.5	95.2	94.3	96.7	99.92
10:90	97.6	91.6	94.2	95.1	99.90

In Table 3, Random Forest reaches the highest accuracy which is 99.9% at the stage of split data with 80% of training set and 20% of testing set. In this ratio, precision reached the highest percentage as well at 98.6%. However, the percentage of Recall and F-measure were not the highest which are 97.9% and 98.9%, respectively. The lowest accuracy in Random Forest in a split data ratio is 10:90, which is 97.6%.

5. Discussion

Based on Tables 1-3 above, this study concludes that Random Forest is the best classifier to process this data as the accuracy of Random Forest training model is higher than the other two classifiers. The highest accuracy rate is 99.9%. While the most unsuitable classifier is Naive Bayes Learner, as the lowest accuracy in the condition of 10:90 of training and testing set is 95.2%.



Row ID	S Topic id	S Term	D Weight
Row0	topic_0	Ebola	342
Row1	topic_0	UK	84
Row2	topic_0	VIDEO	174
Row3	topic_0	vaccine	43
Row4	topic_1	AUDIO	46
Row5	topic_1	The	25
Row6	topic_1	VIDEO	80
Row7	topic_1	brain	34
Row8	topic_2	AampE	83
Row9	topic_2	NHS	73
Row10	topic_2	VIDEO	120
Row11	topic_2	care	79
Row12	topic_3	Mental	40
Row13	topic_3	VIDEO	156
Row14	topic_3	health	157
Row15	topic_3	mental	50
Row16	topic_4	Call	20
Row17	topic_4	VIDEO	44
Row18	topic_4	ban	29
Row19	topic_4	obesity	30
Row20	topic_5	NHS	254
Row21	topic_5	VIDEO	93
Row22	topic_5	plan	23
Row23	topic_5	staff	31
Row24	topic_6	AUDIO	57
Row25	topic_6	VIDEO	106
Row26	topic_6	cancer	134
Row27	topic_6	risk	53
Row28	topic_7	VIDEO	33
Row29	topic_7	baby	39
Row30	topic_7	death	63
Row31	topic_7	hospital	37

Fig. 5 - The LDA clustering result

Fig. 5 above shows the result of the LDA clustering using the highest learning model based on Random Forest. Seven clusters for topic modelling are discovered from this dataset which represents Topic_0 to Topic_7 output. Each output has been set to 4 highest word and shows the highest term and its weight. The highest term used in Topic_0 is term 'Ebola', Topic_1 is term 'VIDEO', Topic_2 is term 'VIDEO', Topic_3 is term 'health', Topic_4 is term 'VIDEO', Topic_5 is term 'NHS', Topic_6 is term 'cancer', and Topic_7 is term 'death'.

6. Discussion

This study discussed the purpose of detecting health news topics from a large-scale microblog short text data set. Although the dataset used in this study is obtained from Twitter, no doubt that today's application model can be applied to any other microblog system as well. In this study, the method of hidden topic modelling is used to effectively solve the problem of data sparseness in short text sets. The method of selecting microblogs that are most likely to discuss news topics will greatly reduce the size of data objects of concern, and to a certain extent eliminate the interference of non-news blogs. But for the work in this study, there is still room for improvement. Moreover, the method in this study is not real-time. A real-time health news discovery system can be achieved by introducing large-scale external data by shortening the time window, and conducting hidden topic mining in the background. On the other hand, because of its own length, it is difficult for Twitter to fully describe a health news issue. Future studies could be conducted on allowing Twitter to describe a health news completely.

Acknowledgement

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme for Research Acculturation of Early Career Researchers (RACER/1/2019/ICT02/UUM/1). This research was supported by Universiti Utara Malaysia (UUM) and Universiti Malaysia Perlis (UniMAP).

References

[1] S. A. Curiskis, B. Drake, T. R. Osborn, P. J. Kennedy, An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit, Information Processing & Management, 57(2), 102034, 2020

[2] F. Millstein, Natural language processing with python: natural language processing using NLTK, Frank Millstein, 2020

[3] A. Karami, A. Gangopadhyay, B. Zhou, H. Kharrazi, Fuzzy approach topic discovery in health and medical corpora, International Journal of Fuzzy Systems, 20(4), 1334-1345, 2018

[4] L. Liu, L. Tang, W. Dong, S. Yao, W. Zhou, An overview of topic modeling and its current applications in bioinformatics, SpringerPlus, 5(1), 1-22, 2016

[5] S. V. Gaikwad, A. Chaugule, P. Patil, Text mining methods and techniques, International Journal of Computer Applications, 85(17), 2014

[6] W. Lam, M. Ruiz, P. Srinivasan, Automatic text categorization and its application to text retrieval, IEEE Transactions on Knowledge and Data engineering, 11(6), 865-879, 1999

[7] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, Text classification using string kernels, Journal of Machine Learning Research, 2(Feb), 419-444, 2002

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research, 3, 993-1022, 2003

[9] R. Feldman, J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge university press, 2007

[10] J. L. Neto, A. D. Santos, C. A. Kaestner, N. Alexandre, D. Santos, Document clustering and text summarization, 2000

[11] J. Han, M. Kamber, J. Pei, Data mining concepts and techniques third edition, The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124, 2011