International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 2 ǁ 2022 ǁ PP. 20-26

Autopsy reports using machine learning Text Classification Techniques in Forensics

Shadab Ahmad Siddique, *M. Tech Scholar, Department of Computer Science and Engineering, Kanpur Institute of Technology, Kanpur, India.*

Pragati Dwivedi, *Assistant Professor, Department of Computer Science & Engineering, Kanpur Institute of Technology, Kanpur, India.*

Abstract: *A forensic autopsy is a surgical process in which experts collect a deceased body's internal and external information. These death certificates are the source of timely warnings of an increase in disease activity. It's only helpful if accurate and quantitative data is available. Therefore, the Classification of plain text medical autopsy reports reduces the time consumption and irregularities. The motive is to design an automatic text classification system that classifies plain text autopsy reports. Therefore, a methodology proposes using different Automatic Text Classification Techniques (ATC). This technique has embedded Feature Extraction, Feature Representation, and Feature Reduction techniques. These techniques use for the construction of classification models that classify the text of autopsy reports. Data sets collected from these types will be helpful in future experiments. Finally, the performance of the classifier measures by using different Evaluation parameters. These Evaluation Measures arePrecision, Recall, Accuracy, and F-measure.*

Keywords: *Machine learning (ML), Text Classification, NB, Forensic Analysis, SVM.*

Date of Submission: 09-02-2022 Date of acceptance: 23-02-2022

I. INTRODUCTION

.Death is the result of severe disease or an injury. The death certificate is a medical and legal document containing the cause of death and other related information of a deceased body for a single person. Death certificates are the essential source for collecting, examining, and interpreting data related to health. The purpose of this health-related data or death certificate is to control or prevent an increase in a disease or injury. The only possibility is if death certificates contain only precise and quantitative data. If any of the more various causes of death is the timely report, it will help health authorities in their disease or injury prevention programs. Examination of autopsy reports gives beneficial information related to health education. It also improves the health care industry in terms of quality. Pathologists are the experts who examine the dead body. They collect information from the internal and external organs of the dead body. Internal examination information about all internal organs like heart, kidney, liver, lungs, abdomen, etc., is collected. While during the external examination, information from head to toe of a deceased body is composed. Pathologists also contain other related information of a dead body like personal information, medical history, and histopathology reports. Initially, the autopsy report is ready within two to 3 three days, while the final report prepares in more than a month. It may take up to 90 days in case of any complexity, depending on a death case. Examination of an autopsy report is very time-consuming and challenging. To assign the cause of death with labor concentrated procedure re- results in several inconsistencies. So, to minimize irregularities, inconsistencies, time, and work, automatic text classification techniques are used to predict the cause of death automatically. Different text classification techniques use for the Classification of autopsy reports from many recent years. Ghulam Mujtaba et al. [1] classifies plaintext medical documents through Automatic Classification Techniques. It automatically predicts the Cause of Death by organizing the plain text using various text classification techniques. Liyana Shuaib et al. [2] proposed the Conceptual Graph-Based Document Representation (CGDR) strategy. Bevan Koopman et al. [3] proposed an Automatic system to identify Cancer from death certificates. Zahra Shams Khoozani el at. [4] Presented an Automated Application that use to predict the CoD. Ghulam Mujtaba et al. [5] proposed an Automatic Multi-Class Classification system. This system uses to assume accident- related COD from Autopsy Reports using Expert Driven Feature Selection. Rafał Wozniak et al. [6] proposed Feature Selection methods used for improving Multi-Label Medical Text Classification. Liyana Shuib et al. [7] classify plain text medical documents

Through Automatic Classification Techniques. It automatically predicts the Cause of Death by organizing the plain text. David Muscatello et al. [8] presented two classification methods to predict the occurrence of some diseases like Diabetes, Influenza, Pneumonia, and HIV, i.e., the Machine Learning method and set of Keyword Matching Rule. Guido Zuccon et al. [9] proposed an Automatic system designed to identify Cancer from Autopsy Reports. Bevan Koopman et al. [10] proposed machine learning methods to classify death certificates related to Cancer. Both Automatic text classification (ATC) technique and machine learning have been pragmatic in many fields like banking, e-commerce, cybercrime detection, and medical document classification. Recently ATC is most used in the Classification of medical documents. If medical papers' Classification manually. The primary purpose of ATC is to give a related class to a specific document from a pre- defined set of classes. In the proposed technique, the plain text autopsy report Classification uses different classification techniques [17]. In Automatic Text Classification (ATC), the Feature Extraction algorithm" Unigrams" is used to convert text document content into useful word features. The extracted features statistically represent to construct Master Feature Vector by using Feature Representation techniques which include Term Frequency (TF) and" Term Frequency Inverse Document Frequency" (TF-IDF). In Master Feature Vector, rows represent documents while columns represent Features. Master feature vector consists of more features. Some Feature Reduction techniques like Chi-Square and principal component analysis use to gain the most discriminate features. The final classification results collect by applying different ATC techniques to the master feature vector.

II. LITERATURE REVIEW

Among different challenges in text classification, the Classification of plaintext medical autopsy report is one of the significant challenges. Autopsy reports contain all information about a deceased body Pathologists collect this information, and several classification techniques use for the Classification of autopsy reports. Several researchers proposed different approaches to classify autopsy reports, but there were some limitations. This technique offers to overcome these limitations. Anne Gallay et al. [11] proposed a Rule-Based method using four Processing steps: standardization rules, splitting causes of death using delimiters, spelling corrections, and dictionary projection. A managed AI strategy utilizing a straight Support Vector Machine (SVM) classifier executed. Mortality exploration is one of the essential significances of general health observation. On account of the Electronic Death Registration System (EDRS), the constant recording of death declarations gives necessary information designed for receptive death observation dependent on medicinal reasons for death in the free-content arrangement. Responsive mortality observation depends on checking of death syndromic gatherings (MSGs). An MSG is a group of fundamental reasons for death that encounters immediate recognition and effect evaluation of general wellbeing occasions. The main point of examination is to execute and gauge the presentation of rule- based techniques. Two administered representations propose the programmed free-content reason for death grouping from death declarations to objectify them for a routine investigation. A standard-based strategy executes four preparing steps: institutionalization rules, parting reasons for death using delimiters, spelling rectifications, and lexicon projection. A manage AI technique utilizing a straight Support Vector Machine (SVM) additionally actualize. Two representations were delivered utilizing various highlights, SVM1 dependent on superficial high- lights and SVM2 consolidating external highlights and MSGs arranged by standard created strategy as highlight directions [16]. Arrangement execution assessment on 7 MSGs (Influenza, Low respiratory illnesses, Asphyxia/anomalous breath, Acute respiratory ailment, Sepsis, Chronic stomach related infections, and Chronic endocrine maladies). Rafał Wozniak et al. [6] proposed feature selection methods designed to improve multi-label medical text classification. In several cases, multi-label Classification is used because documents belong to multiple classes. Multi-label classification algorithms are applied on multidimensional datasets of several aspects and a comparatively lesser number of instances. This type of situation usually occurs in the case of medical reports. The methodology of Multi-label Classification is opposite as compared to the single-label classification methods. During the process of Classification, it predicts multiple classes' labels. Five multi-label classification methods used which are:

Binary Relevance: It simplifies the multi-label problem into binary classifications as it creates a label for every label. All solitary results are combined to form a product. It ignores the label reliance across the classification method.

Label Power set: From the training set, it collects all-new classes from exclusive collections. No matter how many numbers and a variety of labels are included, this technique can be applied.

Classifier chain: This is the enhancement of Binary Relevance. In addition, earlier features are used employing novel features.

Ensembles of Classifier Chains (ECC): This method gives the solution for classifier chains. The final multi-label set is created according to the given threshold value.

Labels Chain (LC): It creates multi-class classifiers. The idea of a classification chain is used in this classification method. Feature Selection consists of four classes in this proposed technique, i.e., Embedded, Filter, Wrapper, and Hybrid techniques.

Embedded: It classifies Feature Selection as an intimate fragment of the learning process compared to the filters and wrapper.

Filter: It selects features individually. Features are they are categorized according to the established evaluation measures. Parts with lowermost scores are distant at the initial level of Classification. It helps in fast computation when the data set is more extensive.

- **Wrapper:** According to the supervised learning algorithm, the best feature subset is selected. Features having the lowest score are eliminated at each set of Classifications.

In the proposed method, two evaluation measures are selected, which are Hamming Loss and Classification Accuracy. The one is considered as the primary classifier, i.e., Classification Accuracy. Classification Accuracy calculates the set of labels that are entirely and correctly predicted. Whereas the Hamming Loss is the fraction of imperfectly confidential labels over entire labels. The best results are collected with the combination of the filter with the wrapper methods. Ghulam Mujtaba et al. [14] proposed term based and SNOMED CT idea for the Classification of forensic death certificates. These days content order has been widely utilized in restorative space to group free content clinical reports. In this examination, content order procedures have been used to decide the reason for death from free content legal postmortem reports utilizing proposed term- based and SNOMED CT idea-based highlights. During this investigation, definite term-based highlights and idea-based highlights extricated since many 1500 legal postmortem examination reports have four death habits while 16 unique reasons for death. All highlights were utilized for the preparation of the content classifier. Then classifier was sent in course engineering: the primary level foresees the way of death, and the subsequent level will anticipate the Cause of Death utilizing designed term-based and SNOMED CT idea-based highlights [18]. In addition, to demonstrate the centrality of the developed methodology, aftereffects of the proposed method are considered with four cutting edges high- light extraction draws near. At last, additionally exhibited the examination of one-level grouping contrasted with two-level characterization. Luke Butt et al. [10] proposed machine learning methods for the Classification of Cancer-associated autopsy reports. The Classification of death certificates related to Cancer is very complex and time-consuming. Researchers [16] offered the automatic recognition of Cancer from plain text death certificates. Three main basic approaches are used for the identification of Cancer. These approaches are:

- Automatic Feature Extraction
- Feature Weighting
- Automatic Classification

III. PROBLEM STATEMENT

Examining the autopsy report to determine Manner of Death and Cause of Death is labor-intensive and subject to inconsistencies. The feature set is affecting the accuracy while classifying autopsy reports.

A. RESEARCH QUESTION

1) How many accuracies can be improved by using the proposed technique for the Classification of Autopsy reports?

2) How many numbers of diseases are classified in autopsy reports?

IV. PROPOSED TECHNIQUE

The procedure employed for the Classification of direct text autopsy reports is briefly discussed. All steps are explained in Fig. 1. A supervised machine learning approach has been used for the Classification of plain text medical reports. They are considering the issues discussed in the last chapter. There is a need to present an approach to accurately classify the understandable text medical reports reducing the feature set size. For the Classification of medical information and death certificates, Automatic Text Classification (ATC) techniques are used. Feature Extraction, Feature Representation, and Feature Reduction techniques are used as ATC techniques. Data is collected and then processed. After processing, ATC techniques are applied. Using a training set classification model is constructed. After this, algorithms are trained and tested one by one to obtain the results. In the proposed methodology, the data set size is increased compared to previous work as a more significant number of autopsy reports are used. More classes are also used. In previous studies, only four categories, i.e., diabetes, influenza, pneumonia, and HIV, are used. In comparison, eight classes are used in the proposed work, i.e., Cardiopulmonary Arrest, Diabetes millets, HIV +ve / HIV AIDS, Pneumonia, Hepatitis A/ B, Tuberculosis, PUL-TB / DUL-TB, Respiratory Failure, Sepsis. The proposed methodology performed better when compared with base papers in terms of accuracy with more classes and less feature set size [15]. Fig. 1 shows the graphical representation of the proposed technique. The processes include collecting Data, Data Pre-Processing, Feature Extraction Technique, Feature Representation Techniques, Feature Reduction Techniques, Classification Models, and the Results.

V. RESULT

In this section implementation of the proposed methodology is thoroughly discussed. Results are presented of the proposed method and then discuss results.

A. DATA SET

The data set is collected from different district health quarter hospitals in Punjab, Pakistan. Total 1501 autopsy reports were collected. They only gave hard copies then digitized. Complete 1501 autopsy reports contain 19 columns. Autopsy reports containing eight causes of death were collected. i.e., Cardiopulmonary Arrest, Diabetes millets, HIV +ve / HIV AIDS, Pneumonia, Hepatitis A/ B, Tuberculosis, PUL-TB / DUL-TB, Respiratory Failure, Sepsis. The whole classes of this data set are 8.

B. EXPERIMENTAL SETUP

For scientific computing, Anaconda could be a free and open-source dispersion of the Python and R programming language. It objects to streamline bundle administration and Dissemination. The transference comprises data-science bundles reasonable for Windows, Linux, and Mac OS.

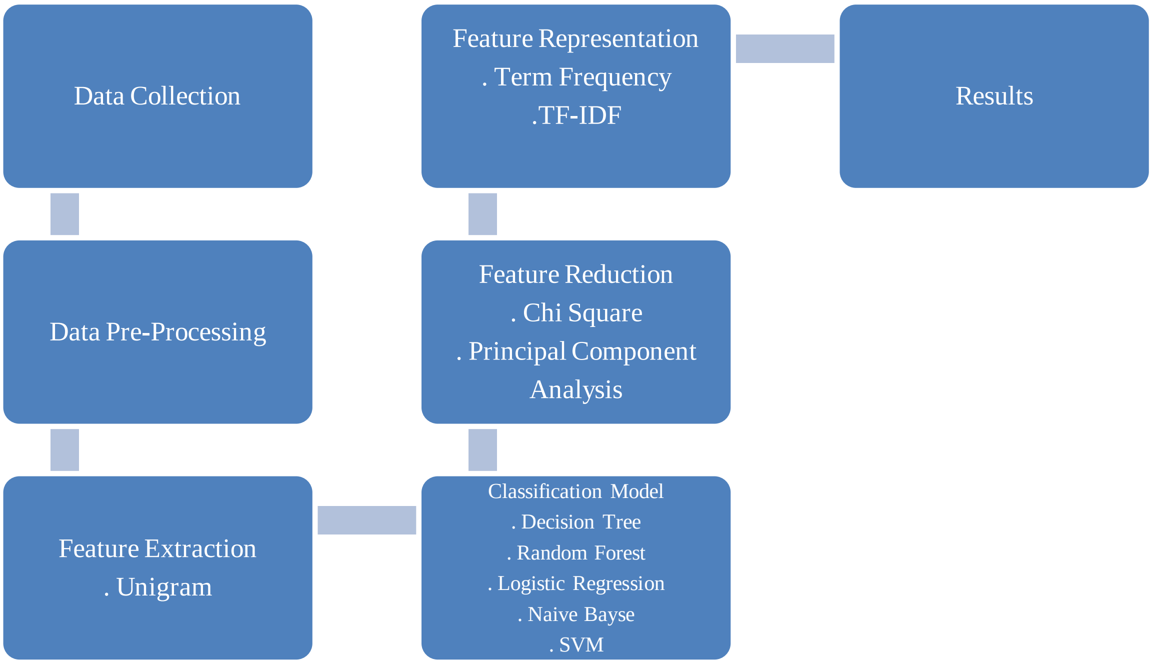

Figure 1. Proposed Technique of Autopsy Report Classification

C. DATA PREPOSSESSING

- *DATA CLEANING:* In data cleaning, all incorrect spellings are corrected through a spellchecker. Then all upper-case letters are converted into lower case letters. In addition, sentences are broken down, and each word is assigned a token. Stop remarks are also removed.

- *Data transformation:* In the digitization process, the data set is obtained in .xls format. That is then transformed to CVS format for further processing.

- *Data partitioning:* The data set is divided into two parts using the Random simple sampling (SRS) technique. For this purpose, I used the" train-test-split" function of the Sklearn Python library. Eighty percent of the data set was chosen to train algorithms, and 20 percent for testing purposes. All proposed algorithms were trained and tested one by one to obtain the results.

D. RESULT

The results and assessments embrace practical standing. These results can assist through a reference for the forthcoming techniques. Results obtained from various combinations are explained here in the form of a table and chart.

- ***Support Vector Machine (SVM):*** A support vector machine is a classification algorithm that paradigms a hyperplane in a high dimensional space. The following flow chart shows the complete working of the SVM classifier. The graphical representation of the SVM classifier working is shown in fig. 2 Table.1 shows a comparison between the proposed technique and existing methodologies based on Text classification techniques. The base papers and the proposed method are estimated by consuming these Evaluation Measures, i.e., Precision, Recall, F-measure, and Accuracy. There is a difference between the feature set size of both base papers and the proposed technique. In the data set of the first base paper, the number of autopsy reports collected and evaluated is 400. In the second base paper, the comprehensive autopsy reports ordered are 482, while in our data set, the total number of autopsy reports is 1501. There are eight classes in our data set, while both base papers contain only four categories.

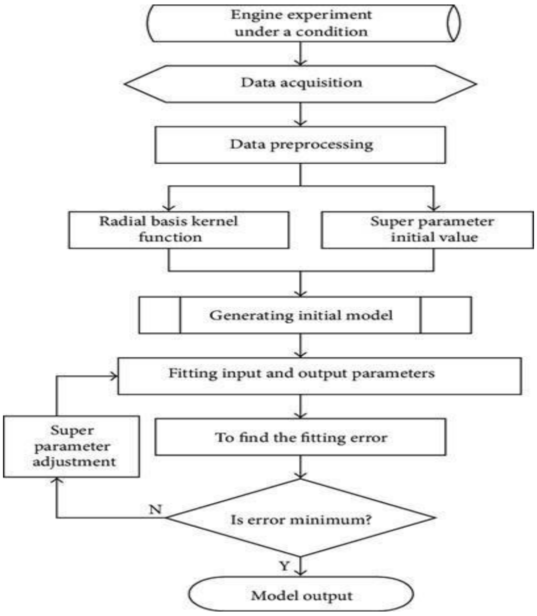
Engine experiment under a condition
Data acquisition
Data preprocessing
Radial basis kernel function
Super parameter initial value
Generating initial model
Fitting input and output parameters
Super parameter adjustment
To find the fitting error
N
Is error minimum?
Y
Model output

Figure 2. Working of SVM Classifier

Evaluation Measures	Base paper 1 (No. of Class = 4)	Base Paper 2 (No. of Class = 4)	Proposed Technique (No. of Class = 8)
Precision	0.781	0.955	0.75
Recall	0.783	0.975	0.81
F- Measures	0.782	0.960	0.77
Accuracy	78.25%	96.33%	80.06%

Table 1. Comparison with Proposed Technique

The accuracy of the first base paper is 78.25%, with 400 autopsy reports and four classes. In comparison, the accuracy of the second base paper is 96.33%, with 482 autopsy reports and four categories shown in Table 1. The accuracy of the proposed technique is 80.06%, which is better than both papers as the proposed technique implemented on the data set contains 1501 autopsy reports and eight classes. As results show, the accuracy of the second base paper is higher than the proposed technique. Still, when analyzed based on classes as the classes of our data set are double, so relatively, the accuracy of the proposed approach is high. Comparison between Performance of SVM for base paper 1 and 2 and proposed technique graphical shown in figure 3.

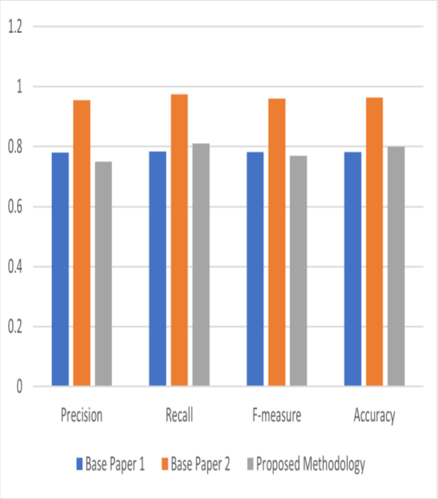

Figure 3. Performance of SVM for base paper 1 and 2 and proposed technique

- ***NB Classifier:*** Naive Bayes is a classification method dependent on Bayes theorem by suppressing individuality amongst interpreters. An NB classifier accepts that a specific component in a class is irrelevant to some additional element. NB classifiers are profoundly versatile, requiring various boundaries linear in the number of factors, i.e., features or predictors in a learning issue. Bayes theorem calculates the posterior probability, P(c—x) from P(c), P(x), and P(x—c). NB classifier accepts that the effect of the value of a predictor (x) for a given class (c) is independent of the importance of additional predictors. This supposition is known as class conditional independence. The results obtained from the NB classifier are shown in the table 2.
  The accuracy of the first base paper is 65.50% which is lower than the proposed technique as the accuracy of the proposed method is 73%, with 1501 autopsy reports and eight classes. Performance of NB Classifier shown in figure 5.

Evaluation Measures	Base Paper 1	Proposed Technique
Precision	0.565	0.66
Recall	0.655	0.73
F-Measure	0.643	0.69
Accuracy	65.50%	73%

Table 2. Comparison with Base Paper

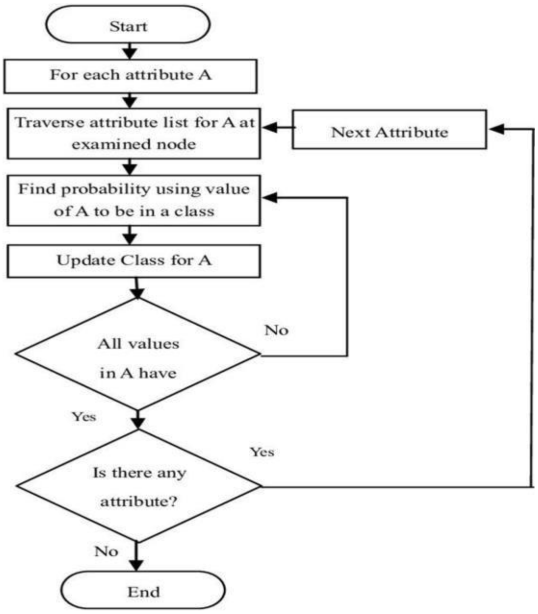

Figure 4. Working of NB Classifier

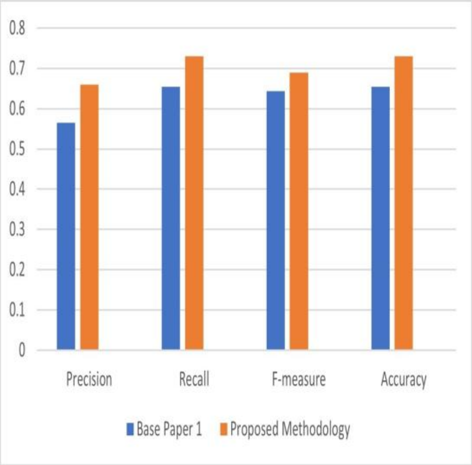

Figure 5. Performance of NB Classifier

VI. Conclusion

Text classification of the plain text autopsy report is done using automatic Text Classification techniques in the proposed method. The experimental results showed that accuracy is improved by reducing feature set size. More autopsy reports are used as a data set as compared to the base paper. More diseases are also used as in base paper, and in previous studies, Classification is done on autopsy reports containing four or a smaller number of infections. While in the proposed method, autopsy reports containing eight causes of death are they are used for classification. When the number of diseases tends to increase, the accuracy usually decreases. The proposed technique holds high-class resolution and better accuracy or performance parameters compared with base papers. The findings of the proposed method contain Para tactical cost and assist as references for the upcoming works. Additionally, existing results will serve as a state of procedures to compare future suggestions with current ATC method. In the future the aim is to predict COD and generate ICD-10 codes. The scope of this work is to classify a more significant number of diseases with more accuracy while reducing future set size.

REFERENCE

[1]. Ghulam Mujtaba, L. S. (2018). Prediction of cause of death from forensic autopsy reports using text classification techniques. Science Direct, 10.

[2]. Bevan Koopman, S. K. (2015). Automatic Classification of diseases from free-text death certificates for real-time surveillance. BMC, 10.

[3]. Bevan Koopman, G. Z. (2018). Extracting cancer mortality statistics from death certificates: A hybrid machine. Science Direct, 10.

[4]. Zahra Shams Khoozani, R. G. (2017). TF-IDF-Based Automated Application for classification Forensic Autopsy Reports to Identification of Cause of Death (COD). IEEE, 6.

[5]. Ghulam Mujtaba, L. S. (2017). Automatic ICD-10 multi-class Classification of cause of death from plaintext autopsy reports through expert driven feature selection. Elsevier, 27.

[6]. Kinga Glinka, R. W. (2017). Improving Multi-Label Medical Text Classification by Feature Selection. IEEE,

[7]. Ghulam Mujtaba, L. s. (2016). Automatic Text Classification of ICD-10 Related COD from Complex and Free Text autopsy reports. IEEE.

[8]. Ghulam Mujtabaa, L. S. (2018). Classification of forensic Autopsy reports through conceptual graph-based. Science Direct, 18.

[9]. Bevan koopman, S, Z. (2015). Automatic ICD-10 Classification of cancers from free-text death certificates. Science Direct.

[10]. Luke Butt, G. Z. (2013). Classification of Cancer-related Death Certificates using Machine Learning. AM, 9.

[11]. Bouree Alix. (2019). "Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France", Google Scholor.

[12]. Whang Yanshan. (2019). "A clinical text classification paradigm using weak supervision and deep representation", Springer.

[13]. Manochander S. (2018). "Scaling feature selection method for enhancing the classification performance of Support Vector Machines in text mining", Google Scholors.

[14]. Ghulam Mujtaba. (2017). "Hierarchical Text Classification of Autopsy Reports to Determine MoD and COD through Term-Based and Concepts-Based Features", Springer.

[15]. Wei-Hung Weng. (2017). "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach", Springer.

[16]. Vijay Garla (2013). "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management", Springer.

[17]. Fleur Mougin. (2016). "Large scale biomedical texts classification: a kNN and an ESA-based approaches"

[18]. Lenivtceva, Iuliia, 2020. "Applicability of Machine Learning Methods to Multi-label Medical Text Classification." International Conference on Computational Science. Springer.

[19]. Danso, S., Atwell, E., & Johnson, O. (2014). A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification. ArXiv, abs/1402.4380.

[20]. Jeblee, S., Gomes, M., Jha, P., Rudzicz, F., & Hirst, G. (2019). Automatically determining cause of death from verbal autopsy narratives. BMC Medical Informatics and Decision Making, 19.