Cancer Literature Classification Methods Performance

Cheng-Yuan Ho
Dept. of Computer Science and Information Engineering,
Asia University,
Taichung City, Taiwan
tommyho@asia.edu.tw

Mariana Syamsudin
Dept. of Computer Science and Information Engineering,
Asia University, Taichung City, Taiwan
Dept. of Information Technology,
Pontianak State Polytechnic, Indonesia
marianasyamsudin@gmail.com

Yueh-Chun Shen
Dept. of Computer Science and Information Engineering,
Asia University,
Taichung City, Taiwan
107121006@live.asia.edu.tw

2020 International Conference on Decision Aid Sciences and Application (DASA) | 978-1-7281-9677-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/DASA51403.2020.9317182

***Abstract*— The literary classification system is the best solution to improve the data search process. In terms of the need, its goal is to compare the relevant biomedical papers and discover novel knowledge to identify potential research issues. This paper will present cancer literature classification performance by comparing three approaches, Naïve Bayes, Neural Network and Linear Classifier with SGD training. The propose approaches classify biomedical literature in five classes of cancer literature type namely, bone cancer, gastric cancer, kidney cancer, skin cancer and papillary thyroid cancer by using 9259 documents. General steps for building classification refer to the classification of scientific literature. The result shows that all algorithms successfully can be used to classify cancer literature. However, for the best performance, it is strongly recommended to use Naïve Bayes and Neural Network.**

***Keywords*—classification performance, naïve Bayes, neural network, linear classification**

I. INTRODUCTION

Regard to the database of the World Health Organization (WHO) annual death registration, cancer is the second leading cause of death globally, 9.6 million deaths in 2018 [12]. Thus, cancer becomes an essential study area for biomedical research. The vast number of biomedical literature has proliferated, provide a rich source of knowledge to improve the development of biomedical research. Figure 1 shows the number of literature on cancer research for the past ten years. The data collected from PubMed using "cancer" as a keyword in the searching of title or abstract.

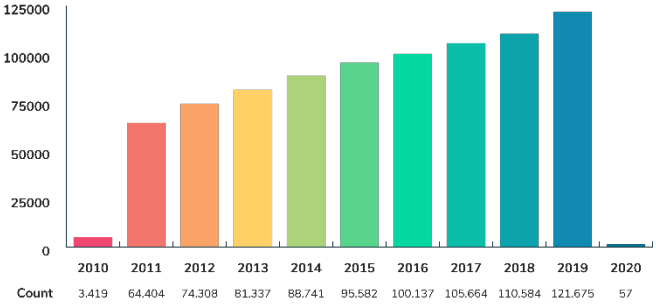

Figure 1. The Number of Cancer Publication in PubMed

Until the beginning of 2020, there are 845.908 cancer publications retrieved from PubMed. Abundant of cancer literature become an essential problem when researchers need to compare the relevant topic papers and discover novel knowledge to identify potential research issues. This study address to help researchers to overcome these difficulties by classifying the cancer literature using text mining methods. Realizing the superiority of the classification of text mining is expected to help cancer researchers find the literature easily and quickly. Sang-Woon Kim [8] used the TF-IDF and LDA scheme to support the paper classification system. This research produces a classification system that has two objectives. The first objective is classifying research papers using keywords and topics with the support of high-performance computing techniques. Then, the papers that have been classified will be applied to search papers in the field of research expeditiously and efficiently. In addition, the classification performance is another concern, and bias can occur if inaccurate disciplinary assignments exist in the scientific classification system especially if there is a significant proportion of multidisciplinary journals in the reference lists [10]. More consideration should be given to the robust and accuracy of classification schemes by consideration of the importance of classification in the construction and analysis of bibliometric indicators [2].

Different from the methods as mentioned above, in the following, we propose to specify literature classification in cancer type that will be remarkably effective to collect and analyze information in biomedical research. This study will also compare the classification performance of three approaches, Naïve Bayes, Neural Network, and Linear Classifier with SGD training. The Naïve Bayes approach was successfully applied to classify scientific literature into predetermined categories, according to the needs of the researchers. Furthermore, this approach increases the effectiveness of work in identifying and solving potential research problems and dramatically facilitates scientific research. [8]. In recent years, neural network-based approaches become potential and dominant methodology in biomedical relation classification. The advantage of Neural network-based approaches that is can effectively and automatically learn the underlying feature representation from the labelled training data [11]. **Stochastic Gradient Descent** (SGD) has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Furthermore, through SGD convex loss functions, such as (linear) Support Vector Machines and Logistic Regression, can be learned in a profoundly efficient approach to discriminative learning of linear classifiers [7].

There are five types of cancers literature considered as a sample of classification problem, namely, Gastric cancer, Skin cancer, Bone Cancer, Kidney cancer, and Papillary Thyroid

The work is supported by the Ministry of Science and Technology of Taiwan, under Grant no. MOST 109-2221-E-468 -008 -MY3, and partially sup-ported by the Ministry of Education of Taiwan, the Artificial Intelligence Talent Cultivation Project, and by Asia University, under Grant no. 107-ASIA-UNAIR-09.

Cancer. All corpus in the form of title and abstracts, in total 9,259 documents.

This project has six sections. It starts from converting data from xml format to csv format, performing vectorization, removing stop words, stemming, doing term frequency-inverse document frequency (TF*IDF), and training as well as testing all data in Jupyter Notebook and Python environment.

Training and testing process will be carried out in three approaches. Respectively, testing for 80% of training dataset and 20% of the testing dataset, 70% of training dataset and 30% of testing dataset, and 60% of training dataset and 40% of the testing dataset.

II. The Procedures for Classification

Referring to the classification of scientific literature, general steps for building a classification model as presented in Figure 2 are [9]:

A. Determining a classification scheme
B. Reading text documents
C. Text preprocessing
D. Classification

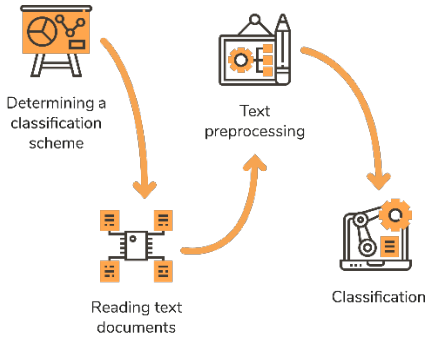

Figure 2. General Steps of Classification Process

A. Determining Type of Cancer Literature

To guarantee the results validity of the design, it is essential to do stages in the design shown in Figure 3.

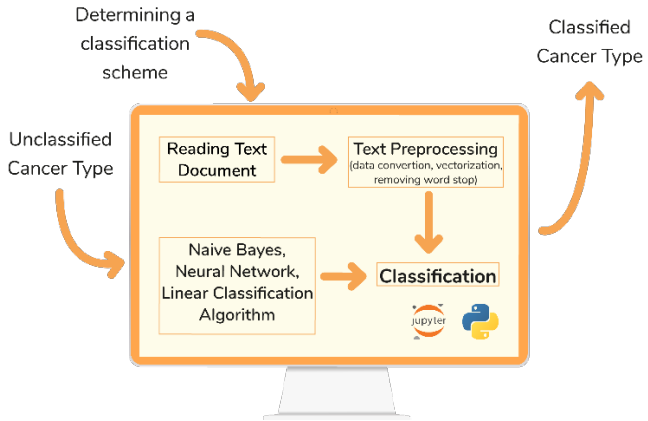

Figure 3. Model Architecture for Cancer Literature Classification

The stage determining a classification scheme is related to the development of a classifier of cancer literature. They are formed according to the five types of cancer literature that will be analyzed in this study are: Bone Cancer, Gastric Cancer, Kidney Cancer, Papillary Thyroid Cancer and Skin Cancer.

The following is a general explanation of each type of cancer. The first type is bone cancer. Many kinds of cancer can begin in the bones. The most common types of bone cancer in children and adolescents are osteosarcoma and Ewing's tumor. Bone cancer is divided into two types, namely primary and secondary bone cancer. Primary bone cancer exists and grows in bone cells, while secondary bone cancer starts from elsewhere and spreads to the bone. The second type of cancer is gastric cancer, which is a type of disease in the stomach lining. The symptoms of gastric cancer can be in the form of indigestion and stomach discomfort or pain. The third type is kidney cancer. It is one of the most common types of cancer in adults in their 60s or 70s. This cancer is also known as kidney cancer since it first appeared in a small tube lining in the kidney. The fourth type is papillary thyroid cancer. It is recognized as an asymptomatic disease with a mass in the neck as the most common symptom. The last type is skin cancer, which occurs because of long-term skin in the sun. There are three main types of diseases such as basal cell carcinoma, squamous cell carcinoma, and myeloma caused by abnormal growth of skin cells [1].

B. Text Documents

This phase is used to present the text documents in a clear word format. All text document resources extracted from PubMed.gov, https://www.ncbi.nlm.nih.gov/pubmed/. The advanced search feature is used to collect tittle and abstract of literature from a set of massive papers efficiently as well as to get specific data by selecting Date – Publication, Title/Abstract, and name of cancer in the searching box.

In this paper, we utilized the title and abstract data of research papers, since these two parts of papers are the most part that users read to catch the main idea of the research before reading other contents in the body of a paper [8].

The number of data collected varies according to availability on PubMed servers and result of data cleaning processing. The data mapping can be seen in Figure 4.

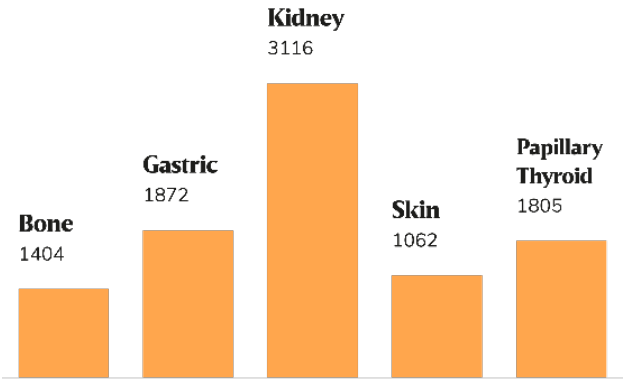

Figure 4. Data Collecting

The data in the form of titles and abstracts were collected, starting from the smallest to the most significant number in the sequence are as follows: 1,062 documents of skin cancer, 1,404 documents of bone cancer, 1,805 documents of papillary thyroid cancer, 1,872 documents of gastric cancer, and 3,116 documents of kidney cancer.

C. Text Preprocessing

The stage of text pre-processing starts from extracting Abstract Title and Abstract Text, then converting xml files into csv files. The sample code and result are presented in Figures 5 and 6 below:

Source Code

```
import glob

from xml.etree import ElementTree

counter = 0
title_list = []
abstract_list = []

directory = '/Users/cg/Google Drive/Codes/Python/data-anna'
for file in glob.iglob(directory + "*.xml"):
    with open('file', 'rt') as f:
        tree = ElementTree.parse(f)
        root = tree.getroot()


for title in root.iter('ArticleTitle'):
    title_list.append(title.text)
for abstract in root.iter('AbstractText'):
    abstract_list.append(abstract.text)


for i in range(0, len(title_list)):
    print(i)
    print('Title: ', title_list[i])
    print('Abstract: ', abstract_list[i])
```

Result

```
0
Title:  Downregulation of Cdc6 inhibits tumorigenesis of osteosarcoma in vivo and in vitro.
Abstract:  Osteosarcoma is the most frequent malignant bone cancer in teenagers. In this study, cell division cycle 6 (Cd
c6) was upregulated in the tumor tissues from patients with osteosarcoma according to the results of RNA sequencing. Cdc6
has been considered as a candidate prognostic marker, which was associated with cell apoptosis and cell cycle. However, t
he mechanisms by which Cdc6 regulates the apoptosis and cell cycle arrest in osteosarcoma remain unclear.
1
Title:  [Study on analgesic effect and mechanism of cinobufagin on rats with bone cancer pain].
Abstract:  Firstly, 3 pairs of osteosarcoma and adjacent normal tissues were subjected to RNA sequencing. In addition, th
e levels of Cdc6 in 30 pairs of patients' tissues and in osteosarcoma cell lines were detected by western blotting and R
T-qPCR. Moreover, the effects of Cdc6 on cell proliferation, apoptosis, cell cycle and invasion were evaluated by cell co
unting kit-8 (CCK-8), flow cytometric and transwell assay, respectively.
2
Title:  [Relationship between expression of high-mobility group box-1 and inflammatory cytokines in patients with bone ca
ncer pain].
Abstract:  Cdc6 was significantly upregulated in patients with osteosarcoma and in osteosarcoma cell lines. Downregulatio
n of Cdc6 inhibited the proliferation and invasion of MG63 cells by promoting apoptosis. In addition, downregulation of C
dc6 induced G1 phase cell cycle arrest. Moreover, inhibition of Cdc6 downregulated the expression of Cyclin D1 and Cyclin
A2, and upregulated the level of p21 in MG63 cells. In vivo study confirmed downregulation of Cdc6 inhibited osteosarcoma
```

Figure 5. The Process of Data Extracting

Data in CSV Format

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	title,abstract,cancertype																				
2	Downregulation of Cdc6 inhibits tumorigenesis of osteosarcoma in vivo and in vitro.,"Our findings indicated that Cdc6 may serve as an oncogene in osteosarcoma. Thus, inhibiting Cdc6 may be a therapeutic approach																				
3	[Study on analgesic effect and mechanism of cinobufagin on rats with bone cancer pain],,bone																				
4	[Relationship between expression of high-mobility group box-1 and inflammatory cytokines in patients with bone cancer pain],,bone																				
5	Incidence and mortality of bone cancer among children, adolescents and young adults of Brazil.,"Osteosarcoma and Ewing Sarcoma are the most incident bone cancers in all Brazilian regions. Bone cancers showed in																				
6	Suppression of WNK1-SPAK/OSR1 attenuates bone cancer pain by regulating NKCC1 and KCC2.,Our preliminary experiment indicated the activation of with-no-lysine kinases 1 (WNK1) in bone cancer pain (BCP) rats.																				
7	A Role for Nf- Rß in Organ Specific Cancer and Cancer Stem Cells.,"Cancer stem cells (CSCs) account for tumor initiation, invasiveness, metastasis, and recurrence in a broad range of human cancers. Although being a																				
8	12b80 - Hydroxybisphosphonate Linked Doxorubicin: Bone Targeted Strategy for Treatment of Osteosarcoma.,"To reply to as yet unmet medical needs to treat osteosarcoma, a form of primary bone cancer, we conce																				
9	Engineering osteoblastic metastases to delineate the adaptive response of androgen-deprived prostate cancer in the bone metastatic microenvironment.,"While stromal interactions are essential in cancer adaptati																				
10	Curcumin Releasing Eggshell Derived Carbonated Apatite Nanocarriers for Combined Anti-Cancer, Anti-inflammatory and Bone Regenerative Therapy.,Bone cancer or osteosarcoma is an aggressive cancer affecting t																				
11	Serum B-cell maturation antigen (BCMA) reduces binding of anti-BCMA antibody to multiple myeloma cells.,"B-cell maturation antigen (BCMA), a tumor necrosis factor receptor (TNFR) family member, is selectively																				
12	Mesoporous superparamagnetic hydroxyapatite nanocomposite: A multifunctional platform for synergistic targeted chemo-magnetotherapy.,"In the present study, the aim was to develop a magneto-responsive na																				
13	Triple- threat activity of PEDF in bone tumors: Tumor inhibition, tissue preservation and cardioprotection against doxorubicin.,"Pigment epithelium-derived factor (PEDF) is known for its osteogenic properties, but it																				
14	Transcript. It is one of the most common growth factors in bone tissue,bone																				
15	High-resolution large-area imaging of nanoscale structure and mineralization of a sclerosing osteosarcoma in human bone.,"Osteosarcoma is the most common primary bone cancer type in humans. It is predominan																				
16	The Endocannabinoid/Endovanilloid System in Bone: From Osteoporosis to Osteosarcoma.,"Bone is a dynamic tissue, whose homeostasis is maintained by a fine balance between osteoclast (OC) and osteoblast (OB																				
17	Xfm1595 ameliorates bone cancer pain in rats via inhibiting p38-mediated glial cell activation and neuroinflammation in the spinal dorsal horn.,Bone cancer pain (BCP) profoundly compromises the life quality of pat																				
18	Incidence and mortality of primary bone cancers in China, 2014.,"This population-based study presents the most recently available estimates on primary bone cancers in China, revealing that the males are 1.34 times																				
19	Fertility optimization in women with cancer: from preservation to contraception.,"Advances in the early diagnosis and treatment of cancer have reduced mortality rates and improved patient survival. For this reason																				
20	(Electroacupuncture intervention relieves pain possibly by promoting MOR endocytosis in locus coeruleus in bone cancer pain rats with morphine tolerance).,"EA intervention can relieve pain and MT in bone cancer																				
21	Hydrophobicity may enhance membrane affinity and anti-cancer effects of Schiff base vanadium(v) catecholate complexes.,"Anti-cancer activities of vanadium compounds have generated recent interest because c																				

Figure 6. The Data in CSV Format

The output documents are then provided for the next phase in text classification that includes:

Tokenization: Before partitioning into the token list, the document is treated as a string in advance.

Removing stop words: The process to remove the insignificant words or stop words such as “the”, “a”, “and”, etc.

Stemming word: This step is the process of conflating tokens to their root form. Converts different word forms into a similar canonical form can be done by applying the stemming algorithm, e.g. change the word “Classification” and “Classified” into the word “Classify”.

The related program codes of this step are depicted in Figure 7. The next step can be seen in Figure 8, and it is responsible for inversing term/document frequency (TF*IDF). The TF-IDF has been widely employed to evaluate the relationship of each word in the collection of documents in the fields of information retrieval and text mining.

D. Classification

The classification will be applied in three treatments with three methods: Neural Network, Naïve Bayes, and Linear Classifier with SGD training. Moreover, in the following subsections, we used a diverse number to indexes each cancer literature, (1) for Bone Cancer, (2) for Gastric Cancer, (3) for Kidney Cancer, (4) for Papillary Thyroid Cancer, and (5) for Skin Cancer, respectively.

Code of Data Removing

```
#Text Preprocessing
columns = ['index','cancertype', 'Abstract']
df_ = pd.DataFrame(columns=columns)
#Lower string
data['Abstract'] = data['Abstract'].str.lower()
#remove email adress
data['Abstract'] = data[str('Abstract')].replace('[a-zA-Z0-9-_.]+@[a-zA-Z0-9-_.]+', '', regex=True)
#remove IP address
data['Abstract'] = data['Abstract'].replace('((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)(\.|$)){4}', '', regex=True)
#remove punctuations and special chracters
data['Abstract'] = data['Abstract'].str.replace('[^\w\s]','')
#remove numbers
data['Abstract'] = data['Abstract'].replace('\d', '', regex=True)
#remove stop words
for index, row in data.iterrows():
    word_tokens = word_tokenize(row[str('Abstract')])
    filtered_sentence = [w for w in word_tokens if not w in stopwords.words('english')]
    df_ = df_.append({"index": row['index'], "cancertype":  row['cancertype'],"Abstract": " ".join(filtered_sentence[0:])},
data = df_

X_train, X_test, y_train, y_test = train_test_split(data['Abstract'].values.astype('U'),data['cancertype'].values.astype('U'
classes  = data['cancertype'].unique()
```

```
#Text Preprocessing
columns = ['index','cancertype', 'Abstract']
df_ = pd.DataFrame(columns=columns)
#Lower string
data['Abstract'] = data['Abstract'].str.lower()
#remove email adress
data['Abstract'] = data[str('Abstract')].replace('[a-zA-Z0-9-_.]+@[a-zA-Z0-9-_.]+', '', regex=True)
#remove IP address
data['Abstract'] = data['Abstract'].replace('((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)(\.|$)){4}', '', regex=True)
#remove punctuations and special chracters
data['Abstract'] = data['Abstract'].str.replace('[^\w\s]','')
#remove numbers
data['Abstract'] = data['Abstract'].replace('\d', '', regex=True)
#remove stop words
for index, row in data.iterrows():
    word_tokens = word_tokenize(row[str('Abstract')])
    filtered_sentence = [w for w in word_tokens if not w in stopwords.words('english')]
    df_ = df_.append({"index": row['index'], "cancertype":  row['cancertype'],"Abstract": " ".join(filtered_sentence[0:])},
data = df_

X_train, X_test, y_train, y_test = train_test_split(data['Abstract'].values.astype('U'),data['cancertype'].values.astype('U'
classes  = data['cancertype'].unique()
```

Figure 7. Code of Data Removing

TF*IDF

```
vectorizer = TfidfVectorizer(analyzer='word',ngram_range=(1,2), max_features=50000,max_df=0.5,use_idf=True, norm='l2')
counts = vectorizer.fit_transform(X_train)
vocab = vectorizer.vocabulary_
```

Figure 8. TF*IDF

At the stage of classification, the researchers divided the data into two parts with different proportions. Specifically, the proportions of training dataset and testing data are 80% and 20% (called 0.2 for short), 70% and 30% (0.3), and 60% and 40% (0.4), respectively. Figures 9 ~ 14 show the results for each method’s model accuracy and confusion matrix form in 0.3 proportion, while the complete result of accuracy is presented in Table 1.

Model Accuracy:0.89

	precision	recall	f1-score	support
bone	0.90	0.84	0.87	409
gastric	0.85	0.87	0.86	560
kidney	0.88	0.93	0.91	943
papillary	0.96	0.88	0.92	325
skin	0.88	0.88	0.88	541
micro avg	0.89	0.89	0.89	2778
macro avg	0.90	0.88	0.89	2778
weighted avg	0.89	0.89	0.89	2778

Figure 9. Model Accuracy of Neural Network with Proportion 0.3

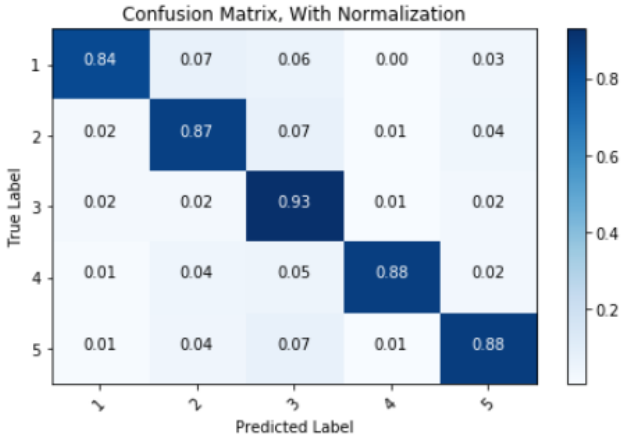

Figure 10. Confusion Matrix of Neural Network with Proportion 0.3

Model Accuracy:0.89

	precision	recall	f1-score	support
bone	0.89	0.83	0.86	409
gastric	0.86	0.86	0.86	560
kidney	0.88	0.93	0.91	943
papillary	0.97	0.87	0.92	325
skin	0.87	0.88	0.88	541
micro avg	0.89	0.89	0.89	2778
macro avg	0.89	0.88	0.88	2778
weighted avg	0.89	0.89	0.89	2778

Figure 11. Model Accuracy of Naïve Bayes with Proportion 0.3

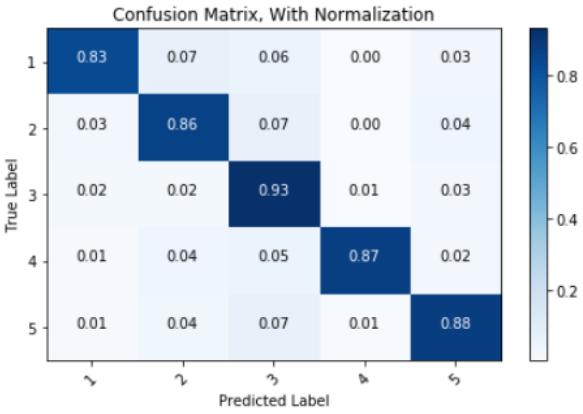

Figure 12. Confusion Matrix of Naïve Bayes with Proportion 0.3

Model Accuracy:0.89

	precision	recall	f1-score	support
bone	0.91	0.83	0.87	409
gastric	0.86	0.87	0.86	560
kidney	0.88	0.93	0.90	943
papillary	0.96	0.88	0.92	325
skin	0.88	0.88	0.88	541
micro avg	0.89	0.89	0.89	2778
macro avg	0.90	0.88	0.89	2778
weighted avg	0.89	0.89	0.89	2778

Figure 13. Model Accuracy of Linear Classifier with Proportion 0.3

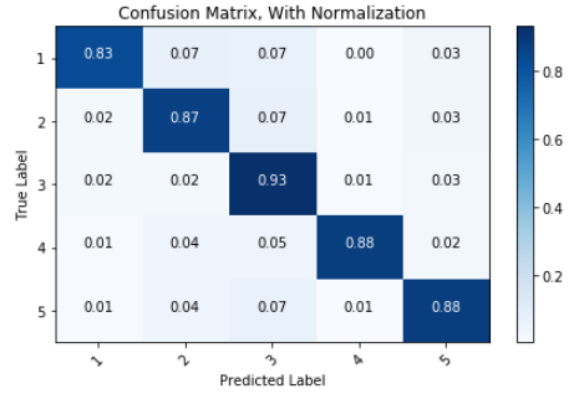

Figure 14. Confusion Matrix of Linear Classifier with Proportion 0.3

TABLE I. RESULT OF ACCURACY FOR 9,259 INSTANCES

Mehtod	Testing Dataset 0.2	0.3	0.4
Naive Beyes	0.89	0.89	0.89
Neural Network	0.89	0.89	0.89
Linear Classification	0.89	0.89	0.89

III. EXPERIMENTAL RESULTS

A. The Analysis of Accuracy Result

After going through the stages of determining the classification scheme, reading text documents, text preprocessing and classification, the results for each condition, i.e., 0.2, 0.3, and 0.4, are obtained as shown in Table 2 below:

TABLE II. RESULT OF ACCURACY FOR HUNDREDS INSTANCES

Mehtod	Testing Dataset 0.2	0.3	0.4
Naive Beyes	0.90	0.97	0.93
Neural Network	0.90	0.97	0.93
Linear Classification	0.90	0.93	0.90

Each method manifests the same performance for an accuracy result of 0.89/1 but produces different confusion matrix that will explain further in the analysis of confusion matric section. This condition presents the initial hypothesis that more number of instances result in more stable accuracy. The researchers also reclassified documents with fewer data to examine the performance of each algorithm. As shown in Table 2, the result of accuracy shows more vary for each method when using hundreds of data instances.

The data describe all methods have the same accuracy performance when using 20% testing dataset, each 0.9/1. While Naïve Bayes and Neural Network present higher performance than Linear Classification when using 30% and 40% testing dataset, each respectively at 0.97/1 and 0.93/1.

B. The Analysis of Confusion Matrix

The result of the confusion matrix is presented in Table 3 below. In this confusion matrix, the system manages to classify each cancer literature with average accuracy 0.84/1 for Bone cancer, Gastric cancer, Papillary cancer and Skin cancer. In contrast, Kidney cancer gets the highest accuracy on average 0.93/1. This condition is highly possible because of a more significant amount of kidney dataset than other cancer literature. For reference, Kidney cancer literature has 3116 datasets, whereas four other types of cancer literature have approximately only a thousand dataset.

TABLE III. RESULT OF CONFUSION MATRIX

Cancer Type	0.2 NB	0.2 NN	0.2 LC	0.3 NB	0.3 NN	0.3 LC	0.4 NB	0.4 NN	0.4 LC
Bone	0.83	0.83	0.83	0.83	0.84	0.83	0.83	0.83	0.83
Gastric	0.84	0.86	0.86	0.88	0.87	0.86	0.89	0.87	0.86
Kidney	0.94	0.93	0.93	0.94	0.93	0.94	0.94	0.93	0.93
Papillary	0.87	0.87	0.89	0.87	0.88	0.89	0.87	0.88	0.89
Skin	0.88	0.88	0.87	0.89	0.88	0.87	0.89	0.88	0.87

Figure 15 shows a comparison of the confusion matrix of three types of the algorithm more precisely with the proportion of 0.3 testing data.

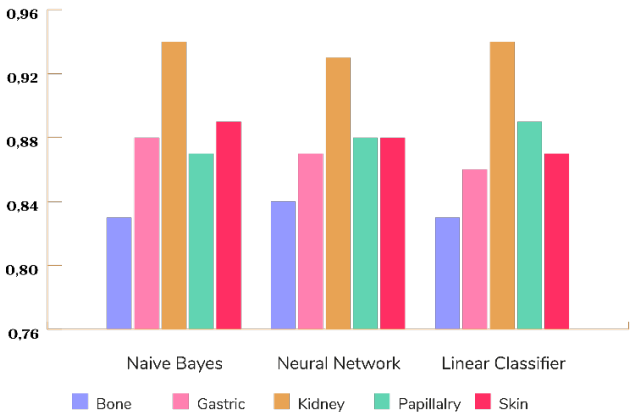

Figure 15. Comparison of Accuracy Performance (0.3)

C. The Analysis of Classification Result

Table 4 depicts information about the mapping of correct and incorrect instances. The data shows incorrect instances for the literature of Gastric cancer, Papillary Thyroid Cancer and Skin Cancer tend to kidney cancer literature in the

number of 0.07/1. In addition, bone cancer literature tends to map inaccurate cases for Gastric cancer as well as Kidney cancer in the number of 0.07/1. On the other hand, incorrect instances of Kidney cancer literature itself tend to skin cancer with the highest missed classification in the number of 0.03/1.

TABLE IV. RESULT OF INCORRECTLY INSTANCES

	0.2					0.3					0.4				
	Bone	Gastric	Kidney	Papillary	Skin	Bone	Gastric	Kidney	Papillary	Skin	Bone	Gastric	Kidney	Papillary	Skin
Bone	0.83	0.07	0.06	0.00	0.03	0.83	0.07	0.07	0.00	0.03	0.83	0.06	0.06	0.00	0.04
Gastric	0.02	0.88	0.06	0.00	0.03	0.02	0.87	0.07	0.01	0.03	0.02	0.86	0.07	0.01	0.04
Kidney	0.02	0.01	0.94	0.00	0.02	0.02	0.02	0.93	0.01	0.03	0.02	0.01	0.93	0.01	0.02
Paillary	0.00	0.06	0.05	0.87	0.02	0.01	0.04	0.05	0.88	0.02	0.00	0.03	0.05	0.89	0.02
Skin	0.00	0.04	0.07	0.00	0.89	0.01	0.04	0.07	0.01	0.88	0.01	0.04	0.07	0.00	0.87

Correctly Instances

The Most Incorrectly Instances

The data above were obtained from the confusion matrix summary of the neural network, with sequential proportions of 0.2, 0.3 and 0.4. Since it is considered as the best algorithm for classification of cancer literature.

IV. CONCLUSION

Based on confusion matrix analysis, the result indicates that type of cancer literature with the most significant number of instances will show the highest accuracy, that is Kidney Cancer literature and followed by other types of cancer literature based on their number of instances. Meanwhile, the mapping of incorrect instances confirms that the highest missed classification of each cancer literature tend to Kidney cancer while Papillary Cancer literature becomes the lowest number of missed classification.

Incorrect instances can be prompted by the relationship between two different cancer types, for example, incorrect literature interpretation of Bone Cancer and Gastric Cancer caused by relational of the research topic of both cancer, that usually Gastric Cancer initially presenting as bone metastasis.

Furthermore, the disambiguation classification of cancer type can also be provoked by strong association from each cancer type, as shown by Kidney Cancer. Considering the treatment for cancer survivor will use a considerable amount of medicine that will harm the patient's liver, Kidney Cancer will be the most common synchronous clinical effect in patients. This case vice versa with Papillary Thyroid cancer, wherein only 0.00 ~ 0.01 will likely to be wrong mapped as Papillary Thyroid because of a weak relationship between papillary cancer with other cancer topics, in this case, Bone Cancer, Gastric Cancer, Kidney Cancer and Skin Cancer.

The results of this study showed that all algorithms successfully could be used to classify cancer literature. However, for the best performance, it is strongly recommended to use Naïve Bayes and Neural Network methods by dividing 70% training dataset and 30% testing dataset.

REFERENCES

[1] American Cancer Society, https://www.cancer.org/cancer/bone-cancer/about/what-is-bone-cancer.html

[2] F. Colas and P. Brazdil, "Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks," in Bramer, M. (Ed.), Artificial Intelligence in Theory and Practice. Springer, Boston, MA, pp. 169-178, 2006.

[3] F. Shu, C.-A. Julien, L. Zhang, J. Qiu, and V. Larivière, "Comparing Journal and Paper Level Classifications of Science," Journal of Informetrics, vol. 13, issue 1, pp. 202-225, February 2019.

[4] W.W.M. Fleuren and W. Alkema, "Application of Text Mining in the Biomedical Domain," Methods, vol. 74, pp. 97-106, 2015.

[5] M. Mowafy, A. Rezk, and HM. El-bakry, "An Efficient Classification Model for Unstructured Text Document," American Journal of Computer Science and Information Technology, ISSN 2349-3917.

[6] P. Jin, Y. Zhang, X. Chen, and Y. Xia, "Bag-of-Embeddings for Text Classification," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2016.

[7] PubMed.gov, https://www.ncbi.nlm.nih.gov/pubmed/

[8] S.-W. Kim and J.-M. Gil, "Research Paper Classification System Based on TF-IDF and LDA Schemes," Human-centric Computing and Information Sciences, vol. 9, no. 30, 2019.

[9] Scikit Learn, https://scikit-learn.org/stable/modules/sgd.html

[10] S. Sulova, L. Todoranova, B. Penchev, and R. Nacheva, "Using Text Mining to Classify Research Paper," in Proceedings of 17th International Multidisciplinary Scientific GeoConference (SGEM 2017), vol. 17, pp. 647-654, 2017.

[11] StackOverflow, https://stackoverflow.com/questions/.

[12] World Health Organization, https://www.who.int/health-topics/cancer#tab=tab_1

[13] L. Zhang, F. Janssens, L. Liang, and W. Glänzel, "Journal Cross-Citation Analysis for Validation and Improvement of Journal-based Subject Classification in Bibliometric Research," Scientometrics, vol. 82, issue 3, pp. 687-706, 2010.

[14] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Sun, B. Xu, and Z. Zhao, "Neural Network-based Approaches for Biomedical Relation Classification: A Review," Journal of Biomedical Informatics, vol. 99, 103294, 2019.