RESEARCH ARTICLE

Selcuk Goksel Toplu[1]
Sengul Cangur[2]

[1]Duzce University, Health Sciences Institute, Department of Biostatistics and Medical Informatics, Turkey
[2]Duzce University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Turkey

Corresponding Author:
Sengul Cangur
Duzce University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Turkey
mail: sengulcangur@duzce.edu.tr
Phone: +90 5375956051

Received: 09.03.2020
Acceptance: 03-04-2020
DOI: 10.18521/ktd.700789

Konuralp Medical Journal
e-ISSN1309–3878
konuralptipdergi@duzce.edu.tr
konuralptipdergisi@gmail.com
www.konuralptipdergi.duzce.edu.tr

Text Mining Method in the Field of Health

ABSTRACT

Objective: Text mining which digitalizes textual data and enables them to be applied for text mining algorithms has a very important place in today's world. The aim of this study was to introduce the text mining method and to show its application on a subject in the field of health.

Methods: The text mining method was applied to the documents obtained separately from the most frequently used Pubmed database under two different titles as "human-and-cancer" and "mouse-and-cancer", and then to the combined documents, through the Knime program. Afterwards, the document classification was made using K nearest neighbor (K-NN) algorithm.

Results: The prominent words were "cell" and "cancer" in tag cloud graphs. In both documents, the words such as "cell", "cancer", "tumor", "patient", whose frequency values were high, were observed to be high rates in the analysis performed after the data was merged. It was found that 255 of 600 test documents belonged to the human-and-cancer class and the remaining belonged to the mouse-and-cancer class, and the accuracy classification was 56.6% for the human-and-cancer-documents and 62.6% for the mouse-and-cancer-documents according to the F-criteria. It was determined that the document classification estimation by the K-NN algorithm was relatively successful with a rate of 59.8% however Cohen's kappa value was 19.7%, meaning that the fit was of a slight level.

Conclusions: It was recommended to use the text mining method and to generalize its use in order to obtain information quickly and reliably in the health field where there were numerous digital and printed documents.

Keywords: Text Mining, Classification, Natural Language Processing, Pubmed

Sağlık Alanında Metin Madenciliği Yöntemi

ÖZET

Amaç: Metinsel verileri sayısal hale getirerek veri madenciliği algoritmalarına uygulanmasını sağlayan metin madenciliği, günümüz dünyasında önemli bir yere sahiptir. Bu çalışmanın amacı, metin madenciliği yöntemini tanıtmak ve sağlık alanında belirlenen bir konuda uygulamasını göstermektir.

Gereç ve Yöntem: Çalışmanın uygulama aşamasında; insan-ve-kanser" ve fare- ve-kanser" şeklinde belirlenen iki farklı konu başlığı altında en sık kullanılan Pubmed veritabanından ayrı ayrı elde edilen dokümanlara ve daha sonra birleştirilmiş dokümanlara Knime programı aracılığıyla metin madenciliği yöntemi uygulanmıştır. Ardından K en yakın komşu (K-NN) algoritması kullanılarak doküman sınıflaması yapılmıştır.

Bulgular: Etiket bulut grafiklerinde öne çıkan kelimeler "cell" (hücre) ve "cancer" (kanser) kelimeleridir. Her iki dokümanda frekans değeri yüksek çıkan "cell", "cancer", "tumor", "patient" gibi kelimelerin veriler birleştirildikten sonra yapılan analizde de yüksek oranla çıktığı gözlenmiştir. 600 adet test dokümanının 255 tanesi insan-ve-kanser sınıfına, geri kalanının ise fare-ve-kanser sınıfına ait oldukları; F ölçütüne göre insan-ve-kanser dokümanları için %56,6'lık, fare-ve-kanser dokümanları için ise %62,6'lık doğru sınıflandırılma yüzdesi tespit edilmiştir. K-NN algoritması ile %59,8 oranında kısmen başarılı bir doküman sınıflama tahmini yapıldığı, ancak Cohen kappa değerinin %19,7 olduğu ve bu uyumun zayıf düzeyde olduğu belirlenmiştir.

Sonuç: Dijital ve basılı dokümanların sayısının oldukça fazla olduğu sağlık alanında hızlı ve güvenilir bir şekilde bilgi elde edebilmek için metin madenciliği yönteminden yararlanılması ve kullanımının yaygınlaştırılması önerilmektedir.

Anahtar Kelimeler: Metin Madenciliği, Sınıflandırma, Doğal Lisan İşleme, Pubmed

INTRODUCTION

Today, there are numerous digital and printed/written documents. These digital documents, which contain large-scale unstructured data, can be exemplified as web pages, e-mails and digitalization of written documents. Processing and analyzing these unstructured data may differ from digital data (1). When it comes to analyzing big data, the first thing that springs to mind is data mining approaches. However, data mining applications are mostly carried out on structural data. Thus, where unstructured data consisting of only text are converted to structured data, text mining comes into play. This method is the process of analyzing text to derive some information from textual and unstructured documents for personal or special purposes (2). In other words, this method structures unstructured data using text-format data and derives numerical data from texts to obtain information It is also called as "text data mining" and "knowledge discovery from textual databases" (3).

In text mining, researchers can analyze regular data as well as textual data from articles, texts on websites, medical reports, invoice details (1). Thanks to their capability to automatically identify various semantic information, text mining techniques may help using the relations between simultaneous concept formats and concepts (4). It has become a method commonly used in the fields of health, education, legal, customer relations, market surveys, and internal security to analyze numerous digital texts in a short time and to reach qualified information quickly (5). In particular, in recent studies on health, this method has been increasingly used (4, 6-13).

The aim of this study was to introduce text mining method which researchers and analysts have used to extract information by analyzing the existing documents in the field of health and to demonstrate its application on a subject in the field of health. In the application of the study, text mining method was applied to the documents obtained separately from the most frequently used Pubmed database under two different titles as "human-and-cancer" and "mouse-and-cancer", and then to the combined documents, through the Knime program. Subsequently, how the document classes were created using the K nearest neighbor (K-NN) algorithm was explained in detail.

MATERIAL AND METHODS

1. Text Mining and Process Phases of Text Mining: Text mining is the process of uncovering unspecific, hidden qualified information in textual data and structuring non-regular data (14). The basic strategy of computation mechanism that processes text-based information is to reduce natural language inputs being too much into a set of small categories (5). Text mining applications may include information retrieval, natural language processing (NLP), named entity recognition, pattern identified entities, coreference, relation, rule, event extractions, and sentimental analysis (9,15). The cross industry standard process for data mining (CRISP-DM), the most commonly used process in data mining, is also preferred in text mining. This process model consists of a 6-phase cycle (15):

I. Determining the Aim of the Study: As in every study, firstly, the aim of the study is determined in text mining.

II. Discovering the Availability and Nature of Data: In this phase; the source of textual data is determined, the accessibility and availability of data are evaluated, the first data set is collected, the enrichment of data is examined, and the certainty and quality of data are evaluated.

III. Preparation of Data: Preparation of data set to be used in the project for modeling purpose involves performance of any modifications (15). Tokenization process is usually required to obtain words in a text. Numbers, punctuation marks, tables, figures, images, repeated and white spaces should be removed from the text to structure the corpus (16). In order to reduce the size of data structures, various pre-processing methods such as filtering, lemmatization or stemming may be used.

a-Filtering method is the removal of words that have no meaning or sentiment status by itself, like prepositions, conjunctions, articles etc. (17).

b-Lemmatization methods are used to convert plural nouns to their singular forms or usually convert conjugations into their infinitive form. Because it is expensive, difficult and error-prone, stemming methods are more preferred in practice (17).

c-Stemming method is used to convert the words into their simple form (17). For this purpose, Zemberek which is an open-source, platform independent, general purpose NLP was designed for Turkish (18). Furthermore, there are stemming algorithms such as ITU-NLP and Kemik developed by some universities in Turkey (19,20).

Pre-processing phase is performed especially to make more precise and qualified analysis by exploring the natural structure of the data, and to produce more useful and meaningful information from the data (21). In particular, besides removal of punctuation marks and transformation of all words to lower cases in Turkish texts; some additional preliminary preparations such as creating and editing wild card words and keywords are required. The dictionary is updated with joker words obtained by the wild card method. Each document including joker words is showed with the weighting of the vector in the size of all words in the dictionary. Many techniques (term frequency (TF), inverse document frequency (IDF), term frequency-inverse document frequency (TF-IDF), term parsing value, probabilistic term weighting, single term accuracy, genetic algorithms) were developed for weighting (22).

IV. Determination and Development of Model: Model can be obtained and developed by using classification, clustering etc. algorithms (3). In this study, one of classification algorithms, K-NN algorithm, was used.

K Nearest Neighbor (K-NN) Algorithm and Vector Space Model

The K-NN is a supervised learning algorithm that allows the query vector to be classified together with the K-NN vector. There are K training points closest to the query point in any query sample (23). Cosine similarity between other documents and query document is computed ($sim(d_i, q)$). Excess of *n* pieces of vectors whose similarity ratio is nearest to 1 is assigned to the document. d_i is the training document vector: $d_i = (wd_{i1}, wd_{i2}, \dots\dots\dots\dots, wd_{ij})$. w_{ij} is the weight of term in the document, q is the vector whose class will be determined.

$$sim(d_i, q) = cos\theta = \frac{d_i.q}{|d_i||q|} = \frac{\sum_j w_{i,j} w_{q,j}}{\sqrt{\sum_j w_{i,j}^2}\sqrt{\sum_j w_{q,j}^2}} \quad (1)$$

$sim(d_i, q) = 1 => d = q$, If $sim(d_i, q) = 0$ there is no term sharing.

The document whose class will be determined and all documents are showed vectorially in line with these rules. Each object here is defined as a vector. The axes of the vector space consist of different qualifications of these defined objects and each object is positioned in the vector space according to their qualifications (24). Three different methods are used to show a text in the vector space model (25):

Binary Vector: In this method, textual data is coded as 1 and 0 according to the presence or absence of words.

Frequency Vector: This method is an identification method considering how many times the word roots in the data are used.

TF-Term frequency - IDF-Inverse Document Frequency Vector: The frequency of the words in each document plays a role in the TF-IDF weighting. TF value shows the frequency information, that is, how many times the term occurs in the data set. The IDF gives a measure about the words that rarely occur in all documents (14). The equations (2) and (3) give TF and IDF calculations, respectively, and the equation (4) gives weight calculation.

$$TF_{ij} = \frac{n_{ij}}{|d_i|} \quad (2)$$

$$IDF_{ij} = \log\left(\frac{n}{n_j}\right) \quad (3)$$

$$W_d = TF_{ij} \times IDF_{ij} \quad (4)$$

The *n* value in the formula TF refers to how many times the j^{th} word root was used in the i^{th} data set. The *d* value refers to the number of all word roots within the data set. The *i* value in the formula is the number of words in the document. The *n* value in the formula IDF refers to the amount of the documents where the term *j* is contained whereas the total amount of documents is n_j. Weighting is made by multiplying these two values (17).

In the vector space model, documents and queries are represented by *m*-dimensional vectors. *m* is the number of terms in the dictionary. In this model, each document is represented by a numerical feature: *w*(*d*) = *(w*(*d,*),..., *w*(*d,*)). Each dimension of the vector includes the weight of the related term in the documents (17).

V. Evaluation of the Results: Before sharing the results, it is required to establish models and validate that all operations have been conducted properly.

VI. Presentation of the Results: It is the final step following the successful realization of the modeling process. The model results may be used several times for a better decision-making process (15).

2. Knime Software: Knime is an open-source coded data analysis platform based on the workflow logic, which ensures processing, interpreting, visualizing and reporting of the data by linking between the nodes under the node repository. This software can be obtained freely from the web site http://www.knime.com. The palladian toolkit should be installed on Knime for using text mining (26).

3. Application: In this stage of the study; the text mining method was applied to the documents obtained separately from the most frequently used Pubmed database under two different titles as “cancer cases in humans (human-and-cancer)” and “cancer researches in mice (mouse-and-cancer)”, and then to the combined documents, through the Knime program. Subsequently, the document classification was made using the K-NN algorithm. This study was prepared in accordance with the rules of research and publication ethics.

Firstly, after the workflow was created in the Knime program, textual document data were obtained from the Pubmed database via keywords using the document grabber node, and these data were entered into a blank folder in the computer and made ready for use. Keywords were written on the “query” section in the options window of the document grabber node as shown in Figure 1. When the “number of results” button was clicked, about three million results were obtained in the Pubmed database for human-and-cancer documents, and the results were limited to one thousand with the “maximal results” tab. Then, the folder, where the documents to be obtained by word tokenization method, document type and categorization would be saved, was created and the data were made ready for processing. The same processes were repeated for the mouse-and-cancer data. Then, these documents were transferred to the Knime workflow page using the document grabber node and made ready for pre-processing (Figure 2).

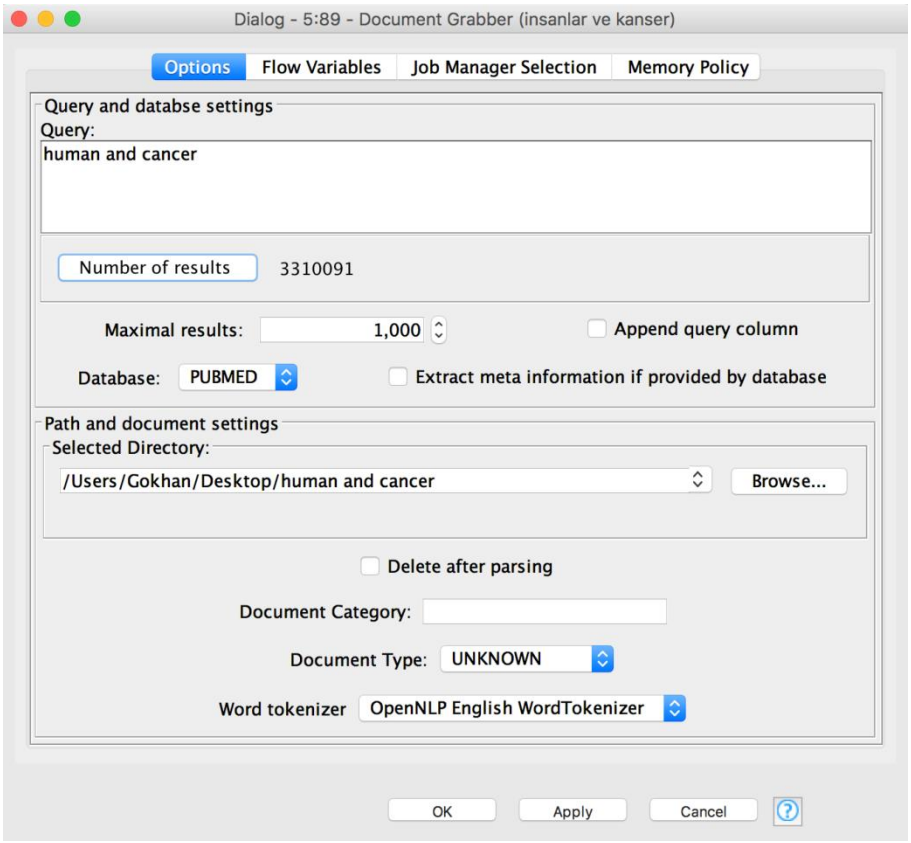


Figure 1. Document Grabber options window.

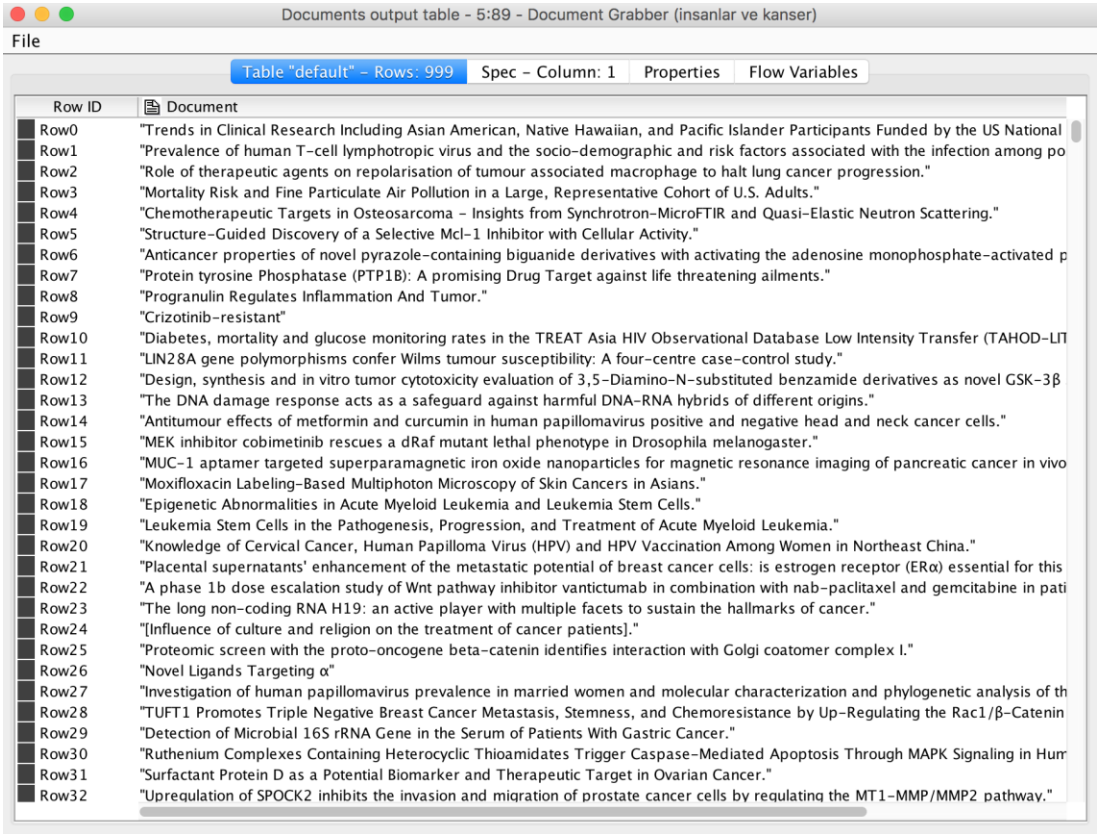


Figure 2. Document output table about human-and-cancer obtained from Pubmed database.

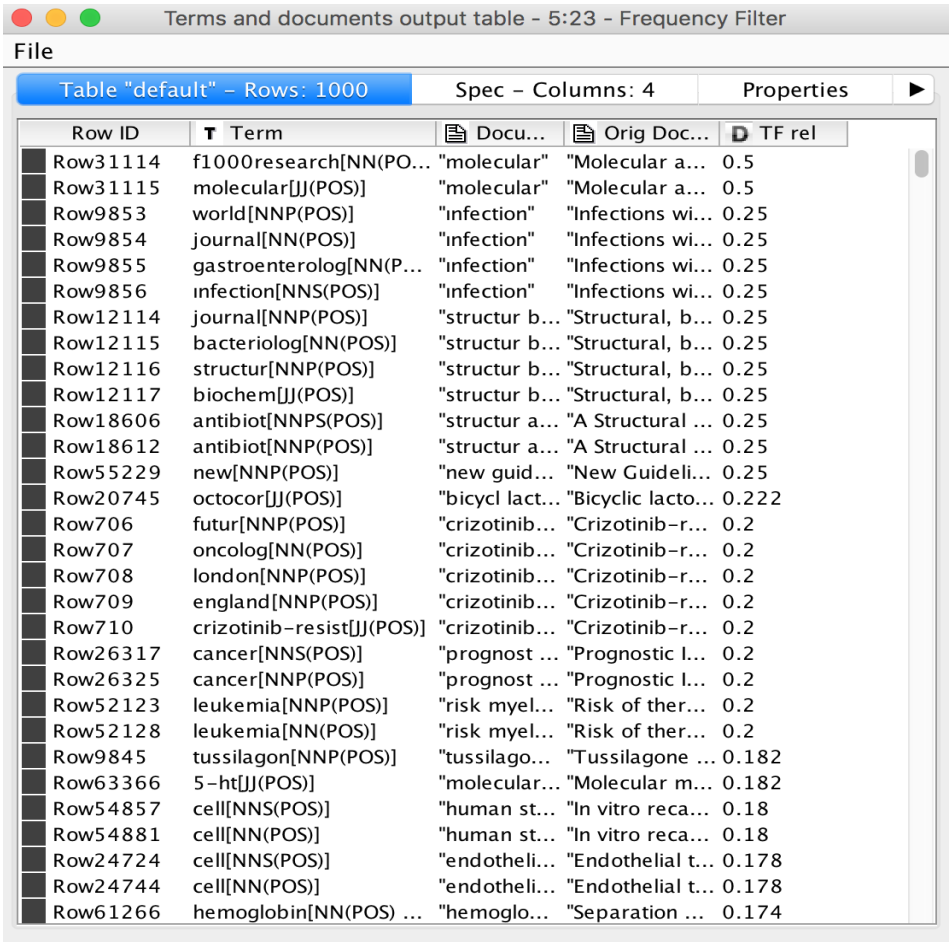
The document grabber node was transferred to the POS tagger and Abner tagger nodes for the named entity recognition phase with the arrows. The data were enriched with these nodes. The enrichment category contained nodes that assign a part of speech tags and recognize the standard named entities (for example, the names of persons, organizations, genes or proteins). The open NLP English word was selected as tokenizer option determining which tag would be assigned to which term. After the POS and Abner tagger nodes were linked to each other, the named entity recognition process was performed.

For pre-processing, punctuation erasure, N chars filter, number filter, case converter and snowball stemmer were applied. Then, the bag of words was created and the frequencies of all terms in the documents were calculated by the TF-term frequency node. The re-filtering process was performed based on these frequency values. The remaining terms were visualized by means of a tag cloud.

Finally, the K-NN algorithm was applied to the merged data set by using partitioning, K nearest neighbor and scorer nodes. In the partitioning node, the input table was divided into training and test data sets. The percentage of rows that would constitute the training data could be created here. The remaining rows would constitute the test data and were added to the K nearest neighbor node. Also, in this algorithm, stratified sampling was selected as the sampling method. For classification, the NN value was determined as K=3. The scorer node compares two columns with the attribute value pairs and shows the confusion matrix. The first column represents the actual class of the data. The second column represents the predicted data classes created by the K-NN algorithm. The scorer node output is the table of confusion matrix and accuracy statistics where each cell includes the number of matches. The confusion matrix where the predicted and actual values of the target quality are compared is commonly used to evaluate the performance of the classification models. The accuracy statistics table includes some statistical measures such as true positive (TP), true negative (TN), false positive (FP), false negative (FN), precision, recall/sensitivity, specificity, F-criteria, accuracy and Cohen's kappa.

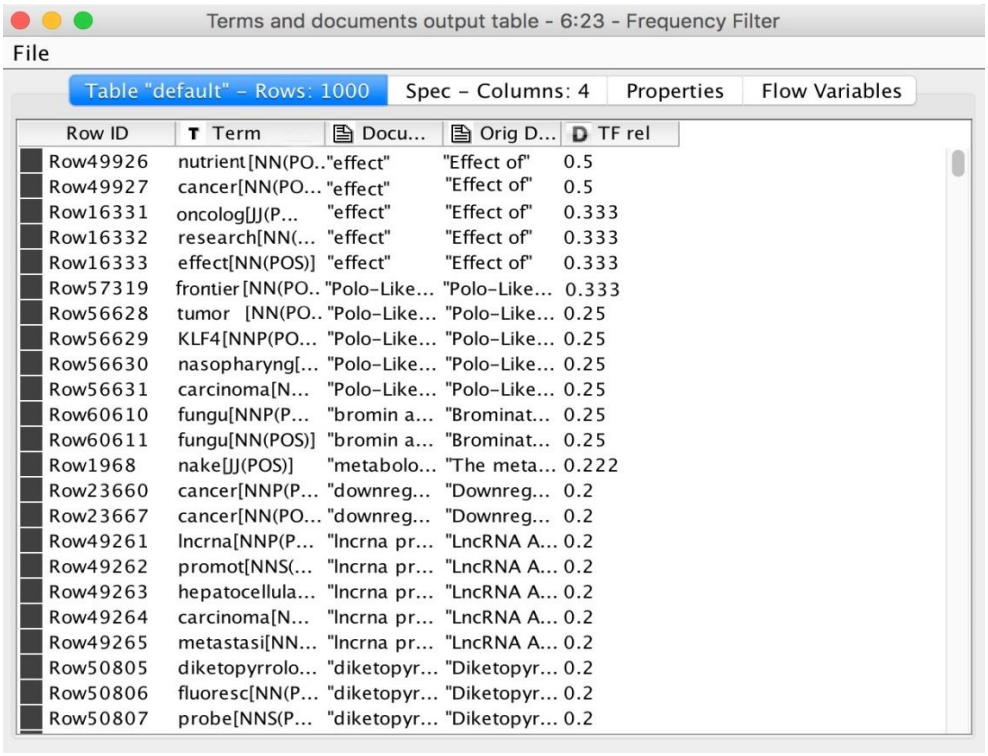
RESULTS

In this study, the documents obtained under two titles in the health field were divided into two categories. The first class consisted of textual documents related to human-and-cancer cases while the second class consisted of textual documents related to mouse-and-cancer studies. The TF values of first and second classes were given in Figures 3 and 4, respectively.



Row ID	Term	Docu...	Orig Doc...	TF rel
Row31114	f1000research[NN(PO...	"molecular"	"Molecular a...	0.5
Row31115	molecular[JJ(POS)]	"molecular"	"Molecular a...	0.5
Row9853	world[NNP(POS)]	"infection"	"Infections wi...	0.25
Row9854	journal[NN(POS)]	"infection"	"Infections wi...	0.25
Row9855	gastroenterolog[NN(P...	"infection"	"Infections wi...	0.25
Row9856	infection[NNS(POS)]	"infection"	"Infections wi...	0.25
Row12114	journal[NNP(POS)]	"structur b...	"Structural, b...	0.25
Row12115	bacteriolog[NN(POS)]	"structur b...	"Structural, b...	0.25
Row12116	structur[NNP(POS)]	"structur b...	"Structural, b...	0.25
Row12117	biochem[JJ(POS)]	"structur b...	"Structural, b...	0.25
Row18606	antibiot[NNPS(POS)]	"structur a...	"A Structural ...	0.25
Row18612	antibiot[NNP(POS)]	"structur a...	"A Structural ...	0.25
Row55229	new[NNP(POS)]	"new guid...	"New Guideli...	0.25
Row20745	octocor[JJ(POS)]	"bicycl lact...	"Bicyclic lacto...	0.222
Row706	futur[NNP(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row707	oncolog[NN(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row708	london[NNP(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row709	england[NNP(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row710	crizotinib-resist[JJ(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row26317	cancer[NNS(POS)]	"prognost ...	"Prognostic I...	0.2
Row26325	cancer[NNP(POS)]	"prognost ...	"Prognostic I...	0.2
Row52123	leukemia[NNP(POS)]	"risk myel...	"Risk of ther...	0.2
Row52128	leukemia[NN(POS)]	"risk myel...	"Risk of ther...	0.2
Row9845	tussilagon[NNP(POS)]	"tussilago...	"Tussilagone ...	0.182
Row63366	5-ht[JJ(POS)]	"molecular...	"Molecular m...	0.182
Row54857	cell[NNS(POS)]	"human st...	"In vitro reca...	0.18
Row54881	cell[NN(POS)]	"human st...	"In vitro reca...	0.18
Row24724	cell[NNS(POS)]	"endotheli...	"Endothelial t...	0.178
Row24744	cell[NN(POS)]	"endotheli...	"Endothelial t...	0.178
Row61266	hemoglobin[NN(POS) ...	"hemoglo...	"Separation ...	0.174

Figure 3. Term frequency (TF) values obtained by frequency filtering for human-and-cancer data.



Row ID	Term	Docu...	Orig D...	TF rel
Row49926	nutrient[NN(PO..	"effect"	"Effect of"	0.5
Row49927	cancer[NN(PO...	"effect"	"Effect of"	0.5
Row16331	oncolog[JJ(P...	"effect"	"Effect of"	0.333
Row16332	research[NN(...	"effect"	"Effect of"	0.333
Row16333	effect[NN(POS)]	"effect"	"Effect of"	0.333
Row57319	frontier[NN(PO..	"Polo-Like...	"Polo-Like...	0.333
Row56628	tumor [NN(PO..	"Polo-Like...	"Polo-Like...	0.25
Row56629	KLF4[NNP(PO...	"Polo-Like...	"Polo-Like...	0.25
Row56630	nasopharyng[...	"Polo-Like...	"Polo-Like...	0.25
Row56631	carcinoma[N...	"Polo-Like...	"Polo-Like...	0.25
Row60610	fungu[NNP(P...	"bromin a...	"Brominat...	0.25
Row60611	fungu[NN(POS)]	"bromin a...	"Brominat...	0.25
Row1968	nake[JJ(POS)]	"metabolo...	"The meta...	0.222
Row23660	cancer[NNP(P...	"downreg...	"Downreg...	0.2
Row23667	cancer[NN(PO...	"downreg...	"Downreg...	0.2
Row49261	lncrna[NNP(P...	"lncrna pr...	"LncRNA A...	0.2
Row49262	promot[NNS(...	"lncrna pr...	"LncRNA A...	0.2
Row49263	hepatocellula...	"lncrna pr...	"LncRNA A...	0.2
Row49264	carcinoma[N...	"lncrna pr...	"LncRNA A...	0.2
Row49265	metastasi[NN...	"lncrna pr...	"LncRNA A...	0.2
Row50805	diketopyrrolo...	"diketopyr...	"Diketopyr...	0.2
Row50806	fluoresc[NN(P...	"diketopyr...	"Diketopyr...	0.2
Row50807	probe[NNS(P...	"diketopyr...	"Diketopyr...	0.2

Figure 4. Term frequency (TF) values obtained by frequency filtering for mouse-and-cancer data.

The tag cloud graphic of the human-and-cancer class was given in Figure 5. The most commonly used tags in the human-and-cancer documents were written in larger font sizes, and the TF values were colored from yellow (the lowest) to red (the highest) using the color manager node. The term frequency values of the frequently used tags might be different.

The tags which were not used very frequently but had a high term frequency value might be small but dark colored while those which were used frequently but had a low term frequency value might be light colored. Therefore, the word which had the highest TF value and was most frequently used in the document was the term "cell". This was followed by the term "cancer". Then, the terms "haplotyp", "journal", "vaccin", "patient", "tumor", "molecular", "breast", "medicin" were weighted by frequencies according to their order of importance, and the most frequently used word groups in these documents were created.

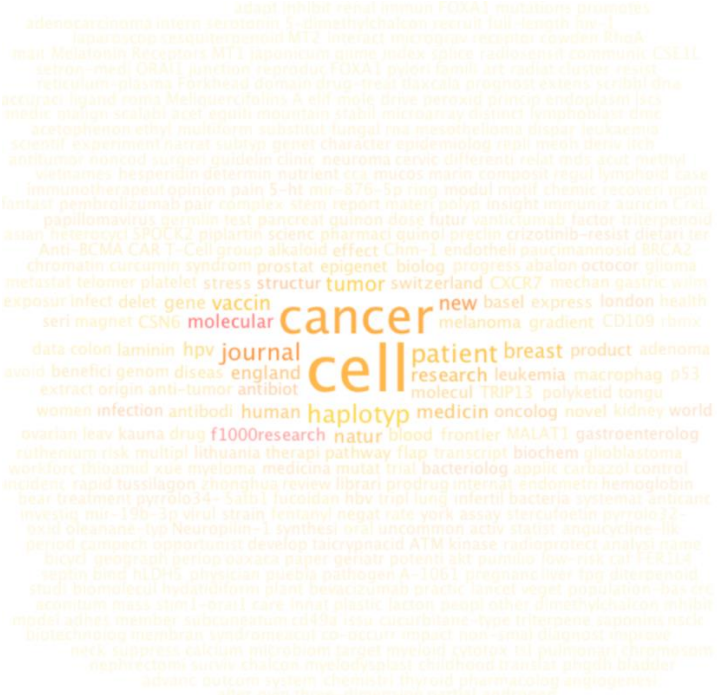


Figure 5. Tag cloud graphic for "human-and-cancer data" using tag filter.

Similarly, the tag cloud graphic of the mouse-and-cancer class was given in Figure 6. The prominent words in this graphic were “cell” and “cancer” words. They were followed by “nutrient”, “effect”, “tumor”, “carcinoma”, “liver”, “model”, “drug”, “patient”, “pancreat”, “oncolog”. In addition to similar words, different words came into prominence in the tag clouds obtained under two subject titles.

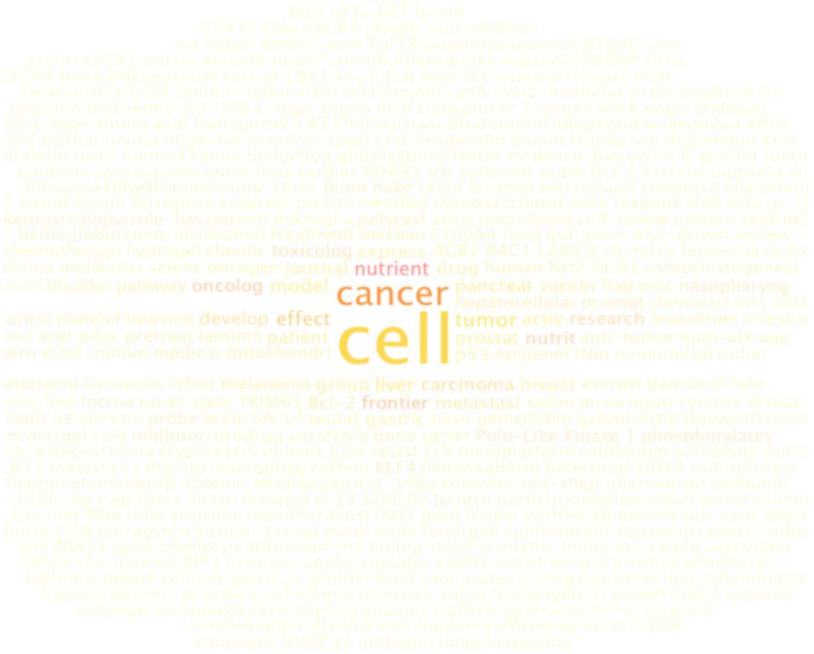


Figure 6. Tag cloud graphic for “mouse-and-cancer data” using tag filter.

Figures 7 and 8 showed the TF values and the tag cloud graph obtained without applying the tag filter after two textual data were merged. Conjunctions such as “and”, “the”, “with”, “were”, “that” used in English were not filtered. Since such conjunctions were used very frequently in the text, such meaningless words were considered as meaningful in the study documents and weighted by their term frequencies and shown with uppercases and dark colored in the tag cloud. However, it was observed that the words with high frequency value in both documents such as “cell”, “cancer”, “tumor” and “patient” had a high level of occurrence in the analysis performed after the data were merged.

Terms and documents output table - 6:30 - Frequency Filter

File

Table "default" - Rows: 1000 | Spec - Columns: 4

Row ID	Term	TF rel
Row24783	human [NNS(POS)]	... ... 0.333
Row47132	cancer[NNS(POS)]	... ... 0.333
Row47137	cancer[NNP(POS)]	... ... 0.333
Row124769	associ [JJ(POS)]	... ... 0.333
Row124770	research[NN(POS)]	... ... 0.333
Row124771	effect[NN(POS)]	... ... 0.333
Row175923	nutrit[NNP(POS)]	... ... 0.333
Row175924	and[CC(POS)]	... ... 0.333
Row175925	cancer[NN(POS)]	... ... 0.333
Row187120	patient [NN(POS)]	... ... 0.333
Row4713	the[DT(POS)]	... ... 0.25
Row55959	medicina[NNP(POS)]	... ... 0.25
Row55960	that[DT(POS)]	... ... 0.25
Row55961	lithuania[NNP(POS)]	... ... 0.25
Row55962	investig[NNP(POS)]	... ... 0.25
Row58904	cancer[NNS(POS)]	... ... 0.25
Row58911	cancer[NNP(POS)]	... ... 0.25
Row52229	cancer[NNP(POS)]	... ... 0.2
Row52238	cancer[NN(POS)]	... ... 0.2
Row68173	futur[NNP(POS)]	... ... 0.2
Row68174	oncologi[NN(POS)]	... ... 0.2
Row68175	london[NNP(POS)]	... ... 0.2
Row68176	england[NNP(POS)]	... ... 0.2
Row68177	crizotinib-resist[JJ(...	... ... 0.2
Row90900	the[DT(POS)]	... ... 0.2
Row10486	the[DT(POS)]	... ... 0.182
Row10488	fungu[NNP(POS)]	... ... 0.182
Row10489	fungu[NN(POS)]	... ... 0.182
Row37610	the[DT(POS)]	... ... 0.182
Row46614	with[IN(POS)]	... ... 0.182

Figure 7. Term frequency (TF) values obtained by frequency filtering without applying tag filter for both data.

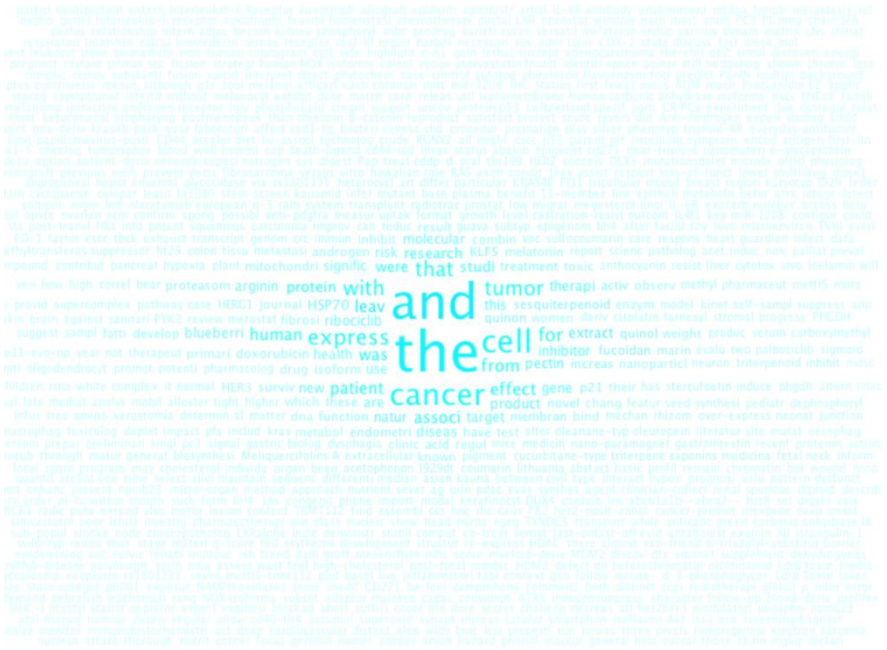
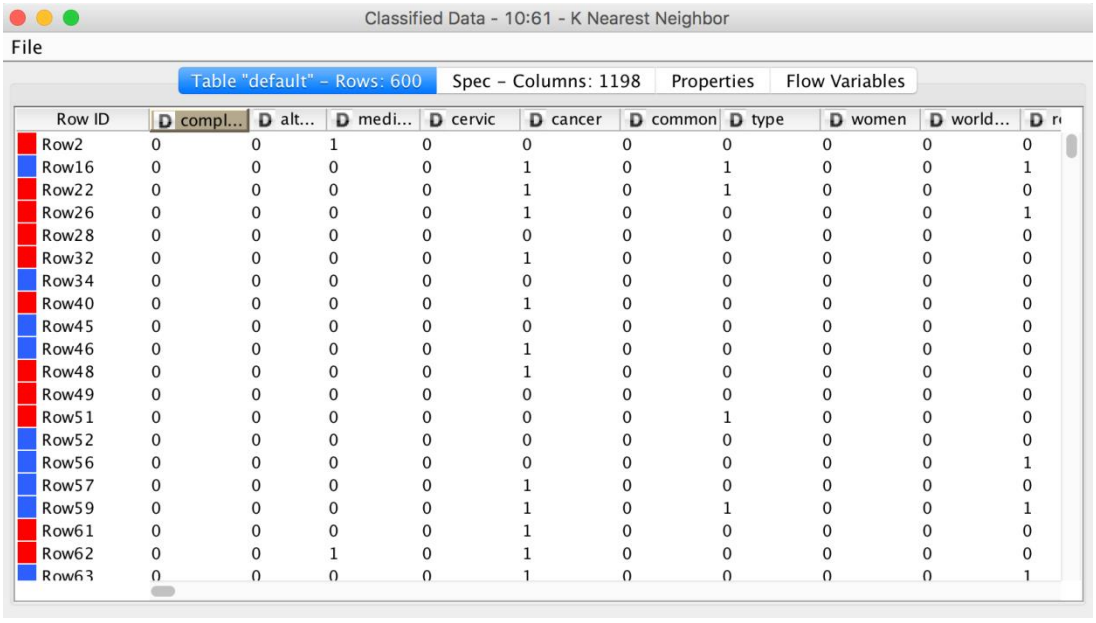


Figure 8. Tag cloud graphic (without applying tag filter).

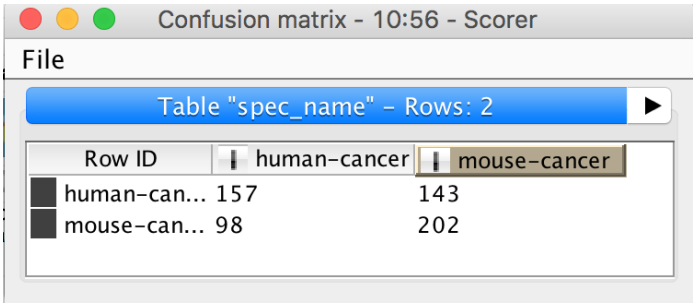
Finally, the classification data output obtained by the K-NN algorithm was given in Figure 9. This output showed the classified word vectors and document rows. The red and blue colored document rows represented the human-and-cancer category and the mouse-and-cancer category, respectively.



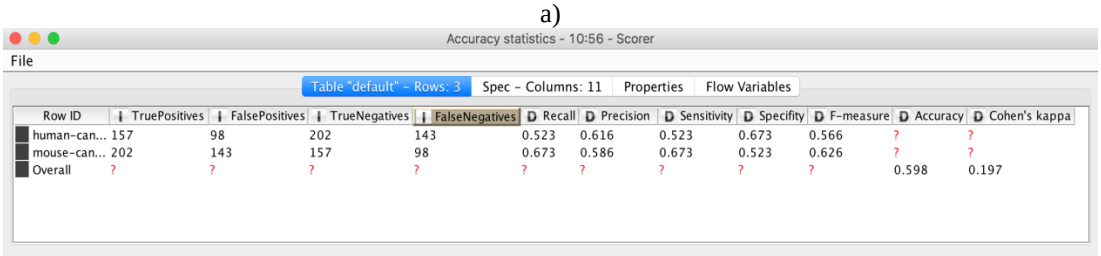
Row ID	compl...	alt...	medi...	cervic	cancer	common	type	women	world...
Row2	0	0	1	0	0	0	0	0	0
Row16	0	0	0	0	1	0	1	0	0	1
Row22	0	0	0	0	1	0	1	0	0	0
Row26	0	0	0	0	1	0	0	0	0	1
Row28	0	0	0	0	0	0	0	0	0	0
Row32	0	0	0	0	1	0	0	0	0	0
Row34	0	0	0	0	0	0	0	0	0	0
Row40	0	0	0	0	1	0	0	0	0	0
Row45	0	0	0	0	0	0	0	0	0	0
Row46	0	0	0	0	1	0	0	0	0	0
Row48	0	0	0	0	1	0	0	0	0	0
Row49	0	0	0	0	0	0	0	0	0	0
Row51	0	0	0	0	0	0	1	0	0	0
Row52	0	0	0	0	0	0	0	0	0	0
Row56	0	0	0	0	0	0	0	0	0	1
Row57	0	0	0	0	1	0	0	0	0	0
Row59	0	0	0	0	1	0	1	0	0	1
Row61	0	0	0	0	1	0	0	0	0	0
Row62	0	0	1	0	1	0	0	0	0	0
Row63	0	0	0	0	1	0	0	0	0	1

Figure 9. Classified data window obtained by K nearest neighbor (K-NN) algorithm.

The accuracy statistics table and the confusion matrix obtained by the K-NN algorithm were given in Figure 10. It was determined that 255 of a total of 600 test documents belonged to the human-and-cancer class and the remaining belonged to the mouse-and-cancer class; the recall/sensitivity, specificity and precision values were 52.3%, 67.3% and 61.6%, for the human-and-cancer documents, and 67.3%, 52.3% and 67.3% for the mouse-and-cancer documents, respectively; according to the F-criteria, the classification accuracy was 56.6% for the human-and-cancer documents and 62.6% for the mouse-and-cancer documents. The document classification estimation obtained with the K-NN algorithm was found to be 59.8% while Cohen's kappa value was 19.7%.



a)



b)

Figure 10. a) Confusion matrix and b) Accuracy statistics table view.

DISCUSSION

With the rapid development of technology and more integration of the internet into daily life, reaching accurate and reliable data in large-scale data sets in the fastest way has become the primary goal. Thus, text mining which digitalizes textual data and enables them to be applied for text mining algorithms has a very important place in today's world. Text mining methods have become a frequently used method recently in the fields of medicine and biology (4, 6-13) as well as in many fields. In a study conducted by Yu et al. (13), investigating the genes and prognostic factors associated with breast cancer considering 708 genes in total which they obtained from the gene expression omnibus database, they created the transcription factor-target regulation and microRNA-target gene network using the text mining approach. Thompson et al. (4), in their study titled "Text Mining the History of Medicine", utilized the ability of text mining methods to recognize various types of semantic information automatically (places, medical conditions, drugs, etc.), synonyms/variant forms of concepts, and relationships holding between concepts (which drugs are used to treat which medical conditions, etc.). Lam et al. (8) used the text mining method to determine the publication trends in journal articles related to sleep disorders published between 2000-2013 and to explore the relationship between sleep disorders and methodological terms. Hoa and Zhang (7) investigated whether the written evaluations of Chinese patients about the doctors, who treated them, on the web and whether the positive and negative comments showed difference depending on the area of specialization of the doctors using the text mining method to provide health services more effectively. Mahgoub et al. (6) converted 100 internet page samples about avian influenza obtained from several sources (BBC, Reuters, yahoo, medical news today, etc.) into extensible markup language (XML) format and investigated the relationship between the keywords and tried to reveal the disease-related characteristics (location, patient's condition, etc.) using a text mining system called extracting association rules from text (EART).

In this study, the text mining method was applied to the documents obtained from Pubmed database under two different titles as "human-and-cancer" and "mouse-and-cancer". In the tag clouds, the most frequently used labels in both "human-and-cancer" and "mouse-and-cancer" documents were written in larger letters and visualized. In both documents, the words "cell" and "cancer" had the highest TF value and were most frequently used. In the mouse-and-cancer studies, in addition to the words "cell" and "cancer", it was seen that the words such as "nutrient", "effect", "tumor", "carcinoma", "liver", "model", "drug", "patient", "pancreat" and "oncolog" are prominent respectively. Similarly, in the tag cloud graphic of human-and-cancer class, in addition to the cell and cancer tags, the word "haplotyp" which is a gene term as well as words such as "journal", "vaccin", "patient", "tumor", "molecular", "breast", "medicin" were found to be the most commonly used word groups in these documents. The most frequently used word in the documents obtained from Pubmed was "cell". The frequency of the words used in the texts about "human-and-cancer" and "mouse-and-cancer" reveals that these two cases have a very strong relationship with human and animal cells and hence the term frequency value is also high. In both categories, the word "tumor" stands out to be effective. Its term frequency is relatively lower than the word "cell". When the same study was conducted without using the tag filter node, conjunctions such as "and",

"the", "with", "were", "that" were not filtered since they were used very often in the text. It is of great importance to filter such type of noisy data in the text mining studies since other terms that need to be significant remain in the background. Otherwise, the study will extend over a longer period time and the efficiency of the study will decrease, the cost will increase due to the use of redundant data and it will be more difficult to achieve the desired outcome. In the results of the K-NN algorithm, 255 of 600 test documents were classified to be in human-and-cancer class while 157 were found to be estimated correctly and precision was calculated as 61.6%. Similarly, of 345 test documents belonging to the mouse-and-cancer class, 202 belonged to the mouse-and-cancer class and the precision was found to be 58.6%. In addition, the recall/sensitivity indicating the ratio of the number of documents in both classes that were accurately predicted by the algorithm to the actual amount of test data was 52.3% for the human-and-cancer class and 67.3% for the mouse-and-cancer class. The high rate in the classification of the mouse-and-cancer documents is marked here. According to the F-criteria, accurate classification percentages of the human-and-cancer and mouse-and-cancer documents were obtained as 56.6% and 62.6%, respectively. Partially successful document classification estimation was found with a percentage of 59.8% with the K-NN algorithm. However, Cohen's kappa value, which shows the probability of total random fit between the actual and the classification results, was found to be 19.7%, and the fit was of slight level according to Cohen's kappa classification (27).

CONCLUSION

The text mining method has many advantages such as having the ability to analyze both structural and non-structural data and providing fast, reliable and accurate information in big data sets. In this sense, it is recommended to use the text mining method and to generalize its use to obtain information quickly and reliably in the health field where there are numerous digital and printed documents.

REFERENCES

1. Cerrito P. Inside text mining. Text mining provides a powerful diagnosis of hospital quality rankings. Health Manag Technol. 2004; 25(3): 28-31.
2. Visa A. Technology of text mining. In: Perner P, editor. Machine learning and data mining in pattern recognition. MLDM 2001. Lecture Notes in Computer Science, vol 2123. Berlin, Heidelberg: Springer; 2001. p.1-11.
3. Sehgal AK. Text mining: the search for novelty in text [PhD dissertation]. Iowa: The University of Iowa, Department of Computer Science; 2004.
4. Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, et al. Text mining the history of medicine. PLoS ONE. 2016; 11(1): e0144717. https://doi.org/10.1371/journal.pone.0144717.
5. Losiewicz P, Oard DW, Kostoff RN. Textual data mining to support science and technology management. J Intell Inf Syst. 2000; 15(2): 99-119.
6. Mahgoub H, Rösner D, Ismail N, Torkey F. A text mining technique using association rules extraction. Int J Comput Intell. 2007; 4(1): 21-8.
7. Hao H, Zhang K. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. J Med Internet Res. 2016; 18(5): e108. doi: 10.2196/jmir.4430.
8. Lam C, Lai FC, Wang CH, Lai MH, Hsu N, Chung MH. Text mining of journal articles for sleep disorder terminologies. Plos One. 2016; 11(5): e0156031. doi: 10.1371/journal.pone.0156031.
9. Hsiao YW, Lu TP. Text-mining in cancer research may help identify effective treatments. Transl Lung Cancer Res. 2019; 8(Suppl 4): S460-3. doi: 10.21037/ tlcr.2019.12.20.
10. Jahanbin K, Rahmanian F, Rahmanian V, Jahromi AS. Application of twitter and web news mining in infectious disease surveillance systems and prospects for public health. GMS Hyg Infect Control. 2019; 14: Doc19. doi: 10.3205/dgkh000334. eCollection 2019.
11. Lebowitz A, Kotani K, Matsuyama Y, Matsumura M. Using text mining to analyze reflective essays from Japanese medical students after rural community placement. BMC Med Educ. 2020; 20(1): 38. doi: 10.1186/s12909-020-1951-x.
12. Sahin K, Durdagi S. Identifying new piperazine-based PARP1 inhibitors using text mining and integrated molecular modeling approaches. J Biomol Struct Dyn. 2020; 1-10. doi: 10.1080/07391102.2020.1715262.
13. Yu Z, He Q, Xu G. Screening of prognostic factors in early-onset breast cancer. Technol Cancer Res Treat. 2020; 19: 1533033819893670. doi: 10.1177/1533033819893670.
14. Soucy P, Mineau W. Beyond TFIDF weighting for text categorization in the vector space model. Proceedings of the 19th International Joint Conference on Artificial Intelligence; July 30-August 2005; Edinburgh-Scotland. San Francisco, CA: Morgan Kaufmann Publishers Inc; 2005. p. 1130-5.
15. Miner G, Delen D, Elder J, Fast A, Hill T, Nisbet RA. Practical text mining and statistical analysis for non-structured text data applications. San Francisco, USA: Academic Press; 2012.
16. Kaşıkçı T, Gökçen H. Metin madenciliği ile e-ticaret sitelerinin belirlenmesi. BTD. 2014; 7(1): 25-32.

17. Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. LDV Forum-GLDV Journal for Computational Linguistics and Language Technology. 2005; 20(1): 19-62.
18. Akın AA, Akın MD. zemberek.googlecode.com [Internet]. Zemberek an open source NLP framework for Turkic languages [cited 2019 March]. Available from: http://zemberek.googlecode.com/.
19. tools.nlp.itu.edu.tr [Internet]. ITU Natural Language Processing Research Group [cited 2019 March]. Available from: http://tools.nlp.itu.edu.tr/.
20. kemik.yildiz.edu.tr [Internet]. YTU Kemik Natural Language Processing Group [cited 2019 March]. Available from: www.kemik.yildiz.edu.tr.
21. İlhan U. Application of KNN and FPTC Based text categorization algorithms to Turkish news reports [master's thesis]. Ankara: Bilkent University, Institute of Engineering and Science; 2001.
22. Pilavcılar İF. Metin madenciliği ile metin sınıflandırma [yüksek lisans tezi]. İstanbul: Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü; 2007.
23. Kutlu F. Categorization in a hierarchically structured text database [master's thesis]. Ankara: Bilkent University, Institute of Engineering and Science; 2001.
24. İlhan S, Duru N, Karagöz Ş, Sağır M. Metin madenciliği ile soru cevaplama sistemi. Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO) 2008; 26-30 Kasım 2008; Bursa. s. 356-9.
25. Çalış K, Gazdağı O, Yıldız O. Reklam içerikli epostaların metin madenciliği yöntemleri ile otomatik tespiti. BTD. 2013; 6(1): 1-7.
26. Knime.com [Internet]. About Knime home [cited 2019 March 22]. Available from: https://www.knime.com/about.
27. Warrens MJ. Five ways to look at Cohen's kappa. Psychol Psychother. 2015, 5(4): 1-4. doi: 10.4172/2161-0487.1000197.