Applied Information Technology And Computer Science Vol. 2 No. 2 (2021) 1238-1254

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

Comparison of Treatment Insomnia Disease Between Prescription Medication and Cognitive Behavioral Therapy Using Rstudio

Siti Aqilah Ibrahim, Zuraidin Mohd Safar*

Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, MALAYSIA

DOI: https://doi.org/10.30880/aitcs.2021.02.02.077
Received 02 August 2021; Accepted 14 November 2021; Available online 30 November 2021

Abstract: Text mining is a process of transforming unstructured text into a structured format or procedure to accumulate data from the knowledge that we got. Text mining can combine information from numerous sources. This project focuses to extract the knowledge to seek out the amounts of occurrences for every term to work out the treatment of insomnia between prescription medication and cognitive behavioral therapy. There are two techniques that have been used which are information extraction and classification. Information extraction is that the process of extracting specific (pre-specified) information from textual sources which is data from prescription medication and cognitive behavioral therapy. There are five algorithms that has been used which are Naive Bayes, PART, IBk, SGD and Random Forest. These five algorithms are used to find the F-measure score. F-measure score determine the diseases whether can treat which prescription medication and cognitive behavioral therapy. Then, F-measure result of every algorithm were compared to understand the acceptable treatment for insomnia disease.

Keywords: Text Mining, Extraction Information, Classification, Insomnia Disease

1. Introduction

The issue in content mining procedures is to remove the use- required data. Concealed data from unstructured to semi-organized information is separated from text mining. This likewise extricates data naturally from different composed instruments and by machine for removing new, already obscure information. In treatment of insomnia disorder illnesses, we had prefer to survey and understand the better treatment of insomnia whether by physician recommended medicine or psychological social treatment [1]. It's hard to work out the least difficult treatment to treat the infections more white utilize the physician recommended drug treatment or intellectual social treatment which is change your day by day exercises.

Insomnia is a type of sleep disorder that can make people hard and trouble falling asleep. It also can cause people to wake up too early and not able to get back to sleep [2]. It is can cause people not getting enough sleep. When people not getting enough sleep, people can feel tired when they wake up. Insomnia can affect people health, mood, energy level and quality of life.

*Corresponding author: zuraidin@uthm.edu.my

Insomnia can be detected when people have symptoms of disease. The symptoms of insomnia is when someone difficulty falling asleep at night, waking up during the night , waking up too early, daytime they will feeling tired or fatigued, problem with concentration or memory, irritability or depressed mood, mood easy to change, difficulty paying attention, focusing on task that they have to do, hard to remembering the task people give, increased error or accident, when they want to sleep they will feel worries, using medication or alcohol to fall asleep and etch [3].

There are two type of Insomnia which are primary insomnia and secondary insomnia. Primary insomnia is when people having sleeping problem but their problem is not link with other health condition or problem, it issue by themselves. Primary insomnia is causes by stress [4]. When people have a lot of thing that they have to think, they will feel stress and pressure it will affect the time for their sleep. The second thing that cause primary insomnia is travel or work schedule. Human body rhythms act as internal clock guiding such as sleep-wake cycle, metabolism and etc. Human body rhythms can lead to insomnia because traveling across multiple time zones, working a late or early shift and etch [5]. Secondary insomnia is when people having trouble to sleep because of heath condition such as depression and so on. Secondary insomnia causes by mental health disorders. Anxiety disorders like post-traumatic stress disorder may disrupt time to sleep to people. Awakening too early can be sign of depression. Secondary insomnia related to other mental health disorders as well. Insomnia can be treated by prescription medication or cognitive behavioral therapy.

In this research, we used 2 technique which is extraction information and classification. In extraction information we used RStudio. First, we need to collect data from the PubMed. The data that we collect is 200, 100 for medication, and 100 for therapy. After that, the data will save in folder and insert into Rstudio. Rstudio will produce one data in cvs format. After done extraction information process we will move to process classification. In classification, we will use the WEKA tool. In this process, we will separate 2 classes, one for medication and another one for therapy. In WEKA we used 5 algorithms to compare the F-measure score.

At the end of this research, we will get the result of the 5 types of algorithm that we used in WEKA. In algorithms, it will be some calculation of precision and recall to produce an F-measure score. It's critical to comprehend what calculation will be use to ask the outcomes. The calculation that was pick should give the yield and doesn't have a misstep. The result that we get will be to determine whether prescription medication or cognitive behavioral therapy is a more suitable treatment for insomnia disease.

2. Related Work

2.1 Introduction

In this chapter, there consist text mining, pre-processing process, diseases treatment, machine learning, and data set that explained further during this literature review. For text mining, the technique used which is information extraction and classification been explained. It also explained in detail the pre-processing process. For the category treatment, there are prescription medications or cognitive behavioral therapy. The machine learning used for this research is Support Vector Machine (SVM) also will explain in this chapter.

2.2 Text Mining

Text mining alludes to the route toward isolating captivating and non-insignificant plans or information from the record [1]. Text mining separated valuable information from the text information. It encourages the clients to search out the information or information in content reports.

The initial step of the text mining measure starts when the archive was gathered from numerous assets. The apparatus of text mining would recuperate a chose archive and pre-measure it by checking arrangement and character sets. The record that has been finding or recover goes to the message examination stage. At that point, text investigation gets the standard data from the report. In some cases, until the standard information is removed content examination has been rehashed over and over. There are two techniques that use which are information extraction and categorization or classification.

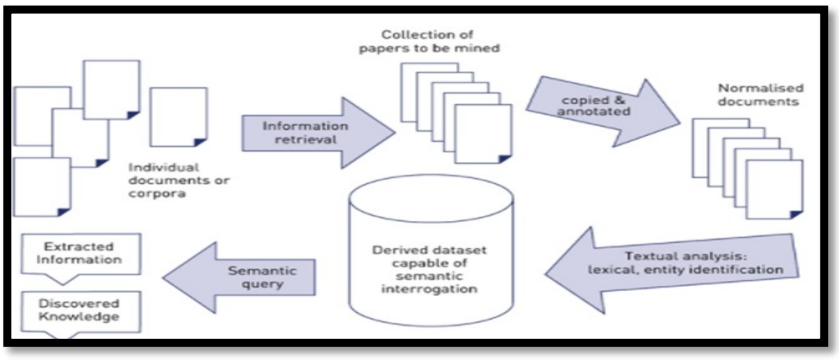

Figure 1: Jisc' model text mining process

2.2.1 Information Extraction (IE)

The capacity of data extraction is to recognize the critical expressions by break down unstructured content in the report. To discover the predefine game plan in content, design coordinating cycle is used. The keywords are parsed and semantically deciphered then need a little of information to enter into the database. The most precise information extraction frameworks include handcraft language processing module has been made in applying data mining technique. This development can be incredibly significant while overseeing colossal volumes of substance or text. Data extraction oversee issue of literary record become corpus to become organized information base. KDD module give a data extraction module to be built [18].

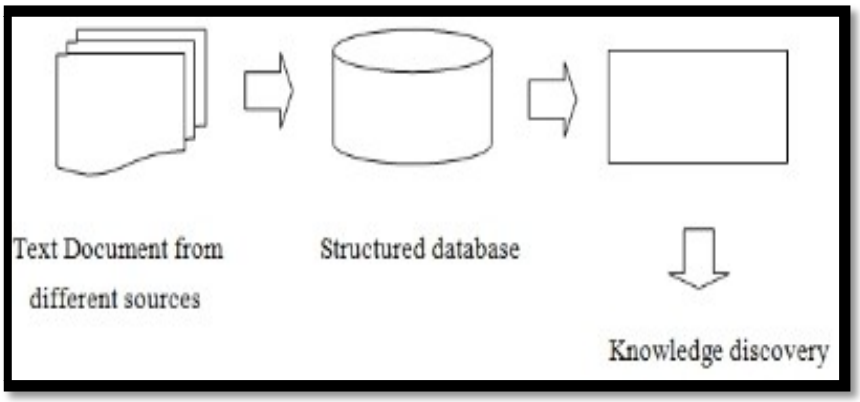

Figure 2: Information Extraction (Sonali, 2014)

2.2.2 Keyword Extraction

Keyword extraction is a text analysis technique that automatically extract the most word and expression that used in text. Keyword extraction help summarize the content of article and recognize the main topic. Research focused on methodologies to normally extricate the watchwords from articles or records as a guide either to suggest catchphrases for a specialist indexer or to deliver rundown for chronicles that would somehow be hard to react. Despite the fact that, the term utilized is extraordinary however it is representing to the most important data that depict in the document. The strategy for watchword extraction that has been used will recognize and clarify all of the information required.

2.2.3 Pre-processing process

Pre-processing process is one of the critical fragments in various content mining calculation. For example, a book arrangement structure contains pre-handling incorporate decision and characterization steps [1] have inspected the impact of pre-handling endeavors particularly in the content zone. It essentially comprises of the technique tokenization and filtering.

2.2.4 Tokenization

Tokenization is the process of separating a piece of data into random string pieces that called tokens, and maybe simultaneously discard certain characters, for example, complement marks. To further processing the list of tokens is used [11].

2.2.5 Filtering

Filtering is generally done on reports to delete some of the words. A regular of sifting is stop-words expulsion. Stop words are the words frequently appear in the substance without having a ton of substance information. Closeness words happening oftentimes in the substance said to have little information to perceive different records. More over words happening now and again are also possibly of no basic essentialness and can be ousted from the records [10].

2.2.6 Classification of Treatments

After the data extraction finish, a grouping method is utilized to approve the outcome. Classification is a process of finding the model that describe data classes and concepts. Text arrangement has been exhaustively thought about in different organizations, for instance, information mining, data set, AI and data recovery, and used in tremendous number of employments in various spaces, for instance, determination of clinical, record affiliation, and others. Allocate predefined classes to message archives is the point of text characterization. Critically, the directed characterization is the base of archive order algorithm.

2.2.7 Support Vector Machine (SVM)

SVM are managed learning classification algorithm where have been comprehensively used in content grouping issues. SVM are a kind of Linear Classifiers. Straight classifiers concerning content files are models that creation a gathering decision relies upon the assessment of the immediate mixes of the report highlights. The SVM from the outset introduced in [3]. SVM attempts to find a “decent” straight separators between various classes. A single SVM can simply isolate two classes, a positive class and a negative class. SVM computation attempts to find a hyperplane with the most outrageous detachment from the positive and negative models. The reports with great ways from the hyperplane are called support vectors and decide the certifiable zone of the hyperplane. If the record vectors of the two classes are not straight recognizable, a hyperplane is settled with the ultimate objective that insignificant number of report vectors are arranged in an unseemly side.

The equation of linear predictor [15]:

$$y = \vec{a}.\vec{x} + b \quad \text{, where} \quad \vec{x} = x_1, x_2, \dots, x_n$$

$$\vec{a} = (a_1, a_2, \dots, a_n)$$

Given:
y = Normalized document word frequency vector

$\vec{a}$ = Vector of coefficient

b = scalar

2.2.8 Algorithm Used In research

Naive Bayes, the name of this algorithm is weka.classifiers.bayes.NaiveBayes. It is class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Update able Classifier (which in typical usage are initialized with zero training instances) if you need the Updateable Classifier functionality, use the Naïve Bayes Updateable classifier. The Naïve Bayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances [12].

Secondly is PART, the name of this algorithm is weka.classifiers.rules.PART. It is class for generating a PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule [8]. IBk, The name of this algorithm is weka.classifiers.lazy.IBk. Is otherwise called as K-nearest neighbor classifier. Can choose suitable estimation of K based on cross-approval. Can likewise remove weighting [1].

SGG, SGD uses stochastic gradient descent to learn linear models (linear SVM, logistic regression and multiple linear regression). Stochastic gradient descent is an incremental anytime algorithm, so it can be applied to data streams or data sets that are too large to fit into main memory [8]. Lastly is Random Forest, The name of this algorithm is weka.classifiers.trees.RandomForest. It is class for constructing a forest of random trees [12].

2.3 Reason selection algorithm

Based on the algorithm that I choose which is Naive Bayes, IBK, PART, SGD and Random Forest. Before I choose this algorithms, I make a research about algorithm that I want use. It's because there are many types of algorithm in WEKA. I make a decision to choose this algorithms because this 5 algorithm give the result that needed in this research. It's give the f measure score different value so we can compare it which is more suitable for treatment of insomnia whether therapy or medical medicine. Other's algorithm have an error to produce the F-measure score. Some of algorithm F-measure score doesn't come out, some of algorithm have 0 F-measure score and other's error. So I choose this 5 algorithm because I can compare the value of F-measure score that hit with the objective of the research which is to identify which is more suitable for treatment. Without F-measure score value the research cannot be completed.

2.4 Diseases Treatment

Insomnia is a disease that people has a common sleep disorder that can make people hard to fall asleep, hard to stay asleep and make people wake up too early and not able to get back to sleep. Insomnia disease have 2 type of insomnia which is primary insomnia and secondary insomnia. Insomnia disease it causes by stress, environment, and change sleep schedule, mental health such as depression, medication for another sickness and etch [6]. There have two ways to treat this disease which is prescription medication and cognitive behavioral therapy.

Prescription medication is taking a sleeping pill to get sleep, stay asleep and not to wake up to early. Usually, doctor do not recommended people relying on prescription sleeping pill. When people start taking sleeping pills, they will rely on sleeping pill for more than a week. They will take a sleeping pill more than a week to help then to get to sleep. Prescription sleeping pill will affect human body when they take the pill too often [7]. The prescription medication sleeping pill that approved for long-term used is eszopiclone (Lunesta), ramelteon (Rozerem), zaleplon (Sonata) and zolpidem (Ambien, Edluar, Intermezzo, Zolpimist). Prescription sleeping pills can affect human body such as can causing daytime grogginess and increasing the risk of falling. Nonprescription sleep medication contains antihistamines that can make people drowsy, daytime sleepiness, dizziness, confusion, cognitive decline and difficulty urinating.

Cognitive behavioral therapy is a therapy or strategies for help people to gel sleep. Cognitive behavioral therapy also knows as CBT-I. CBT-I can help people to control or eliminate negative thought and action that can keep they awake. Usually, doctor will recommend this way to treatment people that have insomnia disease. Generally, CBT-I is more effective than sleep medication but some peoples cannot accept this way and they more prefer medication pill. CBT-I will teach people to recognize and changes beliefs that affect people ability to sleep. This way can help people to control or eliminate negative thought and worries that keep they awake. The therapy or strategies is relaxation techniques. Progressive muscle relaxation, training program and respiration exercises area unit ways in which to cut back anxiety at time of day. Active these techniques will assist management your respiration, heart rate, muscle tension and mood in order that will help they relax. Second strategies is stimulus control therapy. Numerous individuals with sleep deprivation encounter uneasiness at the unimportant prospect of falling snoozing, which can compound and draw out their indications. Jolt control includes a arrangement of steps you'll be able take to decrease these anxieties and develop a positive relationship along with your rest region. These incorporate lying down as it were once you feel tired, employing a bed as it were for rest and sex, and setting an alert for the same time each morning. CBT-I specialists regularly empower sleepers to induce up in the event that they are incapable to drop snoozing after 10 minutes of lying on bed, and to as it were return to bed when they feel tired. Boost control moreover debilitates daytime resting.

2.4.1 Prescription medication

Prescription medication is taking a sleeping pill to get sleep, stay asleep and not to wake up to early. Usually, doctor do not recommended people relying on prescription sleeping pill. When people start taking sleeping pills, they will rely on sleeping pill for more than a week. They will take a sleeping pill more than a week to help then to get to sleep. Prescription sleeping pill will affect human body when they take the pill too often. The prescription medication sleeping pill that approved for long-term used is eszopiclone (Lunesta), ramelteon (Rozerem), zaleplon (Sonata) and zolpidem (Ambien, Edluar, Intermezzo, Zolpimist). Prescription sleeping pills can affect human body such as can causing daytime grogginess and increasing the risk of falling. Nonprescription sleep medication contains antihistamines that can make people drowsy, daytime sleepiness, dizziness, confusion, cognitive decline and difficulty urinating.

2.4.2 Cognitive behavioral therapy

Cognitive behavioral therapy is a therapy or strategies for help people to get sleep. Cognitive behavioral therapy also knows as CBT-I. CBT-I can help people to control or eliminate negative thought and action that can keep them awake. Usually, doctor will recommend this way to treatment people that have insomnia disease. Generally, CBT-I is more effective than sleep medication but some peoples cannot accept this way and they more prefer medication pill. CBT-I will teach people to recognize and changes beliefs that affect people ability to sleep. This way can help people to control or eliminate negative thought and worries that keep they awake. The therapy or strategies is relaxation techniques. Progressive muscle relaxation, training program and respiration exercises area unit ways in which to cut back anxiety at time of day. Active these techniques will assist management your respiration, heart rate, muscle tension and mood in order that will help they relax. Second strategies are stimulus control therapy. Numerous individuals with sleep deprivation encounter uneasiness at the unimportant prospect of falling snoozing, which can compound and draw out their indications. Jolt control includes an arrangement of steps you'll be able take to decrease these anxieties and develop a positive relationship along with your rest region. These incorporate lying down as it were once you feel tired, employing a bed as it were for rest and sex, and setting an alert for the same time each morning. CBT-I specialists regularly empower sleepers to induce up in the event that they are incapable to drop snoozing after 10 minutes of lying on bed, and to as it were return to bed when they feel tired. Boost control moreover

debilitates daytime resting. There are many other strategies or therapy that can help people to get sleep. People have to go hospital or clinic when they have symptom of insomnia disease.

3. Methodology/Framework

3.1 Overview

Methodology plays an important role in implementing this research study accordingly. In this part, all the exercises and work process that are completed in this examination are depicted. The structure is intended to give the general progression of examination including computational system.

3.2 Research Framework

In this research, the stream will rely upon the structure. In the system, it comprises five stages. The subtleties of each stage been clarified totally in this part. The Figure 3 beneath indicated the work process cycle of this research.

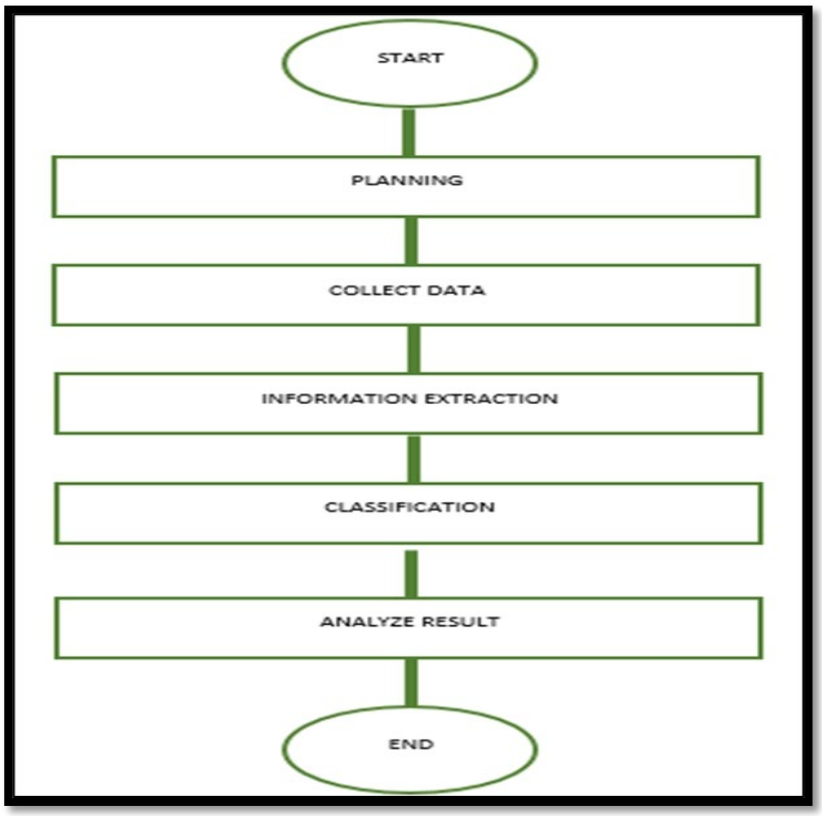

Figure 3: Research Framework

3.2.1 Planning

Firstly, before the research started, we need to create a plan to avoid any necessary errors occur. So, this phase can help us to give explanation about the data that has been collected about treatment of insomnia. This information is derived from several different resources. At the same time, the problem can be distinguished by observing the prior experiments that were featured in such research papers and articles.

3.2.2 Collect Data

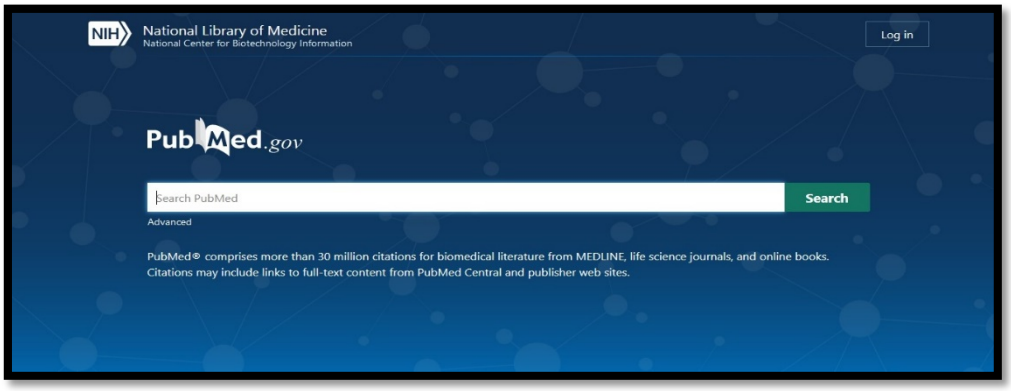

Figure 4: Data Source

To identify medication terms that may have been investigated in insomnia treatment, literature searches were performed. An article that based on original research relating to classical medical terminology and the esteemed doctor who has expertise in it was gathered from various sources. The problem that should be solved is necessary for researchers to consider. In this way, in this process, various information and essential data should be arranged.

3.2.3 Information Extraction

Text mining techniques are used to collect data and information for the information extraction process. The RStudio is used to retrieve the dataset. Using the package for text mining in RStudio. The methods used in RStudio text mining package are tm map. Figure 3 show the text mining that use in RStudio. This process change the Data into .arff format. The data will be stored in csv format after collecting similar data using text mining software using RStudio for information extraction. Then change the csv file to .arff format to be entered in the Support Vector Machine tool, as only .arff format can be used in this tool.

3.2.4 Classification

For classification, the Support Vector Machine approach is used. In SVM, the method used is the Waikato Environment for Knowledge Analysis (WEKA). The diseases have been divided into two forms of treatment by using the Support Vector Machine (SVM). Medicinal medications are the first aid, and therapy are the second. Apply WEKA, after classify the related data using Waikato Environment for Knowledge Analysis, the performance of text mining technique is compared by using F-measure score. There are three algorithm that use in Waikato Environment for Knowledge Analysis to compare the results.

3.2.5 Analyze the Result

In this phase, the result that were produce is been compared. From different algorithms, the result of efficiency, speed or accuracy been proved. The algorithms that we used is Naive Bayes, PART, IBK, SGD and Random Forest. It is because this 5 algorithm have F-measure score that accurate that we need. Some algorithms do not have F-measure score and some algorithm we get F-measure same value, that cannot used in our result. Because we cannot compare the value of F-measure if the value is same. We cannot get the accurate result if the value of F-measure same.

3.3 Performance Measurement

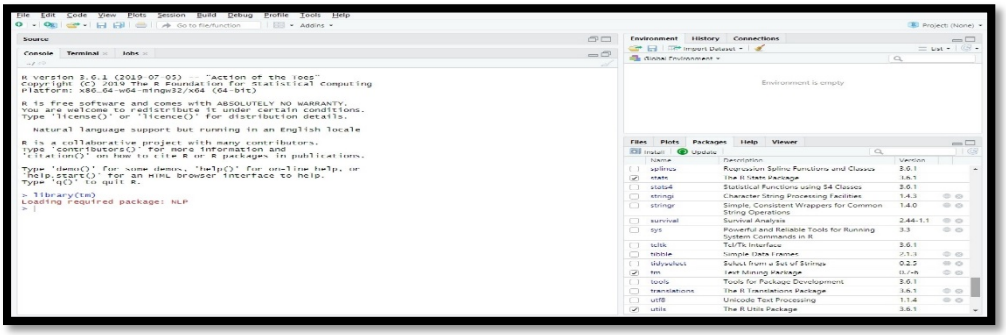

Figure 5: Text Mining tool

To assess the efficiency of the system, there are typical measurement metrics: consistency, recall and f-measure. In extraction data research, these metrics are commonly used (Sampathkumar et al., 2014). Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives. The formula is:

$$Precision = \frac{tp}{tp+fp}$$

Where,

tp is true positive,
fp is false positive.

Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Recall is calculated by using the formula:

$$Recall = \frac{tp}{tp+fn}$$

tp is true positive,
fn is false negative.

To obtain a single metric known as F-measure, which is the weighted harmonic mean of precision and recall, precision and recall can be mixed. The biggest benefit of using F-measure is that a system with one specific ranking may be graded. The formula is:

$$\text{F-}measure = 2 * \frac{precision*recall}{precision+recall}$$

4. Results and Discussion

4.1 Overview

In this chapter, based on the stream handle in past chapter content pre-processing, prepare of classifying for etymological approach which is stage 4 of the method will be clarify more subtle elements counting the Back Vector Machine (SVM) method utilized to classify the content. The apparatuses include in those handle too will be clarify more points of interest in this chapter.

4.2 Text per-processing process

As specified in the previous chapter, pre-processing step acts as an important role in text mining process because it is the step where the raw document is being clean right before the keyword extraction

and classification take place. In this study, data cleaning phase, stop word removal and stemming is the tasks done during text pre-processing.

4.2.1 Data Cleaning

Data cleaning process comprises of few steps, such as removing punctuation, removing numeric, converting characters to lower case and removing white space. All cleaning steps are performed in this prepare utilizing orders. Figure 6 shows the original text document, while Figure 7 until while 10 shows the result after the text document is cleaned in RStudio.

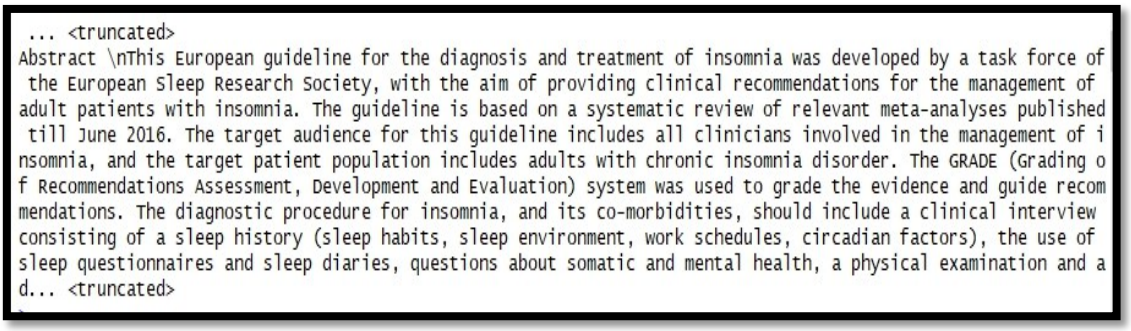

Figure 6: The original text before data cleaning

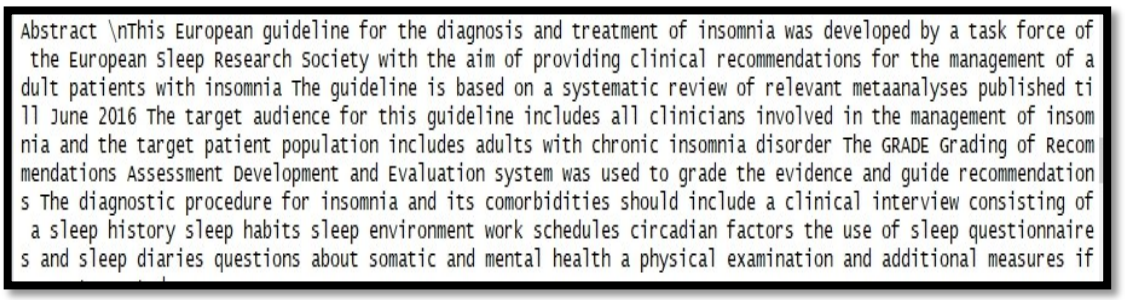

Figure 7: Result of punctuation removal

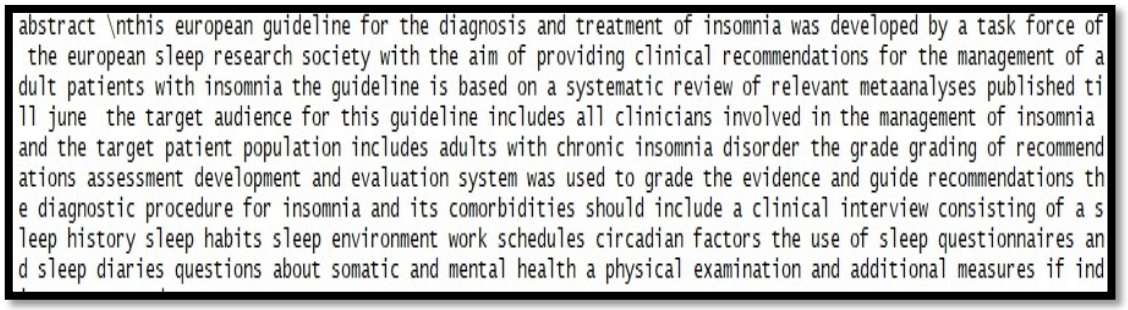

Figure 8: Result of numeric removal

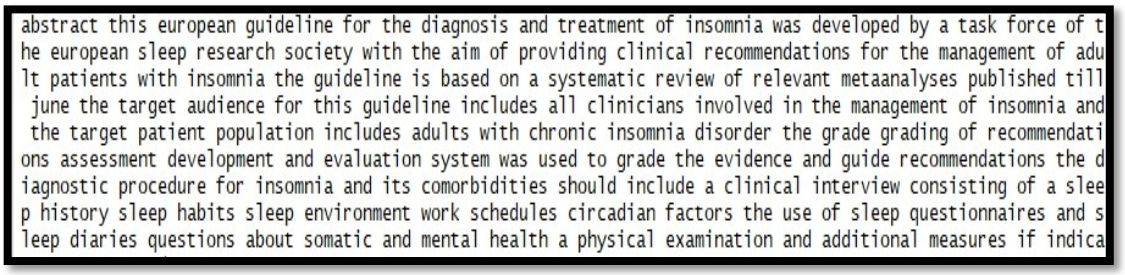

Figure 9: Result of converting character to lowercase

```
Abstract \nThis European guideline for the diagnosis and treatment of insomnia was developed by a task force of
 the European Sleep Research Society with the aim of providing clinical recommendations for the management of a
dult patients with insomnia The guideline is based on a systematic review of relevant metaanalyses published ti
ll June  The target audience for this guideline includes all clinicians involved in the management of insomnia
and the target patient population includes adults with chronic insomnia disorder The GRADE Grading of Recommend
ations Assessment Development and Evaluation system was used to grade the evidence and guide recommendations Th
e diagnostic procedure for insomnia and its comorbidities should include a clinical interview consisting of a s
leep history sleep habits sleep environment work schedules circadian factors the use of sleep questionnaires an
d sleep diaries questions about somatic and mental health a physical examination and additional measures if ind
```

Figure 10: Result of whitespace removal

4.2.2 Stop Word Removal

Since the data cleaning step is completed, there are still meaningless stop word data that involves the removal of stop word processing. In fact, stop word removal is the method in which words frequently occurring in all the corpus documents are eliminated. The process is therefore necessary so as to prevent its presence from affecting the end result (17).

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"
 [56] "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"
 [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
 [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
 [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
 [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
 [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
[101] "who's"      "what's"     "here's"     "there's"    "when's"
[106] "where's"    "why's"      "how's"      "a"          "an"
[111] "the"        "and"        "but"        "if"         "or"
[116] "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"
[126] "against"    "between"    "into"       "through"    "during"
[131] "before"     "after"      "above"      "below"      "to"
[136] "from"       "up"         "down"       "in"         "out"
[141] "on"         "off"        "over"       "under"      "again"
[146] "further"    "then"       "once"       "here"       "there"
```

Figure 11: List of English Stop Word

```
abstract  european guideline  diagnosis  treatment  insomnia  developed  task force  european sleep research
 society  aim  providing clinical recommendations  management  adult patients  insomnia  guideline  based  s
ystematic review  relevant metaanalyses published till june  target audience  guideline includes  clinicians i
nvolved  management  insomnia  target patient population includes adults  chronic insomnia disorder  grade gr
ading  recommendations assessment development  evaluation system  used  grade  evidence  guide recommendations
 diagnostic procedure  insomnia  comorbidities  include  clinical interview consisting  sleep history sleep h
abits sleep environment work schedules circadian factors  use  sleep questionnaires  sleep diaries questions  s
omatic  mental health  physical examination  additional measures  indicated ie blood tests electrocardiogram el
ectroencephalogram strong recommendation moderate  highquality evidence polysomnography can  used  evaluate  sl
e... <truncated>
```

Figure 12: Result of stop word removal

4.2.3 Stemming

Figure above displays the text result until the stemming process is completed. In RStudio, the feature "StemDocument" is used as a command to perform stemming. SnowballC is the sort of stemming utilized in this consider. SnowballC bundle has been stacked to appropriately execute the stemming instruction. This stemmer actualizes the word stemming calculation utilized by Watchman to separate words into a common root. Since stemming is the method of evacuating the fastens from the

word and extricating the word root or stem, it helps to play down the full number of plain terms in a record that influences preparing time and memory space.

```
abstract european guidelin diagnosi treatment insomnia develop task forc european sleep research societi aim pr
ovid clinic recommend manag adult patient insomnia guidelin base systemat review relev metaanalys publish till
june target audienc guidelin includ clinician involv manag insomnia target patient popul includ adult chronic i
nsomnia disord grade grade recommend assess develop evalu system use grade evid guid recommend diagnost procedu
r insomnia comorbid includ clinic interview consist sleep histori sleep habit sleep environ work schedul circad
ian factor use sleep questionnair sleep diari question somat mental health physic examin addit measur indic ie
blood test electrocardiogram electroencephalogram strong recommend moder highqual evid polysomnographi can use
evalu sleep disord suspect ie period limb movement disord sleeprel breath disord treatmentresist insomnia profe
ssion atrisk popul substanti sleep state mispercept suspect strong recommend highqual evid cognit behaviour the
r... <truncated>
list(language = "en")
list()
```

Figure 13: Result of stemming

4.2 Keyword Extraction

In this research, RStudio is used for keyword extraction tool to complete the task. RStudio produce result based on the term frequency for the keyword extraction technique. The RStudio command used to extract keywords in documents as seen in Figure above. In this approach the word frequency was statistical analysis for the extraction process. Term frequency is how frequently a word shows up in a content, and regularly characterizes the setting of the words. "isn" is the title of the corpus utilized to end up input in this process. The generated terminal frequency of the tool will be shown in the format of the document term matrix. "isn.csv" is the file name that extracts the document term matrix result into Microsoft Excel.

```
>
> #transpose of the matrix
> tdm <- TermDocumentMatrix(insomniadocs)
> tdm
<<TermDocumentMatrix (terms: 2340, documents: 200)>>
Non-/sparse entries: 15997/452003
Sparsity           : 97%
Maximal term length: 33
Weighting          : term frequency (tf)
>
> #organize terms by frequency
> freq <- colSums(as.matrix(dtm))
> length(freq)
[1] 2340
> ord <- order(freq)
>
> #export matrix to Excel
> m <- as.matrix(dtm)
> dim(m)
[1]  200 2340
> write.csv(m, file="isn.csv")
>
```

Figure 14: Command to produce term frequency

Doc	abstract	acupunctur	addit	adult	age	agonist	aim	altern	antidepres	antihistami	antipsycho	assess	atrisk	audienc	avail	base	behaviour
m	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	2
m	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
m	1	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1
m	1	0	0	1	0	2	0	0	1	0	1	0	0	0	1	0	0
m	1	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0
m	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
m	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	0
m	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
m	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	1	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0
m	1	0	1	0	1	0	0	2	0	0	0	0	0	0	1	0	0
m	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
m	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
m	1	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
m	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
m	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2	0
m	1	0	1	0	1	4	0	0	2	0	0	0	0	0	1	1	0
m	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
m	1	0	1	0	0	2	0	0	1	0	0	0	0	0	1	0	0
m	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
m	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
m	1	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Figure 15: Extracted keyword in document matrix

4.4 Classification

After information extraction technique was finish, now the process of classification technique will be use. The appropriate classifier which is Naïve Bayes will be inserted by pressing the "Select" button in the Classification section. At that point the gadget utilized to prepare the information is actualized by clicking on the Radio button within the Check Choices zone. In this consider the K-fold Cross Approval strategy is utilized with the esteem k=10. Esteem 10 was selected for being the foremost broadly utilized machine learning instrument. At long last, within the Classifier Yield field the classification result comprising of a portrayal of stratified cross validation, comprehensive course precision, and disarray lattice is shown.

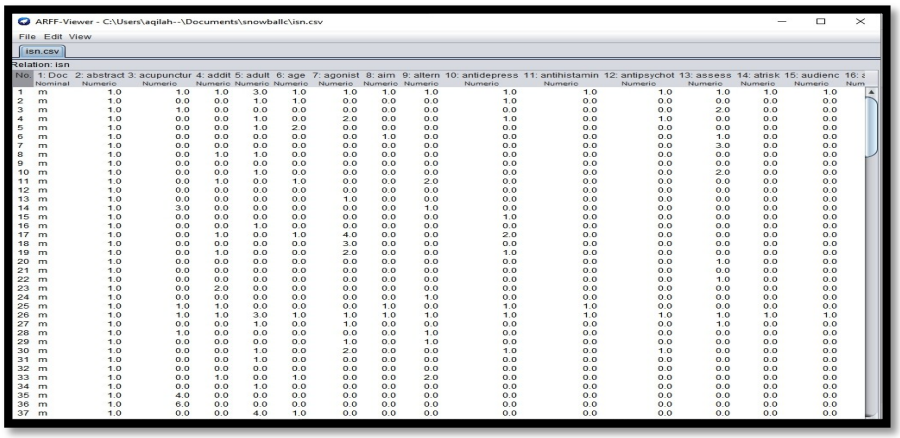

Figure 16: .Arff format

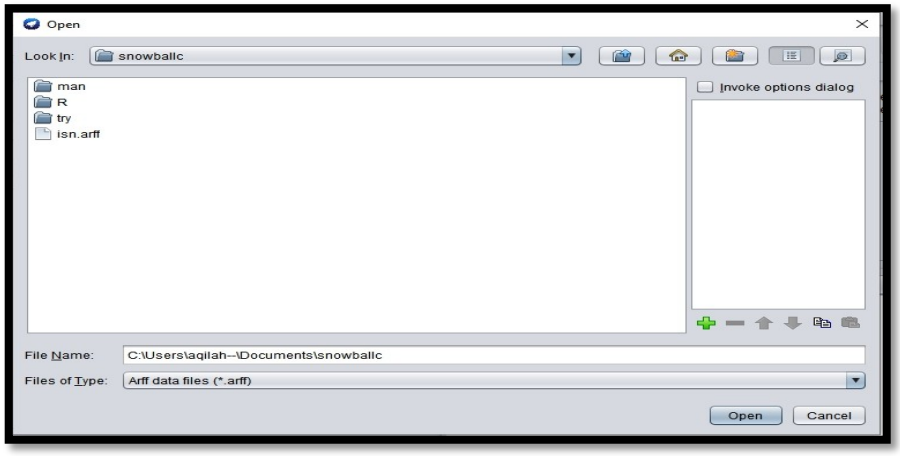

Figure 17: WEKA

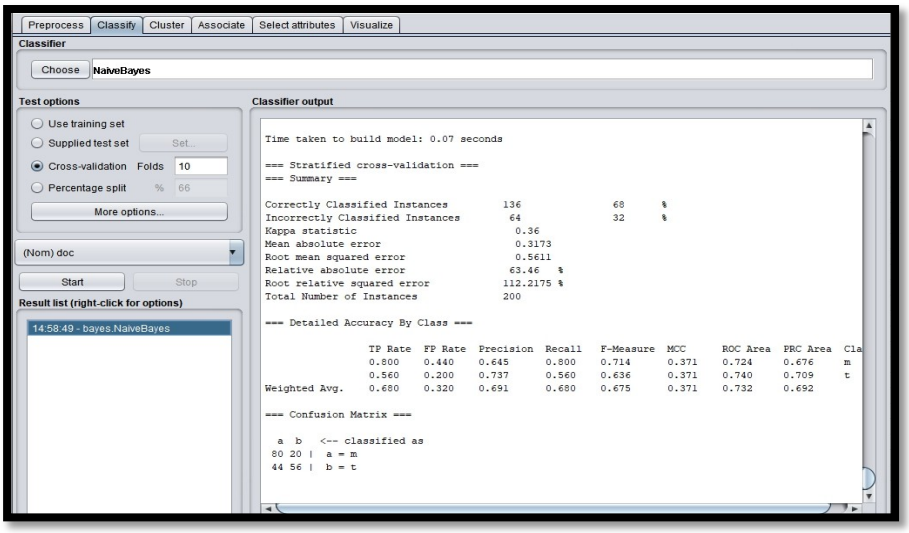

Figure 18: Classified in WEKA

4.5 Result and Discussion

There is a set of standard evaluation metrics used to assess the method's performance: precision, recall, and f-measure. These metrics were widely used in information studies on the extraction. The accuracy of a classifier is calculated by measurement. A higher precision means less false positives, while a lower precision means more false positives. Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Precision and recall can be combined to produce a single metric known as F-measure, which is the weighted harmonic mean of precision and recall. The main advantage of using F-measure is that it is able to rate a system with one unique rating.

Table 1: Results of five algorithm

Naïve Bayes						
Diseases	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Insomnia	medicine	0.800	0.440	0.645	0.800	0.714
	therapy	0.560	0.200	0.737	0.560	0.636
PART						
Diseases	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Insomnia	medicine	0.680	0.240	0.739	0.680	0.708
	therapy	0.760	0.320	0.704	0.760	0.731

Table 1: (cont.)

IBk						
Diseases	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Insomnia	medicine	0.620	0.300	0.674	0.620	0.646
	therapy	0.700	0.380	0.648	0.700	0.673
SGD						
Diseases	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Insomnia	medicine	0.650	0.250	0.722	0.650	0.684
	therapy	0.750	0.350	0.682	0.750	0.714
RandomForest						
Diseases	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Insomnia	medicine	0.680	0.240	0.739	0.680	0.708
	therapy	0.760	0.320	0.704	0.760	0.731

The higher the F-measure value, the better the predictive ability of the classification procedure. A score of 1 means it is a perfect result but when the score is 0 it becomes the lowest possibility. There are five algorithm that has been used. The first is Naïve Bayes classifying method, where insomnia F-measure score for medicine is 0.714 while therapy is 0.636. The second algorithm that has been used is PART classify method, where insomnia F-measure score for medicine is 0.708 while therapy is 0.731. The third algorithm is IBk classify method, which shows insomnia F-measure score for medicine is 0.646 while therapy is 0.673. The fourth algorithm is SGD classify method, with insomnia F-measure score for medicine is seen at 0.684 while therapy is also at 0.714. The fifth algorithm is RandomForest classify method, in which insomnia F-measure score for medicine is 0.706 while therapy is 0.731. For the conclusion, based on the result by using different algorithms we can see that therapy has higher possibilities than medicine. The diseases in this research which is insomnia show that therapy is more suitable for treatment event this method treatment take a long time to recover.

5. Conclusion

The f-measure value are obtained to measure the highest possibilities of treatment. By using different algorithm in Support Vector Machine (SVM), the result that produce is compared to get the best treatment for disease whether medicine or therapy. In this experiment, therapy l has the higher possibility than medicine.

5.1 Problems and limitation of research

There are numerous issues that have been experienced whereas conducting this research. The issues have been confronted are:

i. The place that does the experiment must have great internet coverage because RStudio tool needed high speed internet coverage to make sure that the command of process is running smoothly.
ii. Need a lot of time to study about the technique used in this research including the extraction tools. Lack of knowledge make the process become slowly and have many interruptions because all the process must study first before executing it.

5.2 Improvement and future works

There are a few proposals relating to this investigate for encourage improvement. This suggestion may be forward-looking suggestions or ways to create this research. Any exhortation that will help deliver superior outcomes. The enormous, large of information accessible online nowadays comprises basically of plain text, clinical information, reports, or electronic health records. The text are fundamentally the languages which the human understands it. Although there are plenty of tools to extract the document, they are not all sufficient to extract biomedical text. The inspiration is to progress the tooling method, so the individuals around the world can utilize it effectively and can fit the sort of assortment to form it more widespread.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support and encouragement throughout the process of conducting this study.

References

[1] K.U Alper and G. Serkanl "The impact of preprocessing on text classification" Information Processing & Management 50, 2014, 1 (2014), 104–112

[2] C.Archana "Text Mining Methods and Techniques" November 2014.

[3] C. Corinna and V. Vladimir, "Support-vector networks". Machine learning 20, July 1995, 273–297.

[4] H. David Wolpert. "Stacked generalization" Neural Networks August 1992. 5:241-259.

[5] D. Aha, D. Kibler "Instance-based learning algorithms" Machine Learning November 1991. 6:37-66.

[6] D. Alex , F. Alexa, "Treatments of Insomnia", September 18, 2020, [Online]. Available: https://www.sleepfoundation.org/insomnia/treatment

[7] R. Anis, S. Eric "Insomnia" September 4, 2020 [Online]. Available: https://www.sleepfoundation.org/insomnia

[8] F. Eibe, H. Ian, Witten: Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

[9] H. Lodhi,. "Text Classification Using String Kernels," J. Machine Learning Research, March 2002, vol. 2, pp. 419-444.

[10] S. Hassan, F. Miriam, H. Yulan, and A Harith. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. October 2014.

[11] J. W. Jonathan and K. Chunyu., Tokenization as the initial phase in NLP. In Proceedings of the 14th conference on Computational linguistics-Volume 4.Association for Computational Linguistics, April 1992 ms 1106–1110.

[12] B. Leo “Random Forests”. Machine Learning April 2001. 45(1):5-32.

[13] S. Marc, F. Eibe, H. Mark “ Speeding up Logistic Model Tree Induction”. In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 675-683, 2005.

[14] Mayo Clinic Insomnia [Online]. Available: https://www.mayoclinic.org/diseasesconditions /insomnia/symptomscauses/syc20355167#:~:text=Insomnia%20is%20a%20common%20sleep,tired%20when%20you%20wake%20up.

[15] A. Mehdi “A Brief Survey of Text Mining”: Classification, Clustering and Extraction Techniques July 2017

[16] A. Rashmi , B. Mridula “A Detailed Study on Text Mining Techniques” January 2013

[17] J. K. Raulji, Scholar, R., Ambedkar, B., Saini, J. R., Director, I. / C., and Supervisor, R. Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. International Journal of Computer Applications, January 2016, 150(2), 975–8887. doi 10.5120/ijca2016911462

[18] V.G Sonali “Text Mining Methods and Techniques” March 2014.

[19] weka.classifiers.functions [Online]. Available: https://weka.sourceforge.io/doc.stable-38/weka/ classifiers/ functions/SGD.html