Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7923149

FEATURE EXTRACTION FOR TEXT MINING WEIGHT BASED CORE CORPUS TF-IDF (W2CTF-IDF)

[1] Dr.P.Logeswari, [2] S.Sudha, [3] G.Banupriya, [4]J.Gokulapriya
[1]Assistant Professor, School of Computer Applications,
Lovely Professional University
[2,3,4]Phd Research Scholar , [2,3,4]Department of Computer Science,
[2,3,4] Sri Krishna Arts & Science College,
[1]Phagwara,Punjab,[2,3,4] Coimbatore, Tamilnadu, India.
[1]tppselvalogu@gmail.com, [2]sudhasw89@gmail.com, [3]banu.snmv7@gmail.com
[4]gokulapriyajaganathan1@gmail.com

Abstract - Text mining, otherwise called Intelligent Text Analysis is a significant exploration region. It is extremely challenging to zero in on the most suitable data because of the great dimensionality of information. Highlight Extraction is one of the significant techniques in information decrease to find the main features. Processing a huge measure of information put away in an unstructured structure is a challenging undertaking. Feature extraction is one of the huge pre-processing techniques in information mining that registers features esteem in documents. Thus, productive element extraction techniques term frequency-inverse document frequency (TF-IDF) techniques are regularly used in term weighting. This issue can't mean the accommodation or significance of certain features and diminishes the productivity of characterization. The record server executes stop word expulsion, labelling, and the examination of polysemous words in a pre-processing methodology to make a competitor corpus. Weight-based Core Corpus TF-IDF (W2CTF-IDF) is proposed to the competitor corpus to assess the significance of words in a bunch of documents. The words named of high significance by W2CTF-IDF are remembered for a bunch of keywords, and the transactions of each document are made. The technique is assessed W2CTF-IDF to weight the terms on financial data. The experiments show that W2CTF-IDF further develops the performance evaluation of component extraction as indicated by the maximum worth of the F1 measure.
Keywords: Feature extraction, Text Mining, TF-IDF, Weight-based Core Corpus TF-IDF and Weighting;

1. Introduction

Text Mining is the disclosure by PC of new, previously dark information, by means of normally isolating information from different making resources. A key part is the interfacing together of the removed information together to shape new real factors or new hypotheses to be researched further by additional regular strategies for experimentation. Text mining isn't exactly equivalent to what realize about in web search. In search, the client is ordinarily looking for something most certainly known and has been created by someone else [5]. The issue is pushed aside all the material that at present isn't relevant to your prerequisites to find the appropriate information. In text mining, the objective is to track down dull information; something that nobody yet knows and consequently couldn't have ever yet recorded.

Include extraction (FE) is one of the layered decreases that are routinely used on datasets with tens or countless features. Viewpoint decrease is one of the association used in different applications, for instance, data mining and recovery of information. The principal objective of viewpoint decrease is to diminish high-layered data in a lower layered subspace, while basic features of the primary data are kept whatever amount as could sensibly be anticipated. The best test in highlight extraction comes from different assortments of key terms (include) for each report in the dataset. They are the text pre-processing steps, the association requesting techniques as well as the similarity measure.

Text data is the clearest kind of data which is unstructured in nature. It is made in the massive total by and large. Individuals can doubtlessly see and communicate with unstructured text data anyway it is challenging for machines to get something almost identical. This voluminous text data is a critical wellspring of information and information [14]. Thusly, to use this information removed from text data actually an assortment of applications, techniques and estimations are required. NLP has gained a great deal of thought in past several years because of the colossal proportion of text data gets made in many designs like interpersonal organizations, patient records, media sources, clinical benefits security data, etc in a report created by EMC. It is predicted that, by 2020, the volume of data will foster up to 40 zettabytes. It is difficult for individuals to go through all such text data and find the information of interest and coordinate colossal proportions of data. To enable the strong change and representation of such data, the cycle incorporates computing the word frequencies from the report and in the entire assortment of archives. In this manner, eliminating the needful information from the unstructured text data is critical. Isolating information from text assists in separating the text data for various applications, reports, clinical records, automated terminology the leaders, research with subjecting recognizable confirmation, data mining and focusing on the effect of investigation on them, etc Feature Extraction is a fundamental methodology in dimensionality decrease to remove the huge features.

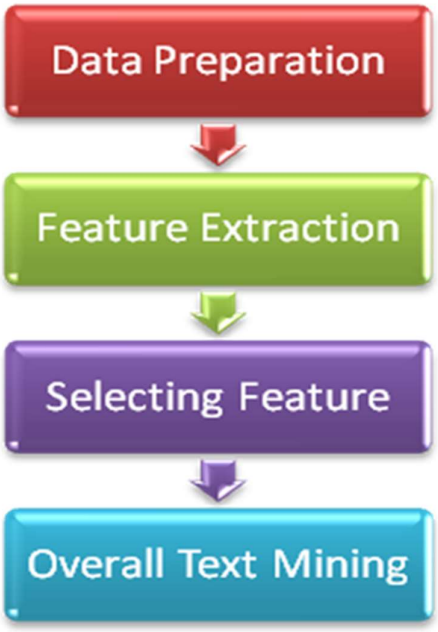

Figure 1 Text Mining for Feature Extraction

Text feature extraction that isolates text information is an extraction to represent a text message; it is the premise of boundless text processing [9]. The fundamental unit of the feature is called text features. Choosing a lot of features from an effective approach to decreasing the part of feature space, the inspiration driving this cycle is called feature extraction. During feature extraction, uncorrelated or pointless features will be eradicated. As a technique for data pre-processing of learning computation, feature extraction can all the more probable work on the exactness of learning estimation and contract the time. Decision from the record part can mirror the information on the substance words, and the computation of weight is known as the text feature extraction. Ordinary techniques for text feature extraction consolidate filtration, combination, planning, and grouping strategy. Standard techniques for feature extraction require great features. To hand-plan, a convincing feature is an extensive cycle, and significant learning can be centered around new applications and straightaway get new suitable characteristic representations from training data.

Text Mining

Text mining is one more area of software engineering which encourages strong relationship with normal language processing, data mining, AI, information recovery and information on the board. Text mining hopes to remove supportive information from unstructured textual data through the recognizable verification and examination of charming examples [2].

"There is gold secret in your association's data" - and data mining promises to assist you with tracking down it. Additionally, in all honesty, various effective applications of data mining show that this is extremely substantial. Regardless, data mining will in general be only a very limited piece of an association's finished data resources: the coordinated information accessible in databases. Presumably, more than 90% of an association's data are never being looked at: letters from clients, email correspondence, records of calls with clients, contracts, specialized documentation, and licenses, With really dropping expenses of mass amassing, organizations accumulate progressively a greater amount of such data on the web. When in doubt, the primary way the data is made usable - beyond very certain applications for subsets of that data - is by making it accessible and accessible in an affiliation's intranet. Notwithstanding, today there is more you can do: text mining assists with recovering the hidden gold from textual information. Text mining hops from old-fashioned information recovery to information and information discovery.

General Framework for Text Mining

Text Mining can be envisioned as containing two stages: (I) Text refining and (ii) Knowledge refining. The text refining stage changes the free construction of text documents into a picked moderate design. Information refining interprets examples or information from a midway design. The Intermediate Form (IF) can be semi-coordinated like the applied chart representation or coordinated like social data representation [11]. The center design can be document based wherein each substance represents a document or thought based wherein each component represents a thing or thought of interest in an express region. Mining a document based IF decides examples and connections across documents. Document clustering/visualization and arrangement are examples of mining from a document based IF.

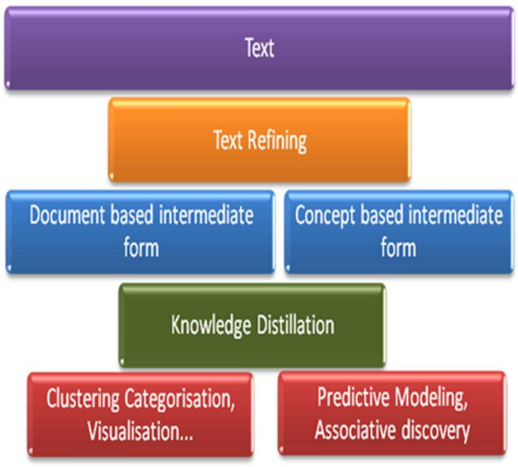

Figure 2.2 General framework for Text Mining

Mining a thought in view of IF decides models and connections across things or thoughts. Data mining tasks, for instance, predictive displaying and cooperative discovery have a spot in this grouping [6]. A document-based IF can be changed over into a thought based IF by removing the significant information as shown by the objects of interest in a specific region. It follows that document based IF is commonly space autonomous and thought based IF is region subordinate. For instance given a ton of reports, text refining starting believers each document into a document based IF. One can then perform information refinement on the document based IF to collect the articles, as per their substance for visualization and course purposes For Knowledge discovery in a particular space the document based IF of the reports can be projected onto an idea set up IF depending with respect to the undertaking prerequisite. For example, one can remove information associated with "product" from the document based IF and structure a product database to decide product-based information. The overall framework for Text Mining is shown in Figure 2.2.

Areas of Text Mining

Text analysis consolidates information retrieval information extraction and data mining strategies including affiliation and affiliation analysis, visualization and predictive assessment. The objective is, fundamentally to turn text (unstructured data) into data (coordinated plan) for analysis, through the usage of natural language processing (NLP) techniques.

Figure 2.3 Text Mining Areas

Text mining is an interdisciplinary field which fuses regions, for example, information retrieval, information extraction, data mining, computational etymology and natural language processing. The areas of text mining is shown figure 2.3.

2. Existing Methodology

1. Yun Ye, XiaoJuan Zhao, XiaoMing Zhang [12] proposed Web Text Feature Extraction with Bean Optimization Algorithm (BOA). This paper presents another site page text feature extraction algorithm. This algorithm relies upon the bean advancement algorithm and the construction of the bean position vector has in like manner been gotten to a higher level. This algorithm can productively glance through complex multi-faceted spaces. This strategy presents a nice idea for diminishing feature size and dealing with the effectiveness of web document processing. The Internet continues to create at a heavenly rate and how much information on the web is overpowering. It provides us with a ton of information resources. Web text feature extraction is viewed as the chief issue in text mining. We use Vector Space Model (VSM) as the depiction of web text and present a novel feature extraction algorithm which relies on the better Bean Optimization Algorithm (BOA). This algorithm will uncommonly chip away at the productivity of web text processing. Presently we are continuing with our experiments about the algorithm and we will convey a more prominent measure of our consequences of the experiments in times to come. Later on, we will focus in on applying this strategy to the fields like web documents separating, request and clustering in agrarian regions.

2.P.C. Barman, Nadeem Iqbal et.al proposed [15] Non-negative Matrix Factorization Based Text Mining: Feature Extraction and Classification. The unlabeled document or text assortments are ending up being progressively huge which is typical and self-evident; mining such data sets are a troublesome task. Using the basic word-document frequency matrix as feature space the mining framework is ending up being more convoluted. The text documents are regularly represented as high layered about scarcely any thousand insufficient vectors with sparsity around 95 to practically 100 percent which in a general sense impacts the effectiveness and the consequences of the mining framework. In this paper, we propose the two-stage Non-

negative Matrix Factorization (NMF): in the principal stage we endeavored to isolate the uncorrelated premise probabilistic document feature vectors by in a general sense diminishing the element of the feature vectors of the word-document frequency from few thousand to two or three hundred, and in the second stage for clustering or grouping. In our proposed approach, it has been seen that the clustering or characterization execution with more than 98.5% exactness. The aspect decrease and arrangement execution have been seen for the Classic3 dataset.

3. Shivaprasad KM, Dr. T Hanumantha Reddy [9] proposed Text Mining: An Improvised Feature Based Model Approach. In this knowledge time, a plethora of textual information is growing rapidly which is normally semi-organized or unstructured data accumulated and set aside in various databases. The Discovery of knowledge from this accessible database isn't clear. As needs be, the programmed feature decision procedure is a ton of fundamental in the processing of this unstructured data. The Feature assurance approach fixates on processing unstructured information into material information and helps with understanding and envisioning the data, it furthermore reduces the training and processing season of immense proportions of data. The feature helps in finishing the text mining task effectively and precisely. Text mining offers various techniques to bring the fascinated data from gigantic databases. Feature decision shows to have a pivotal impact in this cycle. In this paper, the bread model is suggested that processes text documents using the information terms. Considering the primary guidelines of direction application procedure, the model stages are completed that give strong results.

4. A Bodile [1] et.al proposed Text Mining in Radiology Reports by Statistical Machine Translation (SMT) Approach. Clinical text mining has procured growing acclaim of late. Presently days, huge measures of clinical text data are consistently made in prosperity organizations, in any case, positively no point suggests the future as it is a very tedious endeavor. In the Radiology space, by far most of the reports are in free text plan and regularly natural, in this way it is challenging to get to the significant information for clinical experts aside from in the event that authentic text mining isn't applied. There are a couple of frameworks existing for radiology report information retrieval like MedLEE, NeuRadIR, and CBIR anyway relatively few of them use text connected with pictures. This paper proposes a text mining framework to deal with this issue by using a measurable machine interpretation approach. The System stores the text and picture feature to notice the match report. The SVM classifier is utilized in the SMT method for managing checks whether or not entered report present in the database. The framework will return the similar report coordinated with the entered report from the database.

5. Hui He, Bo Chen, Weiran Xu, Jun Guo [4] proposed Short Text Feature Extraction and Clustering for Web Topic Mining. This examination is to present an algorithm with bunch Chinese short texts for mining web subjects considering Chinese pieces. Zeroing in on the characteristics of Chinese short texts, the algorithm uses N-gram feature extraction to get Chinese lumps from texts, which mirror the text semantic construction and character dependence. Then, the RPCL algorithm is applied to perceive text clustering with high accuracy, which doesn't have to know the specific number of packs. Finally, the analysis results demonstrate the way that this approach can astoundingly decline the dimensionality and truly

work on the display of Chinese short texts clustering than standard strategies. Perhaps the best responsibility of this paper is that it is the underlying opportunity to apply factual string decrease algorithm to isolate Chinese pieces as text features from short text assortment. In future work, we could get a remove from the opportunity to solidify more semantic knowledge and weight techniques of different sizes of pieces in our strategy.

3 Proposed Methodology

Text feature extraction is the process of taking out a list of words from the text data and then transforming them into a feature set which is usable by a classifier. The following techniques can be used for extracting features from text data.

Text Mining

Text mining is very well might be described as the method involved with investigating text to extricate information that is helpful for a particular reason. By the by, in present day culture, the text is the most common way for the conventional trade of information. Text mining as a rule manages texts whose capability is the correspondence of real information or feelings, and the upgrades for attempting to extricate information from such text naturally is convincing regardless of whether achievement is just halfway. These are syntactic properties that together represent currently characterized classifications, ideas, faculties or implications [13]. Text mining should perceive, concentrate and utilize the information. Rather than looking for words, we can look for semantic examples, and this is thusly looking at a more elevated level.

Data

The datasets are taken from Kaggle. The dataset is connected with cancer which contains the qualities, genetic mutations brought about by cancer and clinical text. The datasets are given by means of two distinct files - training and test. One is called training/test variations which has information about the genetic mutations. The other training/test text gives the clinical text which is clinical examination papers connected with cancer which are utilized by human specialists used to characterize genetic mutations. The two files can be connected by means of the ID field. The training dataset contains 3000 instances.

Feature Extraction

Words and expressions are the structure blocks of customary documents, and there is routineness in how frequently each term shows up in various documents. Along these lines, recognizing documents with various content can be utilized. Feature extraction (FE) is one of the layered decreases that are normally utilized on datasets with tens or countless features.

FE is a significant methodology it gives a ton of information seeing the text documents like the most elevated and least term frequency for each document. Choosing significant features and determining how to encode them for a learning machine strategy can unfathomably affect the learning machine technique's capacity to remove a decent model [5]. This study will utilize TF-IDF terms weighting to extricate features. Characterize a keyword set for every text document classification, Transform Text to Numerical Feature by counting the quantity of occurrences (called term frequency) in the text document.

Feature Extraction Using TF-IDF Method

Feature extraction is a critical stage in text mining that it gives information as per the texts like the maximum and least term frequency for each document. Choosing the connected features and determining the scope of effect for machine learning. Moreover, the capacity of

separating was a decent feature of the training model. TF-IDF is usually referred to and is utilized as a weighting technique and its performance is still even equivalent to novel techniques. Documents are considered as the elements in the term weighting. Choosing the feature for the feature determination system considers the primary pre-processing process that is expected to file the documents.

It is vital to pick the expected keywords that convey the importance and deny the words that don't uphold acknowledgment between the documents. Separating assortment features by joining individual document features. These means referenced above are portrayed in Figure 3.

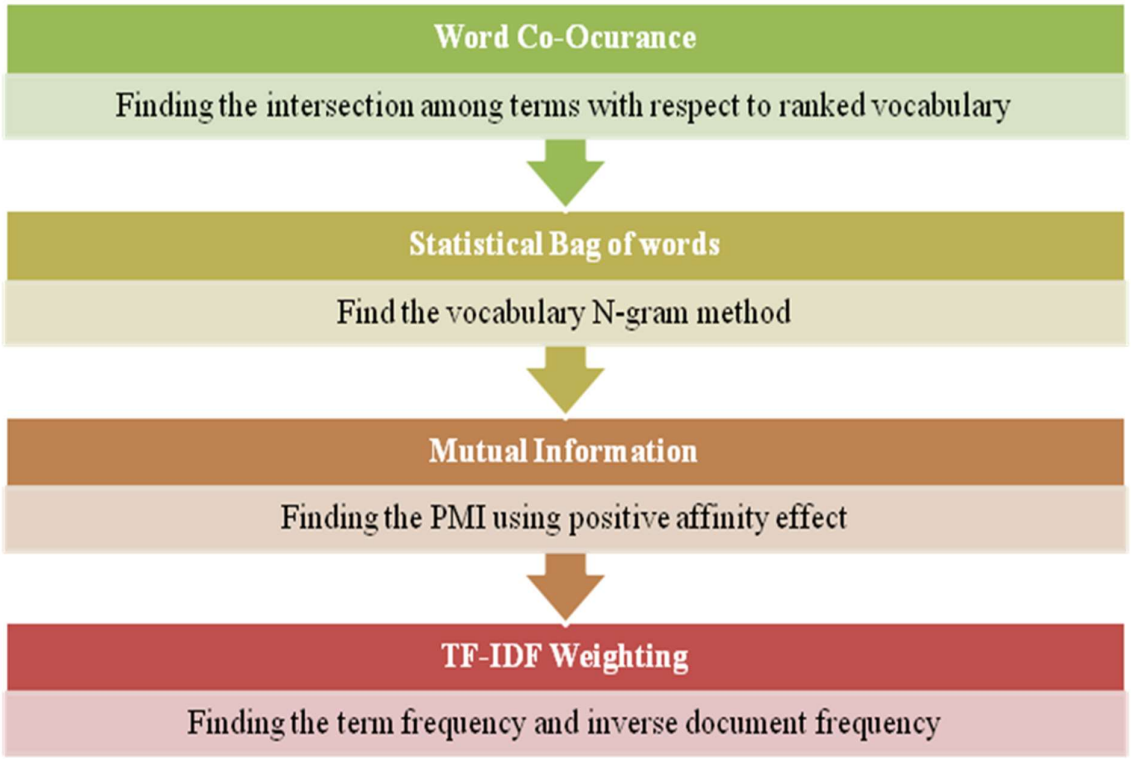

Figure 2.4 Indexing technique feature extraction stages

Term weighting can be basically as straightforward as a parallel representation or as nitty gritty as a blend of terms and existing datasets. TF-IDF is the most well known and utilized weighting technique, and it stays even similar with novel strategies. In TF-IDF term weighting, the text documents are described as transactions. Picking the keyword for the feature determination process is the principal pre-processing process important for the ordering of documents. This study used two different TF-IDF techniques, for example Worldwide and typical, to weight the terms in term-document frameworks of our evaluation datasets. A typical practice to stay away from this changeability or, in any event, lessen the potential effects coming about because of it, is the standardization of the TF-IDF scores for each document in the assortment by utilizing the Euclidean standard determined by utilizing Equations (1) and (2) separately as follows:

$$TF - IDF = In(TF) \times In(IDF)$$
$$TF - IDF = TF \times IDF$$

The TF-IDF technique is utilized to weight and concentrate the keywords from the dataset. A keyword can happen in variation classes that produce changeability. This changeability is the grade of contrasts in keywords regarding the various classes that contain this keyword. A common point is to tackle this fluctuation by reducing the potential impacts resulting from it [100.

The normalization TF-IDF scores for each document in the set via the Euclidean norm are applied. The computations of TFIDF are displayed in Equations 3

$$(TF-IDF)_{ij} = (TF)_i \times \log(IDF)_{ij}$$

$$Neverthless, Equation\ 3\ is$$

$$onlyutilizedinsituationswhen(T-IDF) \geq 1. Otherwise, TF-IDF = 0$$

$$(TF-IDF)_{ij} = \begin{cases} (TF)_i \times \log(IDF)_{ij}\ if(TF), \geq 1 \\ 0 \qquad\qquad\qquad Otherwise \end{cases}$$

Where term frequency (TF) indicates to the standard weighting. It denotes the weight, that indicates to the frequency or the related frequency of the term,t_i within a given tweet, j.
Inverse document frequency (IDF) indicates to the global weighting. It denotes the support of the term tj according to i^{th} belongs to tweets set.

In any case, text pre-processing is used to dispense with the stop words (conjunctions, pronouns, and so forth), URLs, exceptional characters, usernames, accentuations, HTML labels, and so on. Then the keywords are represented in vector space for all the class tweets. The class tweet is changed over completely to numerical keywords by means of computing the quantity of occurrences (called term frequency) as per the keywords in the tweet against the keyword co-occurrences. The keywords are chosen in view of a set edge for each tweet set. A semantic step is extremely influential for select the proposed keywords that conveyed a similar importance of the tweet, and reject different tweets set. Following this, the keywords from the singular tweets are joined. The entire procedure of feature extraction utilizing TF-IDF is shown in Fig. 3.

Composition of Weighted Corpus

A wellbeing document is examined morphologically to permit its partition into different combinations of words. In the pre-processed word combinations, the Term Frequency-Inverse Document Frequency (TF-IDF) esteem is extricated. A word with a high TF-IDF esteem is predicted to be significant in the document. TF-IDF is a strategy for extricating words with high significance from a gathering of different documents. The higher the term frequency (TF) worth of a word, the more significant the word is. The most straightforward method for determining the TF of a word is to utilize a frequency include of the word in a document. In the significance evaluation dependent just upon TF, such as often as possible involved words in a document as "look," "even," "said," and "see" are assessed as significant.

The bigger the quantity of documents containing the word x, the higher the worth ofdf_x. It shows that the word is habitually utilized in various documents. TF-IDF is the worth of TF-IDF for the word x in the document y of the gathered document set N.

$$tf(x,y) = trem(x)withindocument(y)$$

$$idf(x,y) = \log\frac{n}{df_x}$$

$$TF-IDF(x,y) = tf(x,y) \times idf(x,y)$$

Web documents feature a modest quantity of text and those tending to a similar subject are much of the time composed utilizing a comparative subject depending on social conditions. As needs be, the TF-IDF worth of a word connected with a typical subject or interest in a gathered document set is low. In wellbeing large data, "risk" is a word ordinarily found in a corpus and hence its significance is assessed as low. This title is examined morphologically, and afterward core keywords are removed. The hash tag "#" is a word that obviously demonstrates the question of interest contained in a document, and is extricated as a core keyword. An accentuation tag is utilized to design or emphasize a piece of text in HTML5. A core corpus is a bunch of core keywords removed utilizing the title tag, hash tag, and accentuation tag. Throughout assessing the significance of a word removed from a wellbeing document, the degree to which the word is engaged with a core corpus is viewed as the weight.

In the made core corpus, the load for the situation where the word x is checked n times is determined as in Formula (2). This equation presents the weight of the word in the core corpus isolated from the title tag, hashtag, and accentuation tag. Expecting the word x of the competitor corpus removed from 10 archives is tracked down numerous times in the core corpus, its weight is1 + 3/10.

$$w_x = \left(1 + \frac{t_x}{N}\right)$$

$$\mathrm{W2CTF-IDF}\,(\mathrm{x},\mathrm{y}) = \mathrm{tf}(\mathrm{x},\mathrm{y}) \times \mathrm{w\text{-}_x} \times \mathrm{idf}(\mathrm{x},\mathrm{y})$$

$$= tf(x,y \times \left(1 + \frac{t_x}{N}\right) \times \log\frac{N}{df_x}$$

Utilizing the words having a determined W2CTF-IDF esteem more prominent than 2.0, a wellbeing enormous data corpus is made. The wellbeing huge data corpus incorporates the TF, IDF, TF-IDF, and W2CTF-IDF of a word, and is utilized for document information extraction.

Association Analysis between Keywords Using Text Mining

The Apriori mining algorithm is applied to investigate the relationship of keywords. Transactions for affiliation analysis are planned in each document, and things are made with the keywords separated from the health big data corpus. Table 4 presents the health transactions planned in each document [7]. The transaction ID is the document number. Things are made utilizing keywords having a W2CTF-IDF esteem more prominent than 2. The planned health transactions are saved in CSV configuration to permit affiliation analysis and proficient computation.

The keywords that fulfil the base help are utilized to make another candidate set more than once. For affiliation analysis, the data mining instrument Weka 3.8.1 was utilized. In the Apriori algorithm, the base help is a worth more prominent than 2. From the made affiliation controls, the potential relationship of a candidate corpus is investigated and cooperative keywords are found.

Extracting Associative Feature Information

The worth of W2CTF-IDF and associative keywords are utilized to remove associative feature information, which is comprised of the keywords profoundly connected with the keywords of the gathered documents. The words "B" and "C," which have a high relationship with the word "A," the significance of which is high, are likewise exceptionally significant. In this manner, foreseeing keywords in a document and their associative keywords is conceivable. The keywords in the gathered document set are adjusted in the high need request as per their

W2CTF-IDF values. In light of the adjusted keywords, profoundly associative keywords are utilized to make associative feature information.

Algorithm for W2CTF-IDF

Input: {$[D.sup.+], [D.sup.-]$} of Taming Dataset , Discovered features $< T, [DP.sup.+], [DP.sup.-]$,Preliminary term weight function w, Audits A and tips T, Potency function Integer account charge I.

Output: Weight Assigned Terms Collection of

relevant audits C

1: let $n = [absolutevalueof\ [D.sup.+]]$;

2: $T1 = \{t|t\ [memberof]\ pp\ [memberof]\ [DP.sup.+]\}$

3: for each

$$t\ [memberof]\ Tdo$$

4: if t [member of] T1

5: $[MATHEMATICALEXPRESSIONNOTREPRODUCIBLEINASCCI]$

6: then

$$sup\ (t) = d\_sup\ (t, [D.sup.+]);$$

7: else

$$sup(t) = d\_sup(t, [D.sup.-]);$$

8: foreach t [member of] T do

Class = SVM(testdocSpe, TrdocSpe, Label);

9: $Spe(t)T1 = (\{d|d\ [memberof]\ [D.sup.+], t$

$$[memberof]\ D\} - \{d|d$$

$$[memberof]\ [D.sup.-], t\ [memberof]\ D\})/n;$$

10: let

$([T.sup.+], G, [T.sup.-]) = Fclustering\ (T, [DP.sup.+],$

[DP.sup.-], spe());

11: $foreach$

$$t\ [memberof]\ [T.sup.+]\ dow(t) = sup(t) * (1 + spe(t))$$

12: $foreach$

$$t\ [memberof]\ [T.sup.-]\ dow(t)\ =\ sup(t)\ -\ [absolute\ valueof\ (sup(t)\ *spe(t))]$$

13: $pi = [FinTminus, FinG]$;

14: $A = pi;\ C = pi;$

15: while'

$$[absolutevalueofC]\ < Ido$$

16: for all A [member of] A do

17: $Pro\ (A)\ =\ Cov(CUA)\ -\ Cov(C)$

18: $Pri\ (A)\ =\ [beta](1 - Eff(R))) + (1 - [beta]).$

19: end for

20: $[epsilon]\ =\ \{A\ [memberof]\ A:\ Pot\ (CUA)\ [greaterthanor\ equalto]\ [alpha]\}$

21: if ($[epsilon]\ ==\ 0)\ or\ \{max.A\ [memberof]\ Pro\ (A)\ =\ 0$) then

22: break

23: end if

24:$[A.sup.*]\ =\ argmaxA\ [memberof]\ Pro\ (A)/Pri\ (A)$

25: $C\ =\ CUA;\ A\ =\ A\backslash[A.sup.*]$

26: end while

27: return C

Assign n to the count of the number of positive documents. Then every term t that belongs to a positive pattern $[DP.sup.+]$ is assigned to T1. Calculate support values using the deployment support formula depending on whether the term belongs to [D.sup.+] or [D.sup.-] but for all terms belonging to the term set T. Note that we use the modified d_supp formula. The time complexity is $O([n.sup.*][absolutevalueofp])$ where [absolute value of p] is the average size of a pattern, and [absolute value of p]<= |d. Use SVM, a promising classifier to find the appropriate class. Then calculate the specificity value for which the time complexity is O([n.sup.*][absolute value of d]) to calculate spe function. Therefore, the time complexity is

$$O\ ((([[absolutevalueofT].sup.*][n.sup.*][absolutevalueofp]) + ([[absolutevalueofT].sup.*][n.sup.*][absolutevalueofd])) = O([[absolutevalueofT].sup.*]\ [[absolutevalueofd].sup.*]n).$$

Make a function call to FClustering algorithm () to obtain the term categories. Then calculate the term weights depending on the type of term cluster $[T.sup.+]\ or\ [T.sup.-]$. Thus W2CTF-IDF algorithm take a time of

$$O\ [([absolutevalueofT].sup.*]\ [[absolutevalueofd].sup.*]n + [[absolutevalueofT].sup.2])$$

Where the typical document size is $[absoluteworthofd]$ and the count of the pertinent documents is n. The algorithm has numerous recursions expanding the audit assortment C count with new audit A. At every recursion, for a given audit works out the profit (Pro) and cost (Pri). Select $[A.sup.*]$ which has the greatest Profit-to-price ratio. Here $[absolutevalueofT] < [absolutevalueofd]$; so, Algorithm W2CTF-IDF is efficient.

4 Experimental Results

Precision

Dataset	IDF	TF-IDF	Proposed W2CTF-IDF
50	66.45	74.12	87.76
100	69.78	71.89	90.89
150	74.91	67.35	92.41
200	79.33	68.98	95.56
250	86.86	65.33	97.12

Table 1. Comparison tale of Precision

The Comparison table 1 of Precision Values explains the different values of existing IDF, TF-IDF and proposed W2CTF-IDF. While comparing the Existing algorithm and proposed W2CTF-IDF, provides the better results. The existing algorithm values start from 66.45 to 86.86, 65.33 to 74.12 and proposed W2CTF-IDF values starts from 87.76 to 97.12. The proposed method provides the great results.

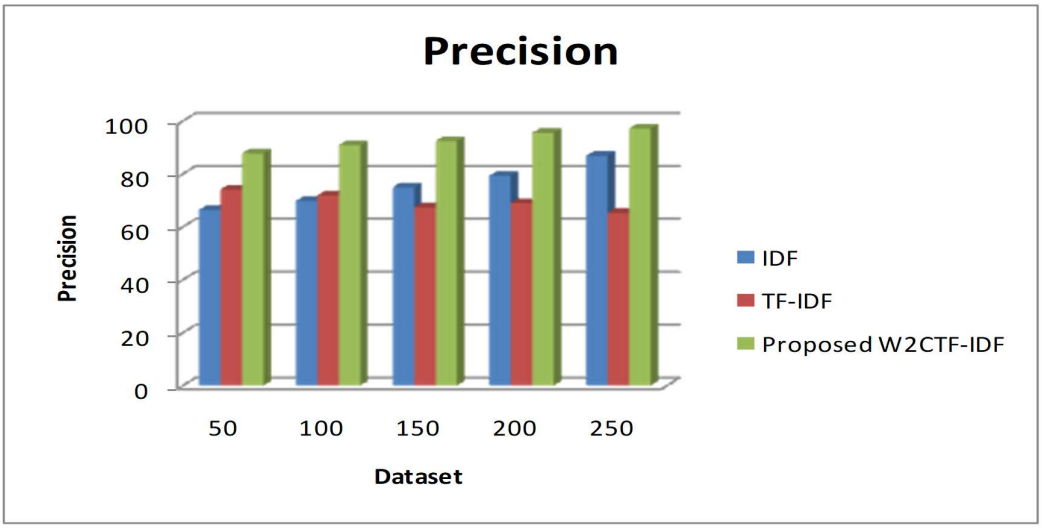

Figure 5 Comparison chart of Precision

The Figure 5 Shows the comparison chart of Precision demonstrates the existing TF-IDF, IDF and proposed W2CTF-IDF. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed W2CTF-IDF values are better than the existing algorithm. The existing algorithm values start from 66.45 to 86.86, 65.33 to 74.12 and proposed W2CTF-IDF values starts from 87.76 to 97.12. The proposed method provides the great results.

Recall

Dataset	IDF	TF-IDF	Proposed W2CTF-IDF
20	0.62	0.72	0.83
40	0.66	0.65	0.87
60	0.70	0.59	0.90
80	0.72	0.62	0.94
100	0.75	0.59	0.96

Table 2. Comparison tale of Recall

The Comparison table 2 of Recall Values explains the different values of existing IDF, TF-IDF and proposed W2CTF-IDF. While comparing the Existing algorithm and proposed W2CTF-IDF provides the better results. The existing algorithm values start from 0.62 to 0.75, 0.59 to 0.72 and proposed W2CTF-IDF values starts from 0.83 to 0.96. The proposed method provides the great results.

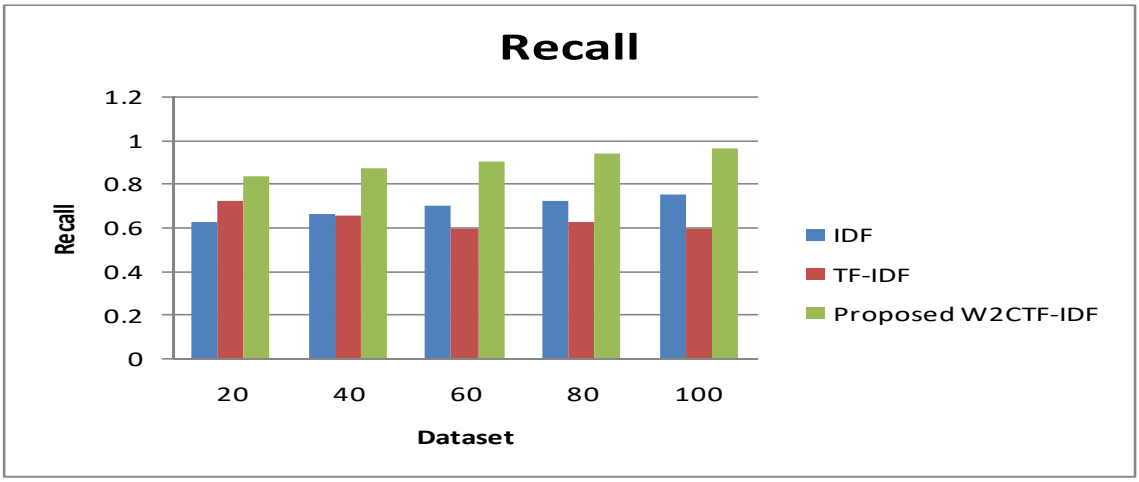

Figure 6 Comparison chart of Recall

The Figure 6 Shows the comparison chart of Recall demonstrates the existing TF-IDF, IDF and proposed W2CTF-IDF. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed W2CTF-IDF values are better than the existing algorithm. The existing algorithm values start from 0.62 to 0.75, 0.59 to 0.72 and proposed W2CTF-IDF values starts from 0.83 to 0.96. The proposed method provides the great results.

F-Measure

Dataset	IDF	TF-IDF	Proposed W2CTF-IDF
100	0.89	0.72	0.98
200	0.85	0.70	0.96
300	0.86	0.67	0.95
400	0.84	0.64	0.93
500	0.82	0.61	0.92

Table 3. Comparison tale of F -Measure

The Comparison table 3 of F -Measure Values explains the different values of existing IDF, TF-IDF and proposed W2CTF-IDF. While comparing the Existing algorithm and proposed W2CTF-IDF, provides the better results. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed W2CTF-IDF values starts from 0.92to 0.98. The proposed method provides the great results.

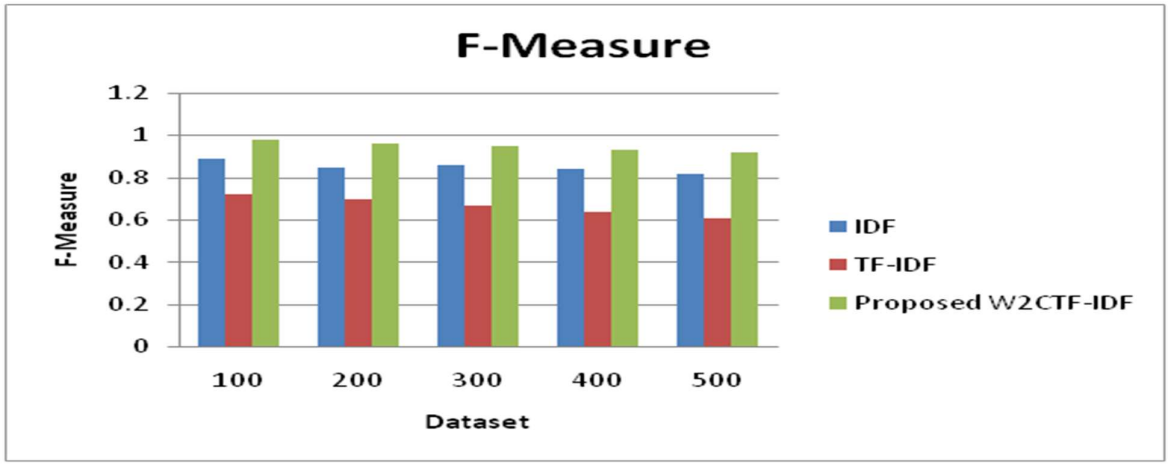

Figure 7 Comparison chart of F -Measure

The Figure 7 Shows the comparison chart of F -Measure demonstrates the existing TF-IDF, IDF and proposed W2CTF-IDF. X axis denote the Dataset and y axis denotes the F -Measure ratio. The proposed W2CTF-IDF values are better than the existing algorithm. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed W2CTF-IDF values starts from 0.92to 0.98. The proposed method provides the great results.

Efficiency

Dataset	IDF	TF-IDF	Proposed W2CTF-IDF
10	69.48	79.54	89.65
20	76.98	82.99	91.38
30	79.71	85.47	93.59
40	81.47	80.67	95.96
50	85.98	89.63	98.01

Table 4. Comparison tale of Efficiency

The Comparison table 4 of Efficiency Values explains the different values of existing IDF, TF-IDF and proposed W2CTF-IDF. While comparing the Existing algorithm and proposed W2CTF-IDF, provides the better results. The existing algorithm values start from 69.48 to 85.98, 79.54 to 89.63 and proposed W2CTF-IDF values starts from 89.65 to 98.01. The proposed method provides the great results.

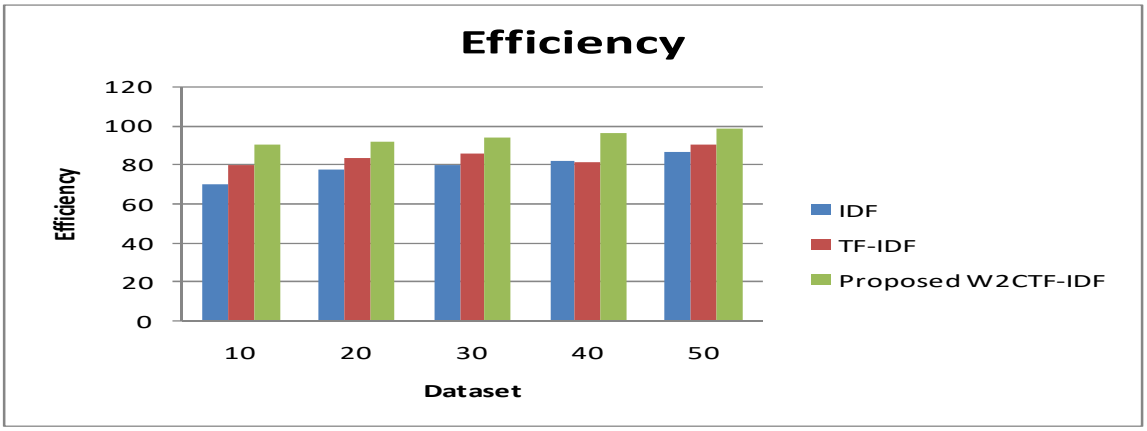

Figure 8 Comparison chart of Efficiency

The Figure 8 Shows the comparison chart of Efficiency demonstrates the existing TF-IDF, IDF and proposed W2CTF-IDF. X axis denote the Dataset and y axis denotes the Efficiency Measure ratio. The proposed W2CTF-IDF values are better than the existing algorithm. The existing algorithm values start from 69.48 to 85.98, 79.54 to 89.63 and proposed W2CTF-IDF values starts from 89.65 to 98.01. The proposed method provides the great results.

5. Conclusion

The present work utilizes feature extraction to extricate the key terms (features) in light of the Reuters-21578 text categorization test assortment. While the TFIDF is utilized to weight the feature in view of the frequency for each term, that is to say, the term in the document is connected with one more term in various classes. Obviously the W2CTF-IDF technique outflanks the F1 measure. Besides, the discoveries uncover that feature weighting techniques influence the effectiveness of financial statement fraud detection. Pushing ahead, the recommended work can be additionally ventured into a bigger framework for text categorization. This permits the incorporation of dimensionality decrease and machine learning draws near. The results showed that the evacuation of stop-words can upgrade the acknowledgment degree among documents and the framework performance for feature extraction.

References

- Bodile and M. Kshirsagar, "Text mining in radiology reports by statistical machine translation approach," 2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 238-241, doi: 10.1109/GCCT.2015.7342797.
- I. Kadhim, Y. Cheah, N. H. Ahamed and L. A. Salman, "Feature extraction for co-occurrence-based cosine similarity score of text documents," 2014 IEEE Student Conference on Research and Development, 2014, pp. 1-4, doi: 10.1109/SCORED.2014.7072954.
- Wang, "Feature Extraction Method of Machine Translation Equivalent Pairs in Chinese-English Comparable Corpus based OCR Recognition," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 899-902, doi: 10.1109/ICOEI51242.2021.9452871.

- H. He, B. Chen, W. Xu and J. Guo, "Short Text Feature Extraction and Clustering for Web Topic Mining," Third International Conference on Semantics, Knowledge and Grid (SKG 2007), 2007, pp. 382-385, doi: 10.1109/SKG.2007.76.
- J. Liu, Y. Ye and Y. Du, "Text Feature Extraction and Clustering Analysis of Events Caused by the Cockpit Crew," 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, 2020, pp. 1018-1022, doi: 10.1109/ICCASIT50869.2020.9368644.
- J. Liu, Y. Ye and Y. Du, "Text Feature Extraction and Clustering Analysis of Events Caused by the Cockpit Crew," 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, 2020, pp. 1018-1022, doi: 10.1109/ICCASIT50869.2020.9368644.
- K. Kawashima, W. Bai and C. Quan, "Text mining and pattern clustering for relation extraction of breast cancer and related genes," 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017, pp. 59-63, doi: 10.1109/SNPD.2017.8022701.
- P. P. Shelke and A. A. Pardeshi, "Review on Candidate Feature Extraction and Categorization for Unstructured Text Document," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 88-92, doi: 10.1109/ICCMC48092.2020.ICCMC-00017.
- Shivaprasad KM and T. H. Reddy, "Text mining: An improvised feature based model approach," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 38-42, doi: 10.1109/ICATCCT.2016.7911962.
- Y. Li, A. Algarni, M. Albathan, Y. Shen and M. A. Bijaksana, "Relevance Feature Discovery for Text Mining," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1656-1669, 1 June 2015, doi: 10.1109/TKDE.2014.2373357.
- Y. Li, Y. Sheng, L. Luan and L. Chen, "A Text Classification Method with an Effective Feature Extraction Based on Category Analysis," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, pp. 95-99, doi: 10.1109/FSKD.2009.304.
- Y. Ye, X. Zhao and X. Zhang, "Web Text Feature Extraction with Bean optimization Algorithm," 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2020, pp. 171-175, doi: 10.1109/ICBAIE49996.2020.00043.
- Yun Yang and Yanan Wu, "The improved features selection for text classification," 2010 2nd International Conference on Computer Engineering and Technology, 2010, pp. V6-268-V6-271, doi: 10.1109/ICCET.2010.5486261.
- Z. Luo, "Extraction Method of Machine Translation Equivalent Pairs in Chinese-English Comparable Corpus based on Text Mining," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1831-1834, doi: 10.1109/ICICCS51141.2021.9432261.

- Barman, Paresh & Iqbal, Nadeem & Lee, Soo-Young. (2006). Non-negative Matrix Factorization Based Text Mining: Feature Extraction and Classification. 703-712. 10.1007/11893257_7.
- S.Sudha & Dr.P.Logeswari (2021). Textual Document Pre-Processing using Hybrid Text Pre-Processing Technique in Text Mining, ISSN: 0011-9342.