IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 8, August 2020

DOI 10.17148/IJARCCE.2020.9805

A Review on Text Analytics Based on Deep Hashing Method

Ankita Kansal[1], Er. Deepak Dudeja[2]

Research Scholar, Department of Computer Science & Engineering, Geeta Engineering College, Panipat, Haryana[1]

Assistant Professor, Department of Computer Science & Engineering, Geeta Engineering College, Panipat, Haryana[2]

Abstract: The analysis of the text content in emails, blogs, tweets, forums and other forms of textual communication constitutes what we call text analytics. Text analytics is applicable to most industries: it can help analyse millions of emails; you can analyse customers' comments and questions in forums; you can perform sentiment analysis using text analytics by measuring positive or negative perceptions of a company, brand, or product. Text Analytics has also been called text mining, and is a subcategory of the Natural Language Processing (NLP) field, which is one of the founding branches of Artificial Intelligence, back in the 1950s, when an interest in understanding text originally developed. Currently Text Analytics is often considered as the next step in Big Data analysis. Text Analytics has a number of subdivisions: Information Extraction, Named Entity Recognition, Semantic Web annotated domain's representation, and many more. This work provides a detailed study on text analytics based on different type of deep learning techniques.

Keywords: Big Data Analysis, Information Extraction, Text Analytics, Deep Learning etc.

I. INTRODUCTION

With regards to TA, Big Data is just a huge volume of composed language information. Be that as it may, where does the boondocks lie between Big Data and Small Data? There has been a culture-evolving reality: while only 15 years prior a content corpus of 150 million words was viewed as gigantic, as of now no under 8.000 million-word datasets are accessible. In addition to the fact that it is an inquiry essentially about size, yet additionally about quality and veracity: information from online networking are loaded with clamor and contortion. All datasets have these issues however they are all the more conceivably genuine for enormous datasets in light of the fact that the PC is a delegate and the human master don't see them straightforwardly, just like the case in little datasets. Subsequently, information purifying procedures expend huge endeavours and frequently after the purging, the accessibility of data to prepare frameworks isn't sufficient to get solid forecasts, as occurred in the Google Flu Trends bombed test [1].

Text mining and text investigation are expansive umbrella terms portraying a scope of advances for dissecting and preparing semi organized and unstructured content information. The bringing together subject behind every one of these innovations is the need to "transform text into numbers" so incredible calculations can be applied to enormous record databases. Changing over content into an organized, numerical arrangement and applying logical calculations require knowing how to both use and consolidate methods for taking care of text, running from singular words to records to whole report databases [2].

Until now, text mining has opposed a progressively complete definition in light of the fact that the field is developing out of a gathering of related yet particular controls, as portrayed in shows the six other significant fields that converge with text mining. Because of the broadness and dissimilarity of the contributing controls, it very well may be troublesome in any event, for text mining specialists to briefly describe. Text mining is something of the "Wild West" of examination, since there are various contending innovations with no unmistakable predominance among them. To additionally entangle matters, various regions of text mining are in various phases of development [3].

The main objective in this part is to carry clearness to the field by giving a system and jargon to talking about the seven diverse practice territories inside content mining. Because of the broadness of text mining, no single book can plan to completely cover the field. Our intended interest group is no master text-mining professionals' investigators who have the specialized skill to deal with difficulties including text yet have restricted understanding or foundation with text handling. Thusly, this book gives a prologue to every one of the seven practice zones, however it covers inside and out just those regions that are available for non experts, yet not omnipresent.

Text Analytics is gaining prominence in many industries, from marketing to finance, because the process of extracting and analysing large quantities of text can help decision-makers to understand market dynamics, predict outcomes and trends, detect fraud and manage risk. The multidisciplinary nature of Text Analytics is key to understand the complex integration of different expertise: computer engineers, linguists, experts in Law, Biomedicine or Finance, data scientists, psychologists, causing that the research and development approach is fragmented due to different traditions,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 8, August 2020
DOI 10.17148/IJARCCE.2020.9805

methodologies and interests. A typical text analytics application consists of the following steps and tasks: Starting with a collection of documents, a text mining tool retrieves a particular document and pre-process it by checking format and character sets [4].

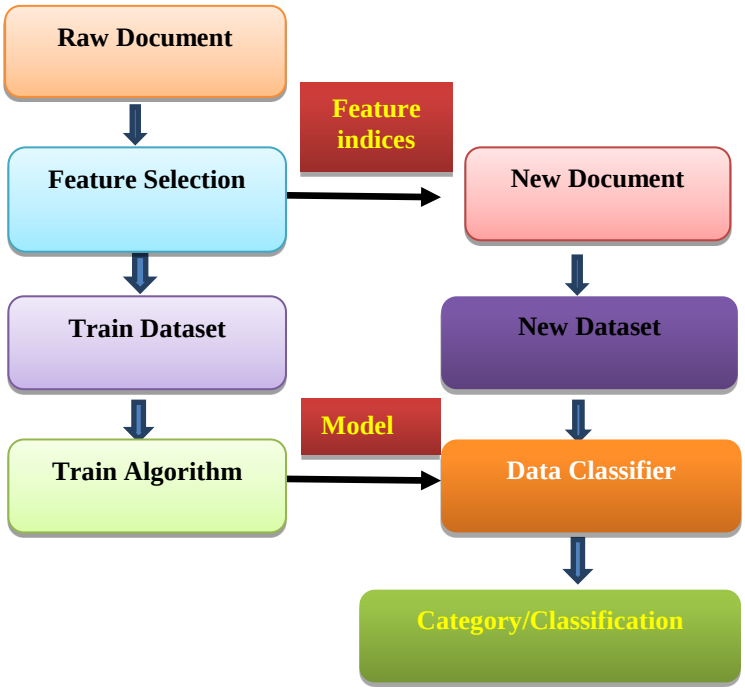

Fig 1: Text classification

Information extraction (IE) software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process usually called pattern matching, typically based on regular expressions. The most popular form of IE is named entity recognition. ER seeks to locate and classify atomic elements in text into predefined categories. ER techniques extract features such as the names of persons, organizations, locations, temporal or spatial expressions, quantities, monetary values, stock values, percentages, gene or protein names, etc. These are several tools relevant for this task: Apache Open NLP, Stanford Named Entity Recognizer, Ling Pipe.
The paper is ordered as follows. In section II, it represents related work in text analytics. Section III presents the text analytics using CNN. Finally, conclusion is explained in Section IV.

II. LITERATURE SURVEY

This section provides the work related to various authors in different fields related to text analytics using different techniques. Table 1 shows the summary of literature survey as shown below:

Table 1: Summary of Literature Survey

Authors	Year	Reviewed Area
N.K. Sawant et al. [13]	2018	Devanagari Printed Text to Speech Conversion using OCR
F. Wei et al. [14]	2018	Empirical Study of Deep Learning for Text Classification in Legal Document Review
F.F. Shahareet al. [10]	2017	Sentiment Analysis for the News Data Based on the social Media
A. Hennig et al. [7]	2017	Big Social Data Analytics of Changes in Consumer Behaviour and Opinion of a TV Broadcaster

L. Bradel et al. [2014] described that Semantic connection offered a natural correspondence instrument between human clients and complex measurable models. By protecting the clients from controlling model boundaries, they centre rather around straightforwardly controlling the specialization, along these lines staying n their subjective zone. n any case, this method isn't inalienably adaptable past many content records. To cure this, t presented the idea of multi-model semantic collaboration, where semantic associations can be utilized to direct different models at numerous degrees of information

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 8, August 2020
DOI 10.17148/IJARCCE.2020.9805

scale, empowering clients to handle bigger information issues. We additionally present a refreshed perception pipeline model for summed up multi-model semantic collaboration. To exhibit multi-model semantic connection, t presented Star SPIRE, a visual book examination model that changes client cooperation's on archives into both little scope show format refreshes just as enormous scope importance-based record choice [7].

N. Medoc et al.[2014] proposed a Visual Analytics device that bolsters circumstances mind-fullness and investigation assignments for text streams. To arrive at this objective, t planned our own information model to encode spilling text n different powerful recurrence networks, taking care of various parts of information. ts perceptions are made first out of two unique Theme Rivers. They permit constant investigation of the considerable number of viewpoints extricated from messages put away n both, presented moment and long-haul cradles. Likewise, pictured the topographical area of messages on a guide. t utilized these perceptions, upgraded by productive client cooperation components, to respond to the inquiries of the third smaller than usual test of 2014 VAST (Visual Analytics Science & Technology) Challenge [8] .

J. Park et al. [2014] presented that self-created java-based visual explanatory apparatus peruses a wide range of text information sources and concentrates significant catchphrases, relations and occasions from them utilizing philosophy and characteristic language preparing strategies. At last, it gives a coordinated and intelligent inquiry interface to clients to encourage their viable and proficient examination for the huge and complex informational collection [9].

A. Salinca et al. [2015] described that the exploration territory of assumption investigation, supposition mining, feeling mining and slant extraction has picked up ubiquity n the most recent years. Online audits are turning out to be significant models n estimating the nature of a business. This paper presents an estimation examination way to deal with business surveys arrangement utilizing an enormous audits dataset gave by Yelp: Yelp Challenge dataset. n this work, we propose a few methodologies for programmed assumption order, utilizing two element extraction strategies and four AI models. it is outlined a near report on the adequacy of the outfit techniques for audits opinion grouping [10].

H S, Chiranjeevi et al. [2016] presented that Digital world was coming, were information as become huge information with ever increment n enormous volume of computerized data accessible as far as text archives. This tends for information extraction, advancement, examination and recovery of text records which were as unstructured nature turns into a significant issue in internet searcher. Generally, text records were the wellspring of putting away our data; either close to home or expert. it was additionally significant for associations including private and open which have been gathering huge volume of space explicit content report data, which may contain national nsight, instruction, clinical data, business and showcasing. n this paper it presented a framework that enhances the data recovery procedure of text archives n internet searcher from unstructured information [11].

A. Hennig et al. [2016] presented that the adjustments n purchaser conduct and conclusions because of the progress from an open to a business supporter with regards to broadcasting worldwide media occasions. By examining TV watcher appraisals, Facebook movement and ts estimation, had planned for sentiment data. It utilized content grouping and visual examination techniques on the business and social datasets. Our primary finding s a reasonable connection between negative supposition and advertisements. n spite of positive change n client conduct, provided a negative impact on customer. n view of media occasions and telecaster speculations, it distinguished generalisable discoveries for every single such progress [12].

R.B. Mbah et al. [2017] presented out work on gathering, breaking down and picturing neighbourhood work information utilizing text mining strategies. We additionally talk about advances utilized, for example, corn occupations for robotization; Java for API information assortment and web rejecting, Elastic search for information sub-setting and watchword examination, and R for information investigation and representation. We anticipate that this work should be of pertinence to an assorted scope of occupation searchers just as managers and instructive establishments [13].

K. S. Sabra et al. [2017] presented that Sentiment Analysis s the way toward recognizing slant from text written n a characteristic language concerning the substance t s alluding to. Feeling dictionaries are utilized to play out this errand. A few dictionaries are accessible to play out this undertaking n English utilizing WordNet. In this paper, it presented another strategy to make a notion vocabulary for the database n Arabic language [14].

F.F. Shahare et al. [2017] described that social Data are increments extremely quick, in each region social information assume a significant job n each edge, internet-based life large information mining territory invited by analysts. A figuring assessment of news information was a critical segment of the online networking large information. The figuring conclusion of news data might be a central point of the internet-based life enormous data. In current web word scope of client utilize internet-based life and interpersonal organization to peruse and peruse news associated data. Ordinary scope of issue territory unit happening and internet-based life impacted the news related with this news [15].

P. Das et al. [2018] presented that Contingent on the varieties of information, enormous information comprises social Data, machine information and exchange-based Data. Social information gathered from Facebook, Twitter and so on. Machine information are RFID chip perusing, GPRS and so on. Exchange based information incorporates retail site's information. Around the varieties of various sorts of information significant part is text information. Text information is organized information. Determining of top-notch organized information from unstructured content is text investigation. Changing over unstructured information into significant information was text examination process. CV parser coordinate up-and-comer's resume with enlistment work process and consequently forms approaching CV's.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 8, August 2020
DOI 10.17148/IJARCCE.2020.9805

M.Maia et al.[2018] presented self-created java-based visual explanatory apparatus peruses a wide range of text information sources and concentrates significant catch phrases, relation and occasions from them utilizing philosophy and characteristic language preparing strategies. At last, it gives a coordinated and intelligent inquiry interface to clients to encourage their viable and proficient examination for the huge and complex informational collection [17].
N. K. Sawant et al. [2018] presented Devanagari text to discourse transformation s accomplished for Marathi printed text. To get the necessary yield the two strategies are executed that are Optical Character Recognition (OCR) and Text to Speech (TTS) framework. OCR s used to change over the content from a picture into editable book which s finished utilizing multiclass Support Vector Machine (SVM) and Text to Speech framework gives the sound yield [18].
F. Wei et al. [2018] presented that Predictive coding has been broadly utilized in legitimate issues to discover pertinent or advantaged records n enormous arrangements of electronically put away data. It spares the time and cost fundamentally. Calculated Regression and Support Vector Machines (SVM) are two famous AI calculations utilized n prescient coding. As of late, profound learning got a ton of considerations n numerous ventures. This paper reports our primer investigations in utilizing profound learning n authoritative archive survey. in particular, it led investigations to contrast profound learning results and results got utilizing a SVM calculation on the four datasets of genuine legitimate issues. Our outcomes indicated that CNN performed better with bigger volume of preparing dataset and ought to be a fit strategy n the content order n legitimate industry [19].
T. Zhang et al. [2018] presented that Text examination has been generally utilized in various areas to find significant information covered up inside a particular book. Regarding power dispatching, a manual consistently contains a lot of unstructured information, which makes it an extreme activity for dispatchers to recollect and comprehend that data. This paper tends to the above issues by receiving text investigation. n view of the possibility of Natural Language Processing, a progression of key advances are received to do the content dissecting occupation, for example, information structure change, proficient word division devices for Chinese and Word2Vec computation, which are useful for dispatchers to manage the dispatching manual [20].
K. Zvarevashe et al. [2018] described that social Data are increments extremely quick, in each region social information assume a significant job in each edge, internet-based life large information mining territory invited by analysts. A figuring assessment of news information was a critical segment of the online networking large information. The figuring conclusion of news data might be a central point of the internet-based life enormous data. In current web word scope of client utilize internet-based life and interpersonal organization to peruse and peruse news associated data [21].
G. Xu et al. [2019] presented that The strategy for text supposition investigation dependent on conclusion word reference frequently has the issues that the slant word reference doesn't contain enough estimation words or overlooks some field assessment words. In this work, an all-inclusive estimation word reference was developed. The all-encompassing slant word reference contains the essential opinion words, the field supposition words, and the polysemic assumption words, which improves the exactness of notion examination. In this manner, the estimation of the polysemic opinion word n the field s acquired. By using the all-encompassing notion word reference and the structured conclusion score governs, the estimation of the content is accomplished. The test results demonstrated that the proposed assumption examination strategy dependent on expanded assessment word reference has certain achievability and precision. The exploration was important for the conclusion acknowledgment of the remark messages [22].
R. Saidi et al. (2019) introduced numerous strategies n the writing, for example, association, crossing point, and altered association. The association and the intersection methodologies can lead once in a while to build the all out number of highlights and lose some significant highlights. In this work, it presented an element choice strategy that consolidates the Genetic Algorithm (GA) and Pearson Correlation Coefficient (PCC). The exploratory outcomes demonstrate that the proposed strategy can be appropriate to improve the exhibition of highlight choice [23].
Y. Peng et al. (2019) proposed a technique for joining shading symbolism with slant investigation to consequently change over everyday portrayals into shading palettes. The calculation involved four stages. In the first place, it characterized influence words as the reason for content grouping. In this examination influence words are CIS picture words. Second, it gathered significant content corpora from Google and Wikipedia. Third, by means of model preparing, it connected word2vec (a word-inserting model) to ascertain the lexical fondness of effect words and hues [24].

III. TEXT ANALYTICS USING CNN

Text analysis is basically the automation of the analysis of a given text in order to determine the feelings conveyed in it. Sentiment analysis and opinion mining have become known as interchangeable terms. sentiment analysis intends to define the feelings of the writer regarding a particular topic based on the writer's opinion. Text analysis is important as it can help to provide insight into different fields.
Text Analysis is a field that is growing fairly rapidly. 81 percent of Internet users (or 60 percent of Americans) have done online research on a product at least once, meaning every year there are more articles targeting different text domains over years, where the reviews represent around the 49.12% of the articles. One would not always want to apply text to product reviews; there are too many other fields. One good example of this that has been experimented is the comparison of Twitter text versus Gallup polls of consumer confidence. The results yielded were positive

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 8, August 2020
DOI 10.17148/IJARCCE.2020.9805

and the correlation was 0.804, inferring that we can use Twitter to measure public opinion. This is precisely what we are going to use Twitter for during this work.
It proposes a method to obtain labelled text data from a social media stream together with a feature representation suitable for dealing with text drifts. Their approach follows the distant supervision paradigm, but instead of relying on emoticons or other text-based text clues, it exploits social behaviour patterns usually observed in online social networks. These patterns are referred to as self-report imbalances and are
defined as follows:
1. Positive-negative text report imbalance: users tend to express positive feelings more frequently than negative ones.
2. Extreme-average text report imbalance: users tend to express extreme feelings more than average feelings.
The data it uses for model experiment is obtained using Google site as the source collection, and a sample of 5000 tweets as the target collection. It pre-process the given corpus before calculating the features. All the tweets are lowercased, tokenized and tagged. The first feature is a nominal attribute corresponding to the POS tag of the word in its context. This feature provides morphological information of the word. There is empirical evidence that subjective and objective texts have different distributions of tags. It visualizes the expanded lexicon intensities of words classified as positive and negative through word clouds.

1. Texts Using Convolutional Neural Networks
Sentence Model
Each word is represented by its d-dimensional word embedding. A sentence (or tweet) is represented by a concatenation of the representations of its n constituent words. This yields a matrix X, which is used as input to the convolutional neural network.

Convolutional Layer
In this layer, a set of m filters is applied to a sliding window of length h over each sentence. Let X denotes the concatenation matrix with features c and filter F:

$$c = \sum X.F \qquad (1)$$

The vectors c are then aggregated from all m filters into a feature map matrix C.

Max Pooling
The output of the convolutional layer is passed through a non-linear activation function relu (x) = max(0, x), before entering a pooling layer. The latter aggregates vector elements by taking the maximum over a fixed set of non-overlapping intervals.

Hidden Layer
A fully connected hidden layer computes the transformation. The output $x \in R^m$ of this layer is the vector of the sentence embeddings for each tweet.

Softmax
Finally, the outputs of the previous layer $x \in R^m$ are fully connected to a soft-max regression layer, which returns the class $y \in [1,K]$ with largest probability.

Network Parameters
Training the neural network consists of learning the set of parameters with X the word embedding matrix, where each row contains the d-dimensional embedding vector for a specific word.

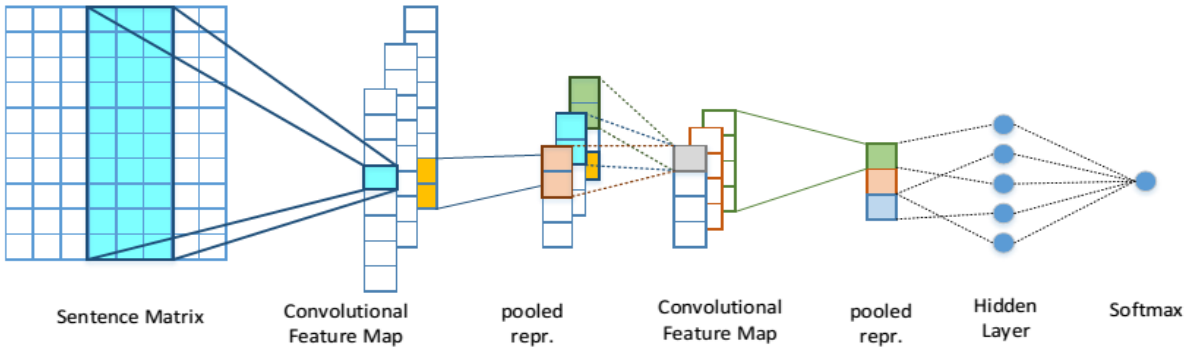

Fig 2: Approach using CNN

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 8, August 2020
DOI 10.17148/IJARCCE.2020.9805

IV. CONCLUSION

Text Analytics is likewise a prescient examination technique. At the point when the preparation informational collections or writings comes, client arranged the writings into various bits or writings for characterization. Text investigation is otherwise called Text Mining. Information mining is extraction of significant important data from enormous measure of information gathered from information distribution center. This work gives a far reaching concentrate on text investigation dependent on various procedures introduced by different creators in their field. The main objective of this work is to provide text analytics using hashing method and compare the performance with conventional deep learning methods.

REFERENCES

[1]. Yi J., Nasukawa T., Bunescu R. and Niblack W., 2003, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques", Third EEE International Conference, pp. 427-434.
[2]. Wang J, ZhaoY, 2010, "Effective feature selection with particle swarm optimization based one dimensional searching", IEEE, pp.4244-6044.
[3]. Bin L., Min Y., 2012, "Analysis Model of Drilling Tool Failure Based on PSO-SVM and its Application, international Conference on Computational and information Sciences", pp. 978-981.
[4]. Liam C, Xue B., 2012, "Binary Particle Swarm Optimization for Feature Selection: A Filter Based Approach, EEE Congress on Evolutionary Computation", pp. 1-8.
[5]. Xue Z, Du P, 2013, "A Novel Classification Technique For Hyper spectral Imagery Based on Harmonic Analysis, SVM & PSO", EEE, pp.1-4.
[6]. K. Atasu & R. Polig ,2013, "Hardware-Accelerated Regular Expression Matching for High-Throughput Text Analytics", EEE, pp.01-07.
[7]. L. Bradel & C. North,2014, "Multi-Model Semantic interaction for Text Analytics", IEEE, pp. 163-172.
[8]. N. Medoc & M. Stefas, 2014, "Visual Analytics of Text Streams Through Multiple Dynamic Frequency Matrices" IEEE,pp.381-382.
[9]. J. Park, 2014, "Integrated Visual Analytics Tool for Heterogeneous Text Data", EEE, pp.325-326.
[10]. A. Salinca,2015, "Business reviews classification using sentiment analysis", IEEE, pp.247-250.
[11]. Chiranjeevi H S& M. Shenoy K ,2016, "DSSM with Text Hashing Technique for Text Document Retrieval n Next-Generation Search Engine for Big Data and Data Analytics", EEE, pp.01-05.
[12]. A.Hennig & A-S Amodt, 2016, "Social Data Analytics of Changes n Consumer Behaviour & Opinion of a TV Broadcaster",IEEE,pp.3839-3848.
[13]. R.B. Mbah & M. Rege, 2017,"Discovering Job Market Trends with Text Analytics",IEEE,pp.137-142.
[14]. K. S. Sabra & R. N. Zantout, 2017, "Sentiment Analysis: Arabic Sentiment Lexicons",IEEE, pp.01-04.
[15]. F.F. Shahare, 2017, "Sentiment Analysis for the News Data Based on the social Media",IEEE, pp.1365-1370.
[16]. P. Das & B. Sahoo,2018, "A Review on Text Analytics Process with a CV Parser Model",IEEE, pp.01-07.
[17]. M. Maia & A. Freitas, 2018, FinSSLx: "A Sentiment Analysis Model for the Financial Domain Using Text Simplification",IEEE, pp.318-319.
[18]. N.K. Sawant & S. Borkar,2018, "Devanagari Printed Text to Speech Conversion using OCR",IEEE, pp.504-507.
[19]. F. Wei & H. Qin,2018, "Empirical Study of Deep Learning for Text Classification n Legal Document Review", EEE, pp.3317-3320.
[20]. T. Zhang & J. Lu, 2018, "The Application of Text Analytics n Electric Power Dispatching", EEE, pp.4186-4189.
[21]. K. Zvarevashe, O. Olugbara, 2018, "A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews", Conference on nformation Communications Technology and Society, pp. 01-04.
[22]. G. Xu, Z. Yu, 2019, "Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary", IEEE,pp. 0-14.
[23]. Saidi R., Bouaguel W, 2019, "Hybrid Feature Selection Method Based on the Genetic Algorithm and Pearson Correlation Coefficient", Springer Nature Switzerland.
[24]. Peng Y., Chou T., 2019, "Automatic Color Palette Design Using Color image and Sentiment Analysis", EEE 4th International Conference on Cloud Computing and Big Data Analytics, pp. 389-392.