iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6026

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

An Analysis of Short Text Detection and Classification Algorithms

Parvathi P[1], Dhanya C. K[*2]
[1, 2]*Dept. of Information Technology, Government Engineering College, Barton Hill, Kerala, India*

Abstract: ***In recent years, there has been an exponential growth within the number of complex documents and texts. It requires a deeper understanding of machine learning methods to be ready to accurately classify texts in many applications. Understanding the rapidly growing short text is extremely important. Short text is different from traditional documents in its length. With the recent explosive growth of e-commerce and online communication, a replacement genre of text, short text, has been extensively applied in many areas. Numerous researches specialise in short text mining. It's a challenge to classify the short text due to its natural characters, like sparseness, large-scale, immediacy, non-standardization etc. With the rapid development of the web, Web users and Web service are generating more and more short text, including tweets, search snippets, product reviews then on. There's an urgent demand to know the short text. For instance an honest understanding of tweets can help advertisers put relevant advertisements along the tweets, which makes revenue without hurting user experience Short text classification is one among important tasks in tongue Processing (NLP). Unlike paragraphs or documents, short texts are more ambiguous .They do not have enough contextual information, which poses challenge for classification. We retrieve knowledge from external knowledge source to reinforce the semantic representation of short texts. We take conceptual information as a sort of data and incorporate it into deep neural networks. Here we are going to study different methods available for text classification and categorisation.***

Keywords: Short-text classification, word embedding, pre-processing, tokenization, categorization, text detection

I. INTRODUCTION

Global internet usage is increasing because of social networking services and advance in e-commerce. Users are able to provide feedbacks, reviews etc. easily. Classification and categorisation of data and comments from customer reviews is the challenging task for business purpose. We have to effectively classify text in less computational time and more accuracy. Vocabulary coverage rate on word embeddings by matching uncovered words with most similar words is to be done.

Short text has been widely used in many fields, such as mobile short message, instant message, BBS title, news title, online chat record, blog comment, news comment, etc. Its main characteristic is the text length .It is very short, no longer than 200 characters. The mobile short message is of less than 70 characters, BBS title and news title is less than 30.Instant messaging (IM) software supports longer message. For sending message quickly and ensuring it safely, IM software also limits its length, such as Windows Live Messenger of Microsoft allows the longest message of 400 characters. Short texts are live messages which are usually received in real time. Real time data are very large in volume as compared to normal text. Therefore, the volume of short text is also high. A short text contains a few words. It does not provide enough words co-occurrence details. Valid language feature extraction is very difficult. The application background has massive amounts of short textual data. It focus on only a small part among the large-scale data. Therefore, useful instances are limited, and the distribution of short text is imbalanced. It is difficult to manually label. Text recognition involves localization of images and reading characters in images. Here we are dealing with text detection from static image and from videos. Short text is not only detected from reviews, comments etc but also to be detected from scenes, movies, news etc .Thus text detection and classification plays an important role in business field.

II. SIMPLE ALGORITHMS AND METHODS IN SHORT TEXT DETECTION AND CLASSIFICATION

A. Different Algorithms for Text Detection (From Static and Dynamic Images)

1) *GISCA:* For multi oriented text, GISCA serves as an end to end trainable network. It is a fully convolutional network. It has four parts. They are GIM, CAM, Pixel level prediction network and post processing module.GIM is the backbone network. It generates feature layers [1].CAM is for extracting the features and to refine it. It removes irrelevant background and highlight the text regions. The output of CAM has detailed description and high level semantic description .Pixel level prediction module is to produce output. Its output is in the form of segmentation map. Post processing module is used to create bounding box for this. For feature extraction VGG16 network is used.

2) *Strokelets:* It represents multiscale representation of characters. It identifies the information about the substructure of characters. With the help of Strokelets, text in different fonts, styles, colors, multiscale text etc can be identified .Idea of Strokelets can be applied to different languages, different texts, etc. We can use character level bounding box for automatic learning [2].Character identification and detection is easier with Strokelets. For character detection, Strokelets identification and Hough map voting is used .Using this technique we can identify characters accurately. Given a set of training images S = {(I, B)}, where I is an image and B is a set of bounding boxes. It specifies the location of the characters in the image I .The Strokelets is to be generated. This is performed from the training set S. Then the prototype is learned.SVM and Random forest are used for classifying these text. Clustering and training of these are done frequently to produce the output. It contains a CCR extraction region.CCR means candidate character region .It means the required text region. It uses Stroke feature transform algorithm. It is used to detect edges in the input image. It then generates a stroke width map and stroke color map. These are generated by pixel wise comparison. This is one of the best method for text detection .But it is very time consuming and highly sensitive to noise.

3) *Superpixel Based Stroke Feature Transform:* The SSFT has 3 steps. This includes Superpixel Segmentation and clustering, background region removal and region refinement. The input image is first resized to a fixed height and a width that preserves the original aspect ratio , and then smoothed .And for smoothening the image ,we use edge preserving filter . Next, it is over segmented into K superpixels .It is done using the simple linear iterative clustering (SLIC) algorithm. It clusters pixels in the combined color and image plane space. Thus uniform superpixels are generated. The low resolution and low contrast images are there. It has under segmented characters. Here we fix the number of superpixels per image. Super pixel size varies for each image. Color descriptor is used. Here, the color descriptor is defined as S = (r, g, b, l, a, b, h, s, v) where (r, g, b), (l, a, b) and (h, s, v) are the means of the colors of the pixels in S in the RGB, CIE Lab and HSV color spaces. Each component of f is linearly normalized into the interval [0, 1].A distance matrix is constructed by computing the Euclidean distance C. Based on this distance matrix, the superpixels are clustered .It is done by using the average linkage hierarchical clustering algorithm. The clustering process is much faster for superpixels than ordinary pixels. The number of classes is automatically determined by grouping all superpixels of similar colors into clusters. Here, when the inconsistency between the classes is below a predetermined threshold then hierarchical clustering procedure stops. Thus the number of clusters differs for each image. Superpixels within the same class have multiple subregions [3]. Overall, the superpixel clusters partition the original image into different regions. Different colors correspond to different regions. A fast edge detection algorithm is used .It predicts local edges by applying structured random decision forests. It is used to directly extract an edge probability map and a gradient orientation map. , Edge probability map value associated with each pixel represents its probability of being be an edge point The edge image E of I is constructed by applying non-maximum suppression to the EP Next, the distance map is constructed by applying a distance transformation on the edge image E. This is done based on the Euclidean distance, ts. The computing element of SSFT is a superpixel and SFT is a pixel. The searching process is on the contour pixels in SSFT. SFT searches on edgepixels. The color comparison in SSFT is performed based on the superpixel's mean color and in SFT it is based on each individual pixel's color. Deep Learning Based Region Classification is used for feature extraction. Geometric features and deep features are extracted using DLRC algorithm. The geometric features include the number of pixels in candidate region ,the area of the candidate region's boundary box, the width and height of the candidate region's boundary box, the width and height of the input image , the number of pixels of the CCR, the mean and variance of the stroke widths of the pixels in the CCR etc. Deep features include global and contextual information of text.

4) Real Time Lexicon Free Scene Text Localisation and Recognition: The efficient sequential selection is performed from a set of extremal region. In the first stage, character candidates are detected as Extremal Regions. It is selected in a two-stage classification process. This is performed on a coarse Gaussian scale space pyramid and on multiple image projections. Extremal Region has its outer boundary pixel a higher values compared to the region itself[4] .We uses RGB and HSI color spaces ..If a character has multiple elements, the elements are combined into a single region .The combining operation is an element-wise addition. Its function is to align the vectors so that the elements correspond to same rows. The computation complexity is constant. Each area is processed separately over a coarse Gaussian pyramid and ERs are detected. Only distinctive ERs which correspond to characters are selected by a sequential classifier. This is to reduce the high false positive rate and the high redundancy of the ER detector. It has two stages .In the first stage, the probability of each ER being a character is estimated by a threshold. This threshold is increased and then incrementally computable descriptors are computed. Features are calculated by a novel algorithm .Only ERs with locally maximal probability are selected In the second stage, the ERs are classified into character and non-character classes .And finally most probable text is selected.

B. Different Algorithms For Text Classification

1) *Bag of Words (BoW):* BoW is an algorithm used in natural language processing. It can be also used for information retrieval. In this model without considering the grammar or the order of the words, text or sentences are represented in a bag. One of the drawbacks with this BoW approach is that it will suffer from the Sparsity of data, if short text is represented by a BoW model. For classification and categorisation a huge volume of data is needed but this method will not capture sufficient contextual information. Bow model discards the semantic relation between words as well as it does not give any importance to the order of words. This model is simple to understand as compared to other models.
2) *Word embedding (US):* WE [5] for short text classification that help enrich semantic relations. Word embeddings represent a set of methods. Here each and every word are checked. In the training document corresponding to each word, a vector representation is present. The words having similar meaning will have a closed semantic vector representation. Therefore word embeddings can assess the semantic similarity between two words.
3) *K-Nearest Neighbour:* The k-nearest Neighbours algorithm (KNN) is a non-parametric technique. It is used for classification. This method is used for text classification applications .Given a test document this algorithm finds the k nearest neighbours [6] of x among all the documents in the training set. The similarity of x and each neighbour's document is the score of the category of the neighbour documents. Multiple KNN documents belong to the same category; If this case occurs, the summation of these scores is defined as the similarity score of class k with respect to the test document x. Black colors belong to one class, red color circles belongs to another class ,blue color triangles form another class .In the figure three different classes are shown. In each class score of each member is calculated .Score of the document or text shows whether it belongs to the same class or different class.

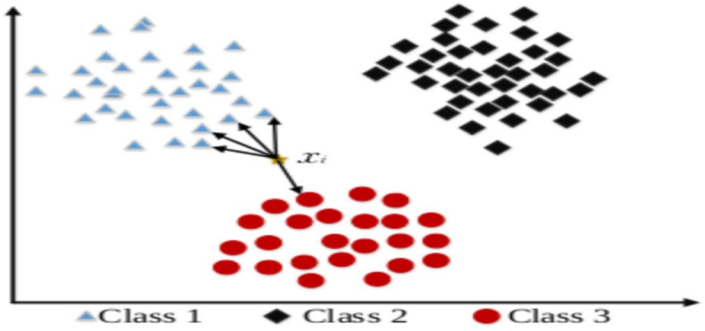


Fig 1. K-nearest Neighbour KNN model

4) *Support Vector Machine (SVM):* The SVM is mostly used in classification problems. It is a supervised learning algorithm. The SVM [7]-[9] was originally designed for binary classification tasks. Work on multi-class problems can also be performed using this.SVM performs non linear classification using kernal tricks. A support-vector machine constructs a hyperplane. It can have more than one hyperplane also. Here we use hyperplane to classify different data objects in an n dimensional plane. A good separation is achieved by the hyperplane .It shows the largest distance to the nearest training-data point of any class .If margin becomes larger error rate will be low. Below figure indicates the linear and non-linear classifier which is used for two dimension datasets. The red color is class 1, the blue color is class 2 and yellow color is miss-classified data points.SVM are used in text categorization and classification of images. There are linear SVM classification and non linear SVM classification. samples on margin are called support vectors. The regions bounded by the hyperplane are called margins.

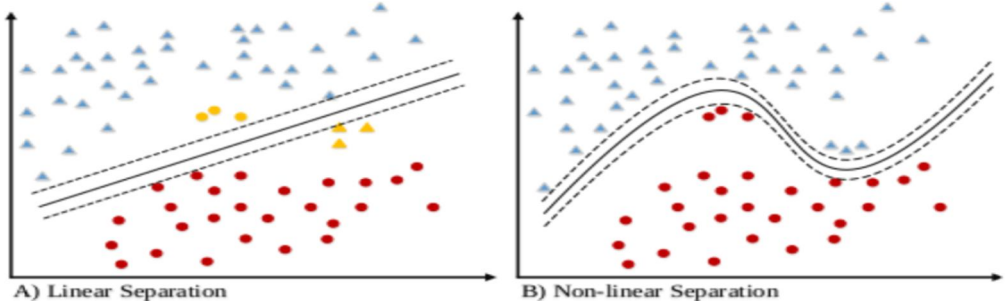


Fig 2: Linear and Non-Linear SVM

5) *Decision Tree:* This is the earliest classification algorithm for text and data mining. It is a decision support tool. Decision tree classifiers (DTCs) [10]-[13] are widely used for classification. It is also used in operational research area. It uses a tree representation to represent data and to achieve the goal in decision making. It consists of leaf nodes and internal nodes. The leaf nodes represent the outputs or the decisions and the internal nodes represent the attributes. The hierarchical decomposition of the data space shows the structure of this technique .The main idea is creating a tree based on the attribute for categorized data points. The main challenge of a decision tree is which attribute or feature could be in parents' level and which one should be in child level. To solve this problem, statistical modelling for feature selection is used in the tree. This is the only method by which we can represent control statements. It has a flow chart like structure. It has the decision rule as its path from root to leaf. Decision trees can also be draw using flow chart symbols. It is a very easy model and people can understand and interrupt easily. But is also considered as unstable model because small change in a data alters the structure of the tree and the entire result will change. Calculation will get complex when the tree structure becomes large. Decision trees can be combined to form random forest but the structure will get complicated and thus error in prediction occurs at a larger rate.

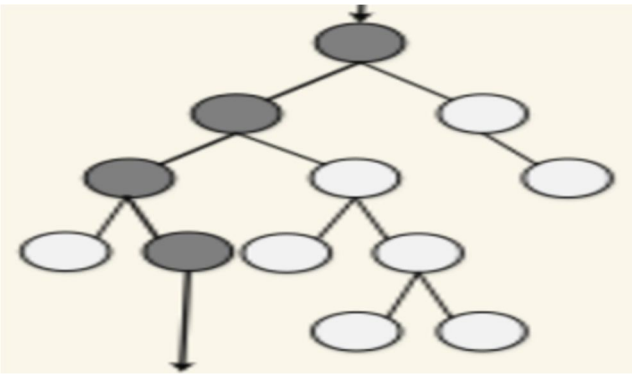


Fig 3. Decision Tree

6) *Random Forest:* Random forest are also known as random decision forest. It is an ensemble learning method which can be used for text classification. It can run on large data sets. It doesn't overfit. It can handle thousands of input values and have methods for handling error. This method constructs t tree as parallel i.e, constructing a multitude of decision trees at training time. The main idea of RF is generating random decision trees. We can run many trees and it does not overfit. After training all trees as forest, predictions are assigned based on voting. Random forest [14] are very fast to train for text data sets but slow in making predictions. For faster structure, the number of trees in forest must be reduced. As number of trees increases, time for prediction decreases. Generated random forest can be stored for later usage. In random forest, many trees are created. Every decision tree has its own observation. The most common observation is the final output. Out of error is one of the methods for error estimation. Errors are estimated in percentage.

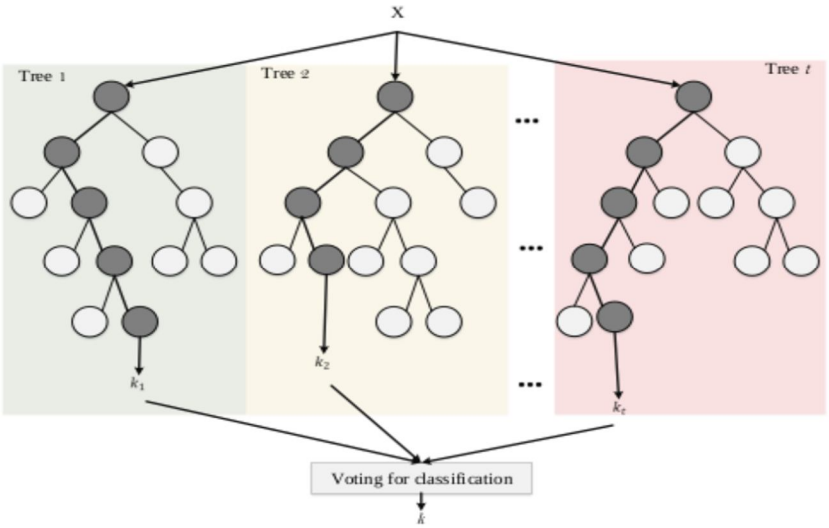


Fig 4. Random Forest

7) *Deep Neural Networks:* Deep neural networks is a network with more than two layers. It is designed to learn by multi-connection of layers .It is widely used in mathematical modelling to process complex data. Every single layer receives the connection from previous layer .It provides connections only to the next layer in a hidden part .It can have multiple hidden layers. Each layer performs sorting and ordering of data. The input denotes the connection of the input to the first hidden layer of the DNN. The input layer is constructed by using TF-IDF, word embedding, etc.The output layer is equal to the number of classes for multi-class classification. Output will be one layer for binary classification. The implementation of DNN [15], [16] is a discriminative trained model. It uses a standard back-propagation algorithm .It has sigmoid as and activation function. The output layer should be a Softmax function for multi-class classification. Deep learning models produce better results than machine learning models. The widely used algorithm for training deep neural networks is back propagation. The prediction accuracy of the network depends on weight and biases used. The difference in generated output and original output is defined as the loss function of the network. Deep neural networks are used in image recognition, object recognition, text processing etc.

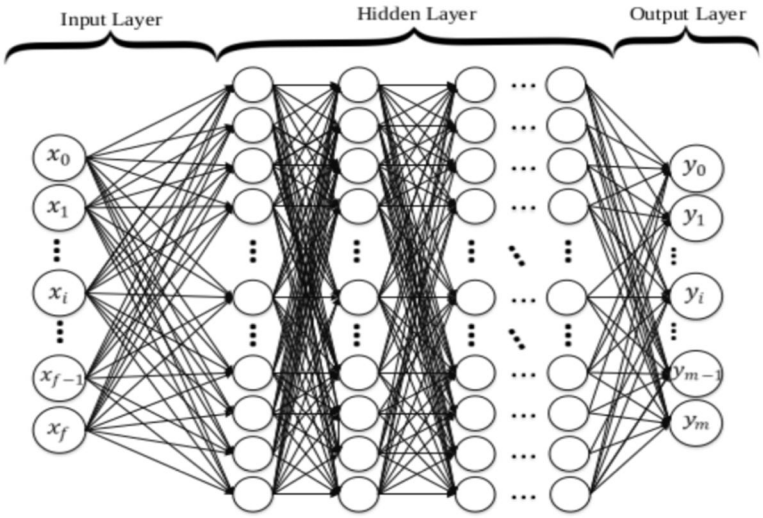


Fig 5. Fully-connected Deep Neural Network (DNN)

8) *Long Short Term Memory:* LSTM is a recurrent neural network which has feedback connections. It can process a single data point as well as a sequence of data. LSTM [6] consist of several gates such as input gate, output gate and forget gate. The gate can store the values at arbitrary time intervals. This is useful for solving the vanishing gradient problem. LSTM is used in speech recognition field. LSTM has a chain-like structure similar to RNN.LSTM uses multiple gates to carefully regulate the amount of information that is allowed into each node state. It is best for prediction and classification in time series data. It consists of a cell. Cells keep track of dependency among data. When errors are back propagating error remains in LSTM cell.GRU is an LSTM without output gate.

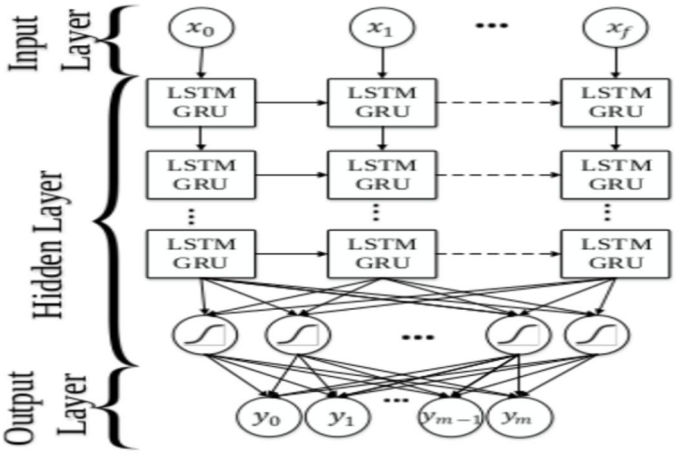


Fig 6. Standard LSTM

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

9) *Convolutional Neural Networks (CNN):* Convolutional Neural Networks (CNN) are also known as ConvNets. It is a deep learning architecture used for hierarchical document classification .It is used for text classification. CNN is a regularised version of networks having perceptron. It performs linear operation or convolutional operation. It takes an input image and assign some weights to various object in that image and using it we can differentiate each object. In CNN [7], an image tensor is used for image processing. Image tensor has a set of kernels of size d × d. These convolution layers are called feature maps .It is stacked to provide multiple filters on the input. CNN uses pooling. This reduces the computational complexity. Receptive field indicates the entire previous layer which is input area. Different pooling techniques are used to reduce outputs .The most common pooling method is max pooling .Pooling reduces the dimension of the data. Here the maximum element in the pooling window is selected.. Weights and bias are called filters. The same filter is used by several neuron. The final layer is a fully connected layer. CNN is used in image recognition, video analysis etc.

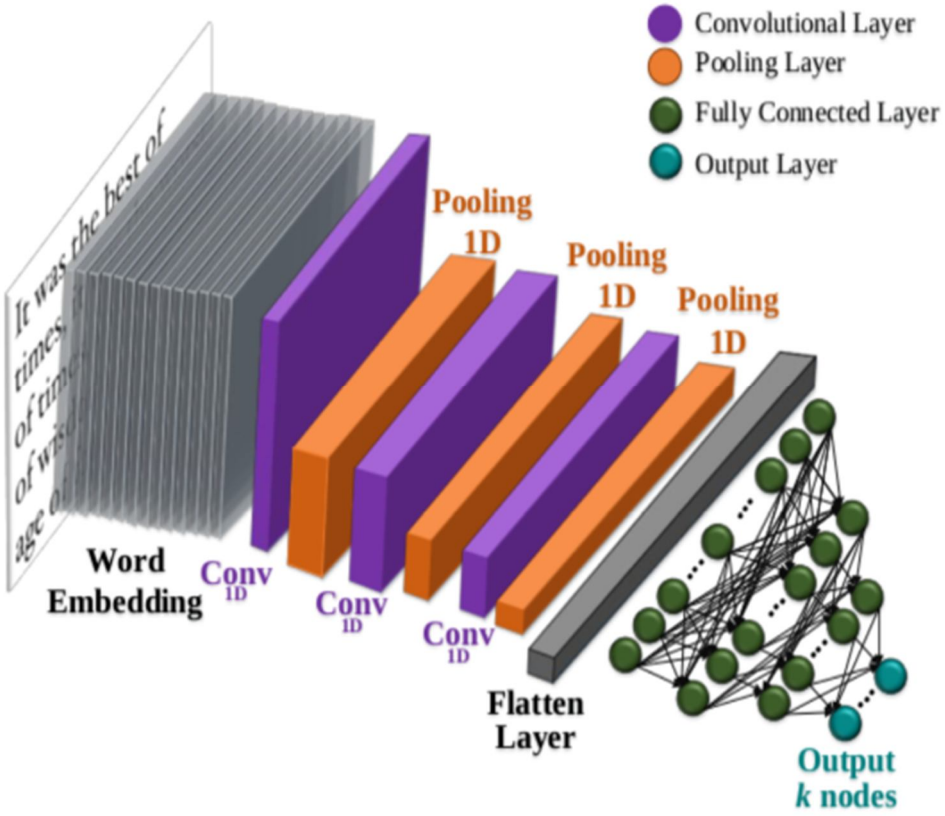


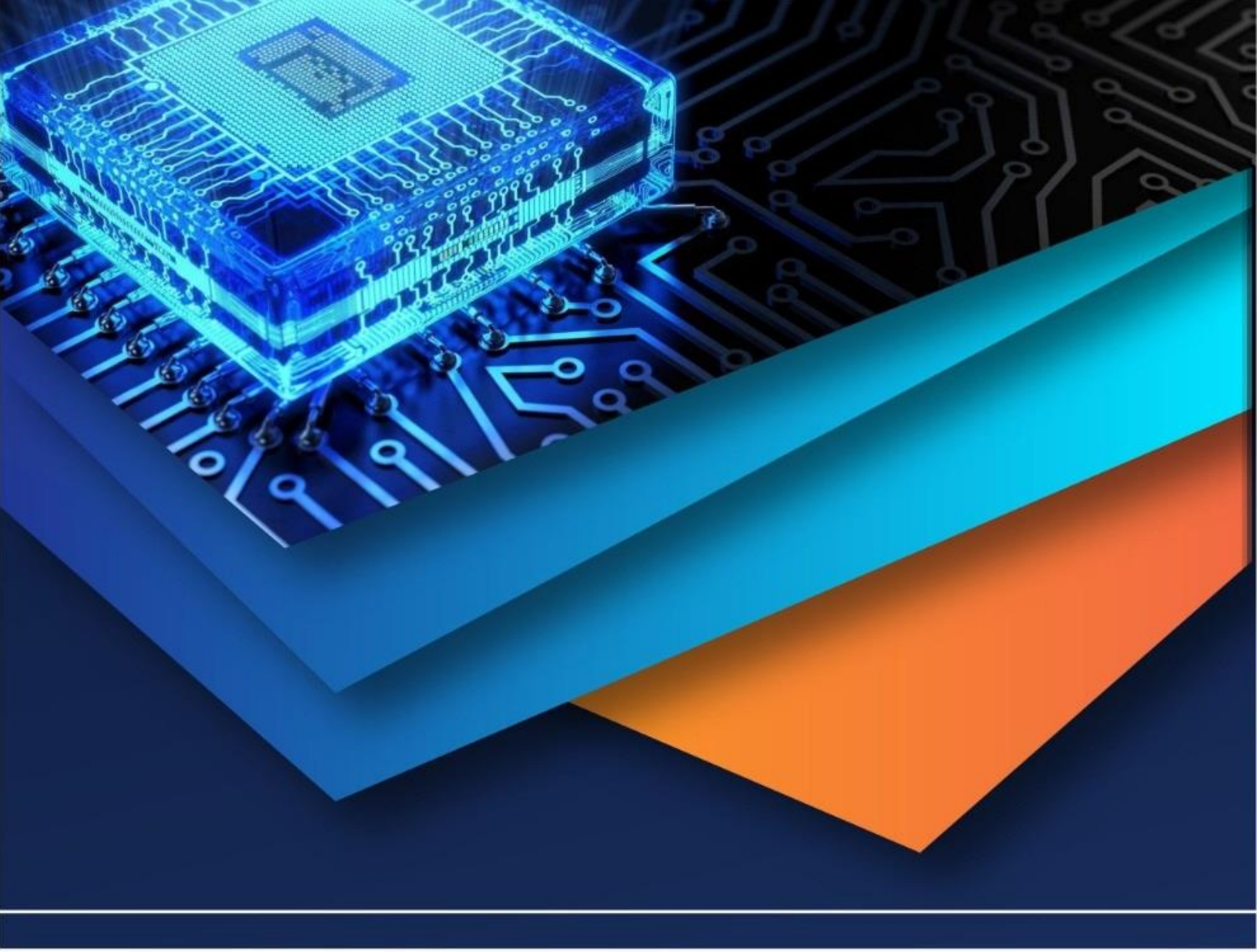
Fig 7. CNN

III.CONCLUSIONS

The short text detection from static and dynamic sources and its classification task is one of the most indispensable problems in machine learning. GISCA and Strokelets concepts define text recognition and detection in an effective manner. In SSFT search is on superpixels and search is in one direction. In SFT search is on pixels and in two directions. In SSFT text location is known and in SFT text location is unknown. SSFT is more robust to noise as colour comparison is on superpixels. Addition of LSTM above CNN effectively extracts features. Failure cases in strokelets such as Extreme Tilt, Heavy blur, Scribbling characters out of alphabet are resolved using superpixel based strokelet methods. Suitable methods are to be choosen for text detection according to the given scenario. As text and document data sets proliferate, the development and documentation of supervised machine learning algorithms becomes an important issue, especially for text classification. Having a better system for categorisation of documents for this information requires understanding these algorithms. However, the existing text detection and classification algorithms work more efficiently if we have a better understanding of feature extraction methods. In this survey, recent techniques and trending of text detection and classification algorithm have discussed.

REFERENCES

[1] Meng Cao,Yuexian Zou, Donmgming Yang: "GISCA:Gradient Inductive Segmentation Network with Contextual Attention For Scene Text Detection," IEEE Transaction, School of Electronics and Computer Engineering ,Japan, May 2019
[2] Xiang Bai, Cong Yao and Wenyu Liu, "Strokelets: A Learned Multi-Scale Mid-Level Representation for Scene Text Recognition," IEEE Transaction on Image Processing, June 2019
[3] Lucas Neuman and Jiri Matas, "Real -Time Lexicon free scene text Localization and Recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence. p. 1776–1781., Sep. 2016
[4] X.H. Phan, L.M. Nguyen, and S. Horiguc,"Learning to classify short and sparse text on web with hidden topics from large-scale data collections," Proceedings of the 17th International Conference on Image Processing, June 2019.
[5] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J, "Efficient estimation of word representations in vector space," arXiv**2013**, arXiv:1301.3781.
[6] Jiang, S.; Pang, G.;Wu, M.; Kuang, L, "An improved K-nearest-neighbor algorithm for text categorization,"Expert Syst. Appl. **2012**, 39, 1503–1509. [CrossRef]
[7] Vapnik, V.; Chervonenkis, A.Y, "A class of algorithms for pattern recognition learning," Avtomat. Telemekh**1964**,25, 937–945.
[8] Boser, B.E.; Guyon, I.M.; Vapnik, V.N, "A training algorithm for optimal margin classifiers," In Proceedingsof the Fifth AnnualWorkshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992;pp. 144–152.
[9] Bo, G.; Xianwu, H, "SVM Multi-Class Classification," J. Data Acquis. Process. **2006**, 3, 017.
[10] Giovanelli, C.; Liu, X.; Sierla, S.; Vyatkin, V.; Ichise, R, "Towards an aggregator that exploits big data to bid onfrequency containment reserve market," In Proceedings of the 43rd Annual Conference of the IEEE IndustrialElectronics Society (IECON 2017), Beijing, China, 29 October–1 November 2017; pp. 7514–7519.
[11] Quinlan, J.R, "Simplifying decision trees," Int. J. Man-Mach. Stud. **1987**, 27, 221–234. [CrossRef]
[12] Ho, T.K, "Random decision forests," In Proceedings of the 3rd International Conference on Document Analysisand Recognition, Montreal, QC, Canada 14–16 August 1995; Volume 1, pp. 278–282. [CrossRef]
[13] Breiman, L, "Random Forests," UC Berkeley TR567; University of California: Berkeley, CA, USA, 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY