Amrita-CEN-SentiDB:Twitter Dataset for Sentimental Analysis and Application of Classical Machine Learning and Deep Learning

Naveenkumar K S
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India.
naveensivakumarr@gmail.com

Vinayakumar R
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India.
vinayakumarr77@gmail.com

Soman K P
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India.

Abstract— **Social media is a platform in which tons and tons of text are generated each and every day. The data is so large that cannot be easily understood, so this has paved a path to a new field in the information technology which is natural language processing. In this paper, the text data which is used for the classification is tweets that determines the state of the person according of the sentiments which is positive, negative and neutral. Emotions are the way of expression of the person's feelings which has a high influence on the decision making tasks. Here we have proposed the text representation, Term Frequency Inverse Document Frequency (tfidf), Keras embedding along with the machine learning and deep learning algorithms for the purpose of the classification of the sentiments, out of which Logistics Regression machine learning based methods out performs well when the features is taken in the limited amount as the features increases Support Vector Machine (SVM) which is also one of the machine learning algorithm out performs well making a benchmark accuracy for this dataset as the 75.8%. For the research purpose the dataset has been made publically available.**

Keywords— Twitter, Sentiment, Sentiment Analysis, Text Representation, Machine learning, Deep learning.

I. INTRODUCTION

Social media resources such as Facebook, Twitter, Whatsapp have become one of the most powerful tool for communication, that rule the entire social media [1]. All these application helps to communicate with person who is known or unknown to us, and also helps to share emotions in different aspects such as text, emojis etc. These text and emojis helps us to understand the behavior and mentality of peoples. These social media texts are collected from various sources activity and many applications are built to understand the mentality of the peoples [2]. These analysis are more helpful in industry sectors such as brand monitoring, mobile phones, clothes, cosmetics etc. Mostly online service like swiggy, dominos, amazon, flipkart improve their services by analyzing the customer reviews, when it comes to understanding the psychological behavior of a person pertaining to a situation in which how the person is being affected and how each person solvers the problems also taken in to consideration [3, 15]. To understand such psychological behavior and peoples need the tweet from the social media is collected and analyzed according to the emotion contents in the messages.

Mostly the social media tweets contains the personal opinion of the individual or group of peoples sharing their feelings as the emotion with respect to the subject matter that as per the situation [4, 6]. As per machine, this method is coined as sentimental analysis. These tweets is a paving path to textual analysis in the field of Natural language processing. In sentimental analysis "sentiment" in the tweet is considered in first place, mainly sentiment comes under the categorization of happy, sad, angry [9, 15].

Sentimental analysis depends on the context of the tweets, in the current scenario, to extract the context of the tweets feature engineering methods are used, by creating a context with proper tone and sentiment indication within the tweets [18, 12]. The main objectives of this work are as follows:

1. To develop a Twitter database for the sentimental analysis.
2. To develop a machine learning based automated system for the sentimental analysis.
3. To perform various text representation methods and to evaluate those methods on the sentimental database.
4. To do a comparative study on the sentimental database by using the machine learning and the deep learning techniques.

The collected dataset are named as Amrita-CEN-SentiDB. The collected data is pre-processed and features are extracted and classified by the methods stated above. The rest of the sections are arranged as follows Related works in Section [II], Background in Section [III], Dataset description in Section [IV], Methodology in Section [V], Statistical methods in Section [VI], Experimental analysis and results in Section [VII], Conclusion in Section [VIII], Future works in Section [IX] and References.

II. RELATED WORKS

In this section various related works from 'classical' method to current advanced methods which are used for sentimental analysis is been discussed. In classical method most word net a lexical database helps in identification of emotion in the context, wordnet mostly measures the distance metrics and meaning of the word with respect to the orientation of

the language [17, 18] Lexical database methods are not more convenient in extracting the exact meaning and sentiment posted in the context. To over such problem many methods like word space model [3], emotinet, sentiword net techniques are implemented by [4, 13] proposing an unsupervised learning approach for the purpose of the analyzing the sentiments in the tweets. The polarity of the tweet is taken into the consideration and they are evaluated by some of the methods such as the senticnet, sentiwordnet and sentislangnet by extracting the slangs in the language and the acronyms [13]. Most of the classical methods fails in extracting an accurate information from the context. Machine learning method is used to overcome such obstacle by feature engineering techniques such as identification of frequency of each characters in the word and their occurrence and creating feature vector for processing the data [5]. Using the vector features [6, 7] have proposed a new model based on bayesian algorithm which helps in defining the dependent features in the context. In [8] the writer has discussed most of the existing methods and feature extraction techniques such as searching, inclusion criteria, exclusion criteria, presentation and pre-processing steps such as parts of speech tagging, stemming and lemmatization, stop words removal also been discussed [8]. The proposed solution for one of the major problem in sentimental analysis which is sentiment polarity categorization. In this paper, dataset for product reviews from online e-commerce website amazon is collected, totally 5.1 million products reviews are collected and categorized into four classes such as beauty products, book reviews, electronic items and home items [8]. Pos-tagging approach is used for feature extraction and machine learning approach such as navies Bayes, random forest and the SVM with the linear kernel and the rbf kernel are used to find the negative phrases of product reviews. In [9] discussed about general information about public opinion's about past and future history of sentiments using mining tools and techniques [10] proposed a novel method which is SLDA-sentence LDA and Aspect and Sentiment Unication Model (ASUM) for information extraction from blind text for product reviews. SLDA is a method of calculating the probability of all words that are generated in a sentence, as an extension of this method ASUM techniques which combines the sentiments and the aspects together for analysis. In this paper electronic dataset comprises of the various electronic items of total 22,000 reviews and the restaurant reviews contains of about total of 30,000 reviews of the various restaurants are used. The [11] author has collected dataset form online news, blogs, e-documents, e-mails etc. The classification process is done by tokenizing, removal of stop words and stemming. Followed by the preprocessing and the feature extraction task is the classification task, many classifiers are used such as KNN, Decision tree, Navies Bayes and SVM. Of these classifiers used SVM outperforms well by getting a result of about 80% [16, 12]. The model is trained were in the glovec is used [13, 15] which creates word vector for text classification task on sentimental analysis were SVM outperforms well with an accuracy of 95% [15]. The dataset uses three languages such as the English, Spanish and Arabic, the dataset is annotated based on emotions such as the joy, fear, angry and sadness. The experimentation are done using Glovec [15] pretrained model which contains 27 billion token and 2 billion tweets. Glovec and the SVD are used for the feature extraction task, and the extracted features are passed to random forest and the SVM for the classification [14].

III. BACKGROUND

A. *Term Frequency Inverse Document Frequency (tfidf)*

Term Frequency Inverse Document Frequency (tfidf) is a vectorizer that helps in conversion of the text to a vector feature by giving weights and helps in information extraction to pass in to the algorithm for smooth processing over the network model. These gives the information about how much that word is important to that document with respect to the corpus. The tfidf works in such a way that it calculates the weights on a value based system and also act as a central tool for finding the rank for the document, tfidf works in such a way that each and every sentence in the document is converted into the vector. The tfidf has the two parts; term frequency (TF) and inverse document frequency (IDF). The function of the term frequency is that it finds out the number of times the word occurs in the document but the inverse document frequency works in a different manner such as they are calculated by taking the log of the number of the documents present in the corpus divided by the number of the documents in which these terms appears. The feature can be extracted by using mindf and maxdf which are nothing but the maximum document frequency and minimum document frequency. The maxdf is a corpus specific parameters that are used for the removal of the frequent word. The mindf are used for checking the minimal occurrence of the words

The mathematical representation of the IDF is given below in Equation 1:

$$idf\left(t\right)=\log\left(\frac{D}{DF\left(t\right)}\right) \qquad (1)$$

B. *Keras Embedding*

Embedding is used for the word or character level extraction of features which are segregation of information [19, 21] from the particular context. It provide a dense representation of the words and their relative meanings which is an improved model over the sparse representations that are used in the bag of words representation. In keras embedding the data is given as the input which has the integer encoded in which each word is represented as the unique integer. The Embedding layer is a set with random weights which learns the words in the training phase of the data [20].

C. *Convolution Neural Network (CNN)*

Convolution Neural Network (CNN) is commonly used in the computer vision. Recently one dimensional CNN are used in the text classification. The working of the CNN are divided into the three parts they are convolution, maxpooling, fully connected. In the convolution the input

vectors are convolved by the means of the filter and the feature map is extracted from them, the extracted feature map is passed on to the maxpooling for the dimensionality reduction. The dimensional reduced output is given to the next layer that is the fully connected layer. All the inputs are connected to the dimensional reduced output in the fully connected layer, activation function is passed for the process of the normalization of the output.

D. Text Pre-Processing

Text pre-processing is one of the predominant approach in Sentimental analysis, pre-processing is done to clean the data in order to pass into the algorithm, the collected text data mostly have repeated letters, punctuation, capitalizations, stop words etc, these unnecessary contents are removed in pre-processing which helps the algorithm to process the raw input in an effective and easy way to understand the emotional context in the text. Normally to extract the useful features following methods are done:

1) Removal of repeated and meaning less characters:

Nowadays in the social media, the texts are not arranged in the proper grammatical order. People type so fast that the grammar in the languages are killed by them, by completely neglecting the grammar. The words are being cut shorted even though for example words like Hmmm, becoz, s, u etc. When it comes to sentence words such as "I need to c u" in these sentence "c" represents "see" and "u" represent "You". These can be easily understood by humans but computer can't understand such. So these words have to be removed for the betterment of the classification so they are neglected using the NLTK library.

2) Removal of Stop words:

Stop words are nothing but unnecessary words and words which are not required for the algorithm to process. Some of the example of stops words are 'these', 'is', 'because', 'can' 'the', 'from' etc. These stop words mostly occupies the space and reduce the computational timing and affects performance. The stop words are removed using NLTK library.

3) Stemming Approach:

Most of words in the English has its noun, adjective and verb. Mostly For example "play", has "play" as noun form, "playing" as verb form, "playful" as adjective which are having "play" as a stem by removing '-p', '-ing', '-ful' etc. Likewise in stop words these words also increase the size of the database and it is more over meaningless. To avoid such kind of unwanted words stemming approach has been used. Stemming is an approach of reduction of words from the root stem to stem words such as, prefixes and suffixes. Using stemming approach we can obtain reduction in data size and increase in the retrieval feature information.

4) Lemmatization Approach:

This approach is mostly similar to stemming approach but in lemmatization the morphological analysis of words is considered and root words have been considered instead of stem words. This process is bit slower when compared to stemming. In order to identify the correct word part of speech tagging is performed in all form such as noun, adjective, adverb and verb before lemmatization. For example: listening, listened, listen is stemmed to "listen" in verb form.

5) Word Tokenizing:

The tokenizing is the method which identify the particular word that tends in the formation of string. It is mainly helps to understand the meaning of the text combination of the relation between the words in a particular text. The words are separated to smaller terms, these conversion of words to smaller unit is known as tokenization. Generally this process is carried to capture each and every word in the text.

6) POS tagging:

Parts of Speech Tagging (POS) extracts the relation between the words by building a Named Entity Recognition (NER) which built by the parse trees. They are used in the process of the separation of the each words in the corpus corresponding to the each parts of speech based on the context of the topic. There are different POS tagging techniques and they are lexical based methods, rule based methods, probabilistic methods and deep learning methods. Here some of the examples of the POS tagging are given below: Noun e.g. Pineapple, Mango, Taj Mahal; Pronoun e.g. He, She, It, They; Adjective e.g. great, beautiful.

7) Extracted feature vectors:

Using the pre-processing techniques, the important features are extracted from the tweets. Some of the extracted feature example are shown in Tables I, II and III: for each classes.

TABLE I. EXTRACTED FEATURES FOR POSITIVE REVIEWS

Positive tweets	Feature words
I hope everyone has an awesome weekend I know that he is giving away some great Apple prizes.	Hope, awesome, giving, great, prizes.
I love that song, Even though she wrote it about Joe Jonas. It is still great and pleasant.	Love, great, pleasant

TABLE II. EXTRACTED FEATURES FOR NEGATIVE REVIEWS

Negative tweets	Feature words
We have been delayed for almost two hrs. I take this airline because I have had good luck but today is really frustrating.	Delayed, frustrating.
I miss my mom and dad with me in this trip, I hate them.	Miss, hate.

TABLE III. EXTRACTED FEATURES FOR NEUTRAL REVIEWS

Neutral tweets	Feature words
Average movie, but one time watchable	Average, movie, but, one, time, watch.
Sorry I was not able to hear you properly.	Sorry, was, not, able, hear, properly.

E. Logistics Regression

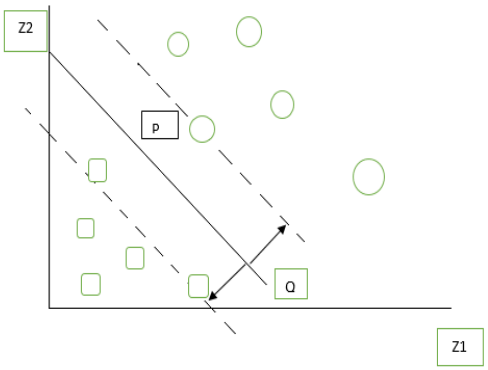
Logistics regression is a predictive analysis method similar to all other regression methods. It is used to find the relationship between the dependent and the independent variables. It helps in finding the maximum likelihood function by minimizing the squared residuals and optimize

best fitting line at a range of value 0 or 1 using logistic function which is sigmoid. This sigmoid function helps in mapping any real value number from 0 or 1.

F. Support Vector Machine (SVM)

Support vector machine (SVM) is one of the most commonly used method of classification for linear and non-linear datasets. This mainly plots the values of classes in n-dimensional space, where n represents the number of features taken for representations. For linearly separable data SVM makes a decision boundary to separate one class from another by making a linear optimal hyperplane along the classes. By finding the right hyperplane of the class which has maximum margin along the boundary and maximum distance between the data points of the classes. The classification process of the SVM is shown below in Fig I.

FIGURE I. SUPPORT VECTOR MACHINE



Were P stands for the optimal hyperplane and Q stands for the maximum margin. Mathematically they are defined as shown in Equation 2:

$$w.x + b = 0 \qquad (2)$$

Were w is the weight vector W = w 1 , w2, ..., wn. X is a training tuple and b is a scalar.

IV. DATASET DESCRIPTION

In this paper, dataset used here is twitter dataset that has been collected from the various tweet reviews. Here the sentimental analysis are done so for implementation of that task the tweets are categorized into three types they are; positive tweet, negative tweet and neutral tweet. In which class 0 as negative, class 1 as positive, class 2 as neutral are taken for analysis. These categorization of class are collected from various resources such as movie review, iphone review, airline reviews, electronics product reviews and general tweets such as Facebook, Twitter etc. The dataset is split into 70% for training and 30% for testing. These is the only dataset which contains all types of reviews, and helps to understand the sentiment and opinions of peoples in a single application. The collected dataset is named as the Amrita Cen-Sentimental Database, detailed description of the collected database split is shown in Table IV.

TABLE IV. DATASET DESCRIPTION

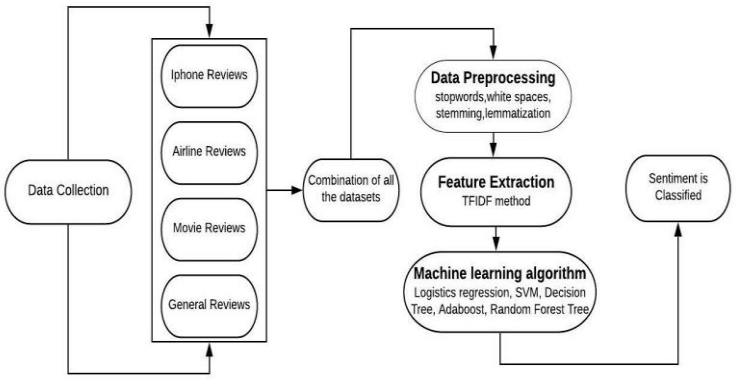
Dataset	Positive	Negative	Neutral	Total
Training	43,841	38,834	24,875	1,07,550
Testing	18,788	16,643	10,661	46,092

The collected dataset is raw data which consists of punctuations, special characters, white spaces, symbols, emojis, retweets, hashtags, lower case, numbers etc, this unwanted things are removed using some pre-processing steps which is been explained in methodology section.

V. METHODOLOGY

Three types of the sentiment are considered in this work such as the positive, negative and neutral. The data are collected which constitutes the reviews that are clubbed together to form a database of the mixture of these reviews. These contents consists of verbs, adverbs, adjective, phrase etc. For example “We are travelling in the aircraft for the very first time, so excited and overwhelmed” – The keyword “excited and overwhelmed” here refers to the positive state of the person based on which the sentiment is classified. In order to extract the emotional content information from the collected dataset text pre-processing is done. The detail proposed work is been shown in fig. The collected data is pre-processed by the steps and then passed to the tfidf method for the feature extraction and then to the various machine learning and the deep learning algorithms for the classification Fig II shows a full representation of the task done.

FIGURE II. FLOWCHART OF WORKDONE



VI. STATISTICAL MEASURES

In this work to measure the performance of the machine learning and the deep learning trained models, we estimate the confusion matrix. The confusion matrix is the one which decides the quality of the model that is trained by the means of comparing with the actual class with that of the predicted class, here the actual class refers to the data that is belonging to that particular group, were as the predicted class refers to the trained model when tested the machine predicts one class depending upon the training. They are also called as the error matrix, they give the actual classification rate from

the test data with the symbolic representation as that of the true positive, true negative, false positive, false negative. Now the terms are being explained below in a Table V:

TABLE V. EXPLANATION OF THE CONFUSION MATRIX

Terms	Definition
Positive (P)	The result is positive.
Negative (N)	The result is negative.
True Positive (TP)	Actual label is positive and the predicted is also positive.
True Negative (NP)	Actual label is negative and the predicted is also negative.
False Positive (FP)	Actual is negative but the predicted is positive.
False Negative (FN)	Actual is positive but the predicted is negative.

So, by this methods the data can be decided to which they belong to. Even the terms like the Accuracy, Precision, Recall, F1 score are derived from the confusion matrix. The accuracy is calculated by TP and TN and dividing them by the sum of TP, TN, FP, FN multiplied by 100. The model is judged based on the accuracy. Now the Precision is calculated by TP divided by the sum of the TP and the FP multiplied by 100. The Recall is calculated by the TP divided by the sum of the TP and the FN multiplied by 100. Then finally the F1-score is calculated by double the times of the precision multiplied by the recall divided by the sum of the precision and recall multiplied by 100. So the statistical measures decides the model performance by confusion matrix.

VII. EXPERIMENTAL ANALYSIS AND RESULTS

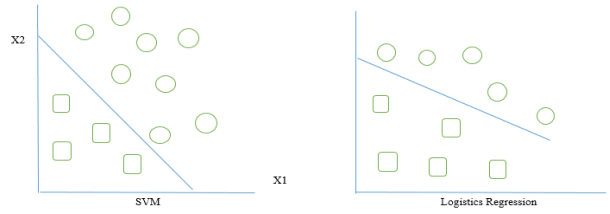
We used the train data and trained the model with the features extracted using TFIDF method, the extracted features are passed into various machine learning algorithms at different feature length of 10,000 to 40,000. Different feature length is selected to analyze how feature selection is important for learning important information and how its helps in increasing the accuracy of the model. For each selected features various machine learning algorithms. In which at 10,000 to 20,000 feature level logistic regression outperforms well compared to all other algorithms, in 30,000 to 40,000 feature level SVM outperforms well compared to all other algorithm on the taken dataset as show in Table VI.

TABLE VI. RESULTS FOR MACHINE LEARNING

Features	Classifiers	Accuracy	Precision	Recall	F-Score
10,000	Decision Tree	64.2	64.2	64.2	64.2
	Adaboost	65.2	65.2	65.2	64.8
	Randomforest Tree	72.7	72.7	72.7	72.7
	SVM	72.1	72.2	72	72
	Logistics Regression	**75.3**	**75.2**	**75.3**	**75.3**
20,000	Decision Tree	64.2	64.2	64.2	64.3
	Adaboost	65	65	65	64.8
	Randomforest Tree	70.2	70.2	70.2	70.8
	SVM	72.5	72.5	72.5	72.5
	Logistics Regression	**75.6**	**75.6**	**75.6**	**75.6**
30,000	Decision Tree	64.7	64.6	64.7	64.7
	Adaboost	65.7	65.7	65.7	65.4
	Randomforest Tree	71.4	71.4	71.4	71.4
	SVM	**75.7**	**75.7**	**75.7**	**75.7**
	Logistics Regression	70.2	70.2	70.2	70.2
40,000	Decision Tree	64.7	64.7	64.7	64.7
	Adaboost	65.3	65.3	65.3	65.1
	Randomforest Tree	70.2	70.2	70.2	70.2
	SVM	**75.8**	**75.8**	**75.8**	**75.8**
	Logistics Regression	69.2	69.3	69.2	69.2

The major difference between logistic regression and SVM is that loss function, LR function comes up with logistic loss and SVM with hinge loss. In logistics regression the diverging of loss is greater than the hinge loss, in our experiment we analyzed that with less feature LR out performs well. But when it comes to high features SVM perform with higher accuracy this is due to LR is more sensitive and it doesn't make zero when the class is been classified to the nearest value, these leads to degradation of accuracy. As show in fig we can conclude that the point which is been classified over the hyperplane space is been more accurate compared to LR. SVM maximize the closest point of class and predict only 0 or 1, and moreover SVM use different kernels tricks which helps in getting better accuracy when it comes to larger data and larger features. More over complex problems are deals with linear SVM which mostly have a better hyperspace liftment in classification of classes through the hyperplane. So in our experiment we have concluded that SVM linear perform well when it comes to higher features. Comparison of the SVM and LR are shown in the Fig III for the better understanding.

FIGURE III. COMPARISON OF SUPPORT VECTOR MCHINE AND LOGISTICS REGRESSION



Class 1 – circle; Class 2 - Rectangle

In deep learning only the CNN with 1D convolution layer is used, and the result obtained is been shown in Table VII ,and the result is been not good compared to machine learning approach this is due to the proper hyper tuning is been not done, but after several trail run experimentation this results are obtained.

TABLE VII. RESULTS FOR DEEP LEARNING

Algorithm	Accuracy	Precision	Recall	F-score	Time for computing
CNN	45.25	45.25	44.61	44.68	1,440 minutes

In CNN approach more over other than hyper tuning the neutral dataset is been less and its leads to misclassification of classes and decrease in accuracy. The problem which is been analyzed in deep learning approach is been applicable for machine learning. The misclassification result for machine learning approach is been explained by showing the number of actual class and the predicted class as show in Table VIII.

TABLE VIII. CONFUSION MATRIX

Architecture		Confusion Matrix			
		Predicted Class			
			0	1	2
Decision Tree	Actual Class	0	**64**	15	21
		1	21	**63**	16
		2	25	22	**53**
Adaboost		0	**65**	16	19
		1	11	**70**	19
		2	26	21	**53**
Random Forest tree		0	**72**	13	15
		1	12	**63**	25
		2	27	23	**50**
SVM		0	**75**	12	13
		1	13	**72**	15
		2	16	22	**62**
Logistics Regression		0	**69**	17	14
		1	19	**69**	12
		2	29	6	**65**

Here, in this table the confusion matrix is drawn between the actual class and the predicted class. 0 refers to Negative data, 1 refers to positive data and 2 refers to Neutral data. It is a multi-class classification task.

VIII. CONCLUSION

The twitter data is taken which has been named as the Amrita-CEN-SentiDB which contains a mixture of the tweets that are nothing but the mixture of the reviews positive, negative and neutral. Various text representation methods are being followed along with the machine learning and the deep learning approach. Of which the support vector machine from the machine learning outperforms well by creating a bench mark accuracy of about 75.8%. This is the benchmark accuracy for this dataset that has been made publically available for the research purposes. The link for the publically available dataset is https://vinayakumarr.github.io/Amrita-CEN-SentiDB/ .

REFERENCES

[1] J. Kamps, M. Marx, R. Mokken and M. De Rijke, ‘Using wordnet to measure semantic orientations of adjectives’, 2004.

[2] C. Fellbaum, ‘Wordnet: An electronic lexical database (language, speech, and communication)’, 1998.

[3] D. Pucci, M. Baroni, F. Cutugno and A. Lenci, ‘Unsupervised lexical substitution with a word space model’, Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence, Citeseer, 2009.

[4] A. Balahur, J. Hermida and A. Montoyo, ‘Building and Exploiting Emotinet, a knowledge base for emotion detection based on the appraisal theory model’, Affective Computing, IEEE Transactions, vol. 3, 188101, 2012.

[5] Neethu M, S, Rajasree R,’Sentiment analysis in Twitter using Machine Learning Techniques’, 4th ICCCNT , 2013.

[6] Z. Niu, Z. Yin and X. Kong, ‘Sentiment classification for microblog by machine learning,’, Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, vol. 286289, 2012.

[7] Asghar, Muhammad Zubair, et al. "A review of feature extraction in sentiment analysis." Journal of Basic and Applied Scientific Research 4.3 (2014): 181-186.

[8] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 5.

[9] Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis." IEEE Intelligent Systems 28.2 (2013): 15-21.

[10] Jo, Yohan, and Alice H. Oh. "Aspect and sentiment unification model for online review analysis." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[11] Khan, Aurangzeb, et al. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology 1.1 (2010): 4-20.

[12] Maas, Andrew L., et al. "Learning word vectors for sentiment analysis." Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. Association for Computational Linguistics, 2011.

[13] Pandarachalil, Rafeeque, Selvaraju Sendhilkumar, and G. S. Mahalakshmi. "Twitter sentiment analysis for large-scale data: an unsupervised approach." Cognitive computation 7.2 (2015): 254-262.

[14] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre-processing in sentiment analysis." Procedia Computer Science 17 (2013): 26-32.

[15] George, Anon, Barathi Ganesh HB, and K. P. Soman. "Teamcen at semeval-2018 task 1: global vectors representation in emotion detection." Proceedings of the 12th international workshop on semantic evaluation. 2018.

[16] Khan, Aurangzeb, et al. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology 1.1 (2010): 4-20.

[17] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010.

[18] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Language in Social Media (LSM 2011). 2011.

[19] Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. "Detecting malicious domain names using deep learning approaches at scale." Journal of Intelligent & Fuzzy Systems 34.3 (2018): 1355-1367.

[20] Vinayakumar, R., Prabaharan Poornachandran, and K. P. Soman. "Scalable framework for cyber threat situational awareness based on domain name systems data analysis." Big Data in Engineering Applications. Springer, Singapore, 2018. 113-142.

[21] Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. "Evaluating deep learning approaches to characterize and classify malicious URL’s." Journal of Intelligent & Fuzzy Systems 34.3 (2018): 1333-1343.

[22] Vinayakumar, R., Alazab, M., Srinivasan, S., Pham, Q. V., Padannayil, S. K., & Simran, K. (2020). A Visualized Botnet Detection System based Deep Learning for the Internet of Things Networks of Smart Cities. IEEE Transactions on Industry Applications.

[23] Venkatraman, S., Alazab, M., & Vinayakumar, R. (2019). A hybrid deep learning image-based analysis for effective malware detection. Journal of Information Security and Applications, 47, 377-389.

[24] Naveenkumar, K. S., Vinayakumar, R., & Soman, K. P. (2019, July). Amrita-CEN-SentiDB 1: Improved Twitter Dataset for Sentimental Analysis and Application of Deep learning. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[25] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., & Venkatraman, S. (2019). Robust intelligent malware detection using deep learning. IEEE Access, 7, 46717-46738.

[26] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525-41550.