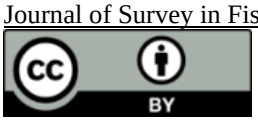

Two Level LSTM For Sentiment Analysis Using Lexicon Embedding And Polar Flipping

Mrs. Sunitha Surarapu[1]*, Mrs. Srilatha Puli[2], Aakash Bonagiri[3], Kanagala Lahari[4], Sachin Bhukya[5], Shivani Pandala[6], Chittimalli Alekhya[7]

[1]*Assistant Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
sunithasurarapu@sreyas.ac.in
[2]Assistant Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
srilatha.puli@sreyas.ac.in
[3]Department of CSE, Sreyas Institute of Engineering and Technology,Telangana,India.
aakashbonagiri1103@gmail.com
[4]Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
laharikanagala5@gmail.com
[5]Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
bhukyasachin22@gmail.com
[6]Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
shivanipandala@gmail.com
[7]Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
chittimallialekhya@gmail.com

***Corresponding Author: -** Mrs. Sunitha Surarapu
*Assistant Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.
sunithasurarapu@sreyas.ac.in

Abstract
Sentiment analysis is a key component in various text mining applications. Numerous sentiment classification techniques, including conventional and deep-learning-based methods, have been proposed in the literature. In most existing methods, a high-quality training set is assumed to be given. Nevertheless, constructing a high-quality training set that consists of highly accurate labels is challenging in real applications. This difficulty stems from the fact that text samples usually contain complex sentiment representations, and their annotation is subjective. We address this challenge in this study by leveraging a new labelling strategy and utilizing a two-level long short-term memory network to construct a sentiment classifier. Lexical cues are useful for sentiment analysis, and they have been utilized in conventional studies. For example, polar and negation words play important roles in sentiment analysis. A new encoding strategy, that is, ρ hot encoding, is proposed to alleviate the drawbacks of one-hot encoding and, thus, effectively incorporate useful lexical cues. Moreover, the sentimental polarity of a word may change in different sentences due to label noise or context. A flipping model is proposed to model the polar flipping of words in a sentence. We compile three Chinese datasets on the basis of our label strategy and proposed methodology. Experiments demonstrate that the proposed method outperforms state-of-the-art algorithms on both benchmark English data and our compiled Chinese data.

Keywords: Text classification, sentiment analysis, LSTM, lexicon embedding, flipping

1. INTRODUCTION

Text is important in many artificial intelligence applications. Among various text mining techniques, sentiment analysis is a key component in applications, such as public opinion monitoring and comparative analysis. Sentiment analysis can be divided into three problems according to input texts, namely, sentence, paragraph, and document levels. This study focuses on sentence and paragraph levels Text sentiment analysis is usually considered a text classification problem. Almost all existing text classification techniques are applied to text sentiment analysis. Typical techniques include bag-of-words (BOW)-based, topic model-based, deep learning-based, and lexicon based (or rule-based) methods. Although many achievements have been made and sentiment analysis has been successfully used in various commercial applications, its accuracy can be further improved. The construction of a high-
accuracy sentiment classification model usually entails the challenging compilation of training sets with numerous samples and sufficiently accurate labels. The reason behind this difficulty is two-fold. First, the sentiment is somewhat subjective, and a sample may receive different labels from different users. Second, some texts contain complex sentiment representations, and a single label is difficult to provide. We conduct a statistical analysis of public Chinese sentiment text sets in GitHub. The results show that the average label error is larger than 10%. This error value reflects the degree of difficulty of sentiment labelling.

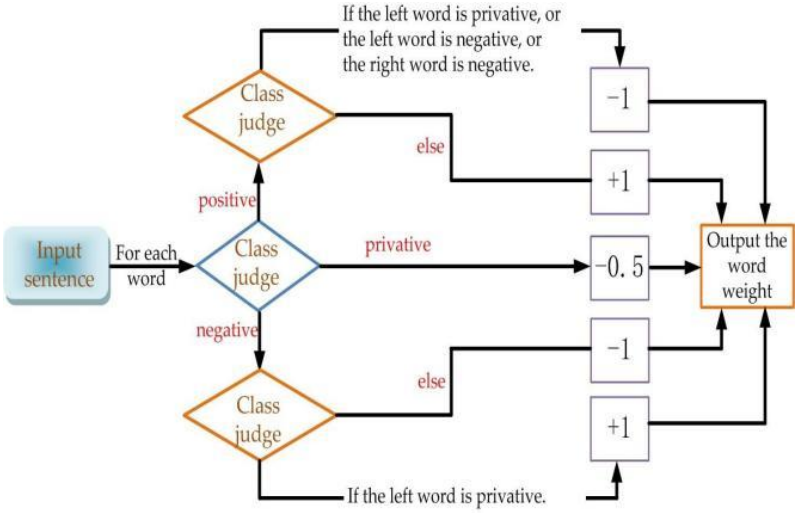

Figure.1: A lexicon-based approach for sentiment classification

Negation and interrogative sentences are difficult to classify when deep-learning-based methods are applied. Although lexicon-based methods can deal with particular types of negation sentences, their generalization capability is poor. We address the above issues with a new methodology. First, we introduce a two-stage labelling strategy for sentiment texts. In the first stage, annotators are invited to label a large number of short texts with relatively pure sentiment orientations. Each sample is labelled by only one annotator. In the second stage, a relatively small number of text samples with mixed sentiment orientations are annotated, and each sample is labelled by multiple annotators. Second, we propose a two level long short-term memory (LSTM) [6] network to achieve two-level feature representation and classify the sentiment orientations of a text sample to utilize two labelled datasets. Third, in the proposed two-level LSTM network, lexicon embedding is leveraged to incorporate linguistic features used in lexiconbased methods. Finally, the labels in a word-polar dictionary usually contain noise and the polarity of a word can also change in different contexts. A flipping model is proposed to model the sentiment polarity flipping of a word in a sentence. Three Chinese sentiment datasets are compiled to investigate the performance of the proposed methodology. The experimental results demonstrate the effectiveness of the proposed methods.

2. LITERATURE REVIEW

Twitter as a corpus for sentiment analysis and opinion mining

Probabilistic topic models have been widely used for sentiment analysis. However, most of existing topic methods only model the sentiment text, but do not consider the user, who expresses the sentiment, and the item, which the sentiment is expressed on. Since different users may use different sentiment expressions for different items, we argue that it is better to incorporate the user and item information into the topic model for sentiment analysis. In this paper, we propose a new Supervised User-Item based Topic model, called SUIT model, for sentiment analysis. It can simultaneously utilize the textual topic and latent user-item factors. Our proposed method uses the tensor outer product of text topic proportion vector, user latent factor and item latent factor to model the sentiment label generalization. Extensive experiments are conducted on two datasets: review dataset and microblog dataset. The results demonstrate the advantages of our model. It shows significant improvement compared with supervised topic models and collaborative filtering methods.

Convolutional neural networks for sentence classification

We report on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks. We show that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance. We additionally propose a simple modification to the architecture to allow for the use of both task-specific and static vectors. The CNN models discussed herein improve upon the state of the art on 4 out of 7 tasks, which include sentiment analysis and question classification.

Lexicon based methods for sentiment analysis:

We present a lexicon-based approach to extracting sentiment from text. The Semantic Orientation Calculator (SO-CAL) uses dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation. SO-CAL is applied to the polarity classification task, the process of assigning a positive or negative label to a text that captures the text's opinion towards its main subject matter. We show that SO-CAL's performance is consistent across domains and in completely unseen data. Additionally, we describe the process of dictionary creation, and our use of Mechanical Turk to check dictionaries for consistency and reliability.

Context-sensitive lexicon features for neural sentiment analysis

Sentiment lexicons have been leveraged as a useful source of features for sentiment analysis models, leading to the state-of-the-art accuracies. On the other hand, most existing methods use sentiment lexicons without considering context, typically taking the count, sum of strength, or maximum sentiment scores over the whole input. We propose a context-sensitive lexicon-based method based on a simple weighted-sum model, using a recurrent neural network to learn the sentiments strength, intensification and negation of lexicon sentiments in composing the sentiment value of sentences. Results show that our model can not only learn such operation details, but also give significant improvements over state-of-the-art recurrent neural network baselines without lexical features, achieving the best results on a Twitter benchmark.

3.METHODOLY

The construction of a high-accuracy sentiment classification model usually entails the challenging compilation of training sets with numerous samples and sufficiently accurate labels. The reason behind this difficulty is two-fold. First, the sentiment is somewhat subjective, and a sample may receive different labels from different users. Second, some texts contain complex sentiment representations, and a single label is difficult to provide. We conduct a statistical analysis of public Chinese sentiment text sets in GitHub. The results show that the average label error is larger than 10%. This error value reflects the degree of difficulty of sentiment labelling. Negation and interrogative sentences are difficult to classify when deep-learning-based methods are applied. Although lexicon-based methods can deal with particular types of negation sentences, their generalization capability is poor.

Disadvantages:

1. some texts contain complex sentiment representations, and a single label is difficult to provide
2. generalization capability is poor
3. The average label error is larger than 10%. This error value reflects the degree of difficulty of sentiment labelling.

We address the above issues with a new methodology. First, we introduce a two-stage labelling strategy for sentiment texts. In the first stage, annotators are invited to label a large number of short texts with relatively pure sentiment orientations. Each sample is labelled by only one annotator. In the second stage, a relatively small number of text samples with mixed sentiment orientations are annotated, and each sample is labelled by multiple annotators. Second, we propose a two level long short-term memory (LSTM) network to achieve two-level feature representation and classify the sentiment orientations of a text sample to utilize two labelled datasets. Third, in the proposed two-level LSTM network, lexicon embedding is leveraged to incorporate linguistic features used in lexiconbased methods. Finally, the labels in a word-polar dictionary usually contain noise and the polarity of a word can also change in different contexts. A flipping model is proposed to model the sentiment polarity flipping of a word in a sentence.

Advantages:

1. Effectively incorporate useful lexical cues.
2. Achieved the highest accuracies on these three data corpora.

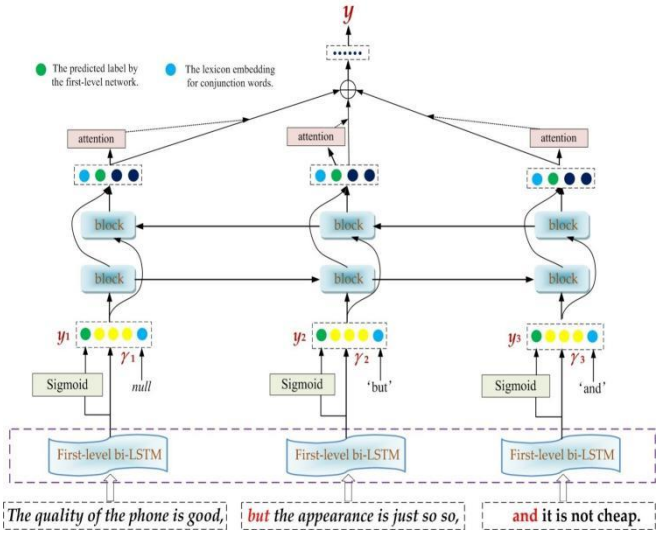

Figure.2: The whole two-level LSTM network with lexicon embedding inboth the input and attention layers.

MODULES:

To implement aforementioned project we have designed following modules

- Data exploration: using this module we will load data into system
- Processing: Using the module we will read data for processing
- Splitting data into train & test: using this module data will be divided into train & test
- Model generation: Random Forest - Decision Tree - KMM - Support Vector Machine - Voting Classifier - CNN - CNN+LSTM - LSTM - BiLSTM - RNN - CNN with KFoldVaildation.

- User signup & login: Using this module will get registration and login
- User input: Using this module will give input for prediction
- Prediction: final predicted displayed

4. IMPLEMENTATION

ALGORITHM 1: ρTl-LSTM

Input: Training sets T1 and T2; dictionary of key lexical words; POS for each word; dictionary of conjunction words; character/word embeddings for each character/word; parameters $\lambda 1$ and $\lambda 2$.

Output: A trained two-level LSTM for sentiment classification.

Steps:

1. Construct the ρ-hot-based embedding vector for each word (including punctuation) in the clauses in T1. The embeddings include the character/word and lexicon embeddings of each character/word;
2. Construct the embedding vector for each conjunction word in each clauses as the input for the second-level LSTM;
3. Train the two-level LSTM on the basis of the input embedding vectors and labels of polar words, the T1 text clauses, and the T2 text samples.

5. EXPERIMENTAL RESULTS

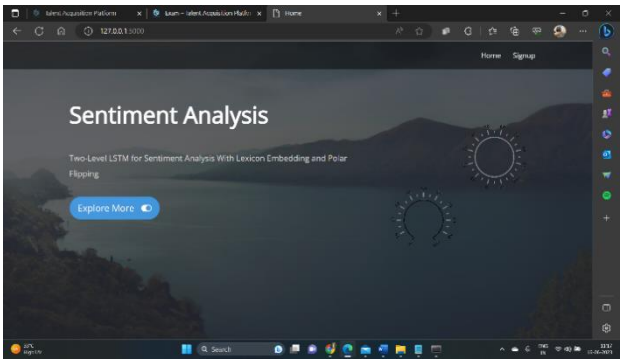

Figure 3: Home screen

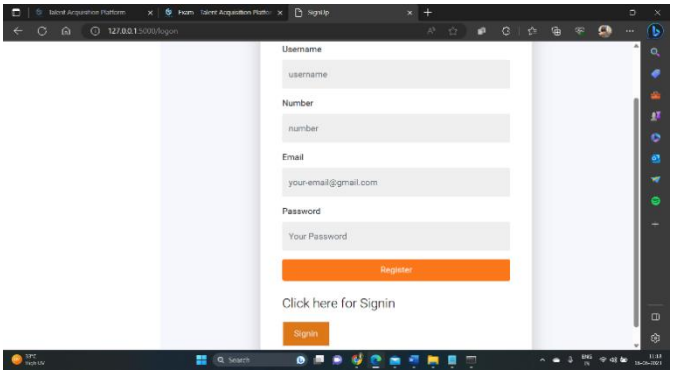

Figure 4: User registration

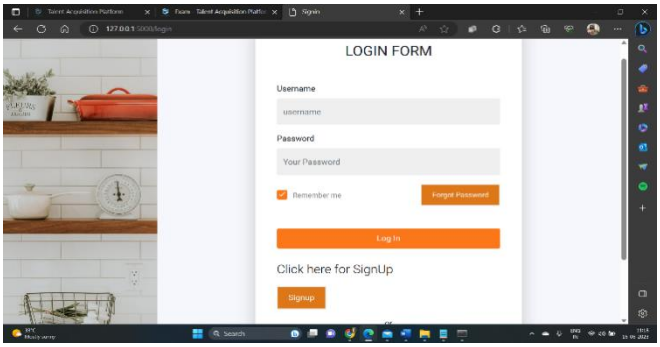

Figure 5: User Login

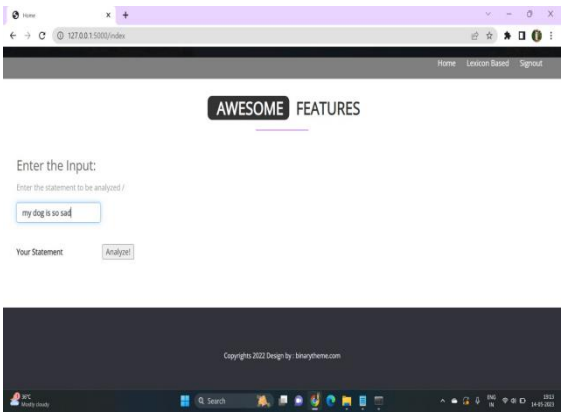

Figure 6: User input

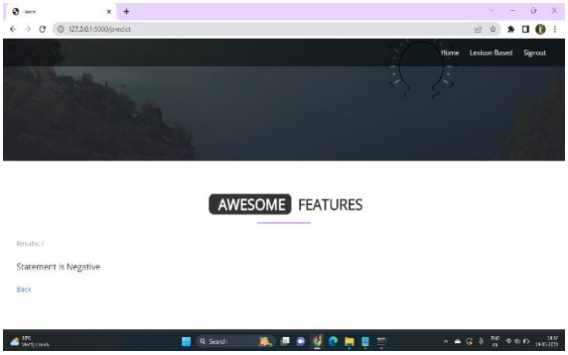

Figure 7: Prediction result

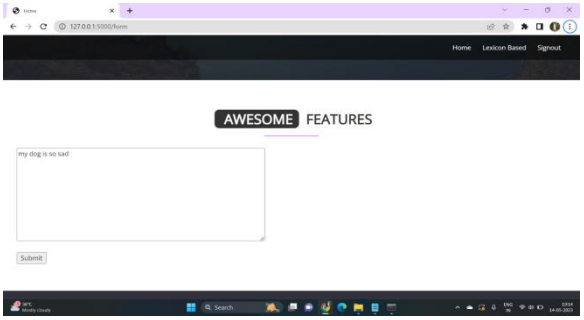

Figure 8: User input for lexicon based

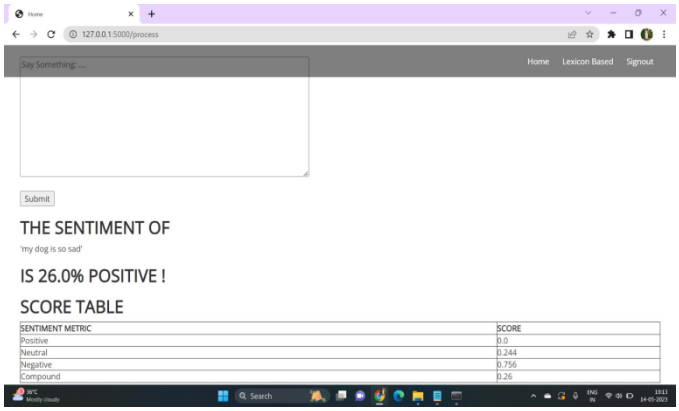

Figure 9: Output for lexicon based

6. CONCLUSION

High-quality labels are crucial for learning systems. Nevertheless, texts with mixed sentiments are difficult for humans to label in text sentiment classification. In this study, a new labelling strategy was introduced to partition texts into those with pure and mixed sentiment orientations. These two categories of texts were labelled using different processes. A two-level network was accordingly proposed to utilize the two labelled data in our two-stage labelling strategy. Lexical cues (e.g., polar words, POS, and conjunction words) are particularly useful in sentiment analysis. These lexical cues were used in our two-level network, and a new encoding strategy, that is, ρ-hot encoding, was introduced. ρ-hot encoding was

motivated by one-hot encoding. However, the former alleviates the drawbacks of the latter. Due to labelling noise or context, the polarity of a word varied in different texts. A flipping model was proposed to model the polarity flipping process. Three Chinese sentiment text data corpora were compiled to verify the effectiveness of the proposed methodology. Our proposed method achieved the highest accuracies on these three data corpora. On English data corpora, the proposed method outperformed state-of-the-art algorithms. The proposed two-level network and lexicon embedding can also be applied to other types of deep neural networks.
In our future work, we will extend our main idea into several networks and text mining applications.

7.REFERENCES

[1] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[2] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proc. Int. Conf. Lang. Resources Eval. (LREC), 2010, pp. 17–23.
[3] F. Li, S. Wang, S. Liu, and M. Zhang, "SUIT: A supervised user-item based topic model for sentiment analysis," in Proc. AAAI Conf. (AAAI), 2014, pp. 1636–1642.
[4] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. Int. Conf. Empirical Methods Nat. Lang. (EMNLP), 2014, pp. 1746–1751.
[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexiconbased methods for sentiment analysis," Comput. Linguist., vol. 37, no. 2, pp. 267–307, 2011.
[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[7] Z. Teng, D.-T. Vo, and Y. Zhang, "Context-sensitive lexicon features for neural sentiment analysis," in Proc. EMNLP, 2016, pp. 1629–1638.
[8] Q. Qian, M. Huang, J. Lei, and X. Zhu, "Linguistically regularized LSTMs for sentiment classification," in Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL), 2017, pp. 1679–1689.
[9] W. Zhao et al., "Weakly-supervised deep embedding for product review sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 30, no. 1, pp. 185–197, Sep. 2018.
[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD), 2004, pp. 168–177.
[11] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in Proc. Int. Conf. Empirical Methods Nat. Lang. (EMNLP), 2004, pp. 412–418.
[12] S. Liu, X. Cheng, F. Li, and F. Li, "TASC: Topic-adaptive sentiment classification on dynamic tweets," IEEE Trans. Knowl. Data Eng., vol. 27, no. 6, pp. 1696–1709, Dec. 2015.
[13] G. Paltoglou and M. Thelwall, "A study of information retrieval weighting schemes for sentiment analysis," in Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL), 2010, pp. 1386–1395.
[14] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in Proc. Int. Conf. Mach. Learn. (ICML), 2011, pp. 513–520.
[15] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target dependent sentiment classification," in Proc. Int. Conf. Comput. Linguist. (COLING), 2016, pp. 3298–3307.
[16] SRILATHA PULI, A MACHINE LEARNING MODEL FOR AIR QUALITY PREDICTION FOR SMART CITIES, DESIGN ENGINEERING || ISSN: 0011-9342 | YEAR 2021 - ISSUE: 9 | PAGES: 18090 – 18104
[17] SRILATHA PULI, QUALITY RISK ANALYSIS FOR SUSTAINABLE SMART WATER SUPPLY USING DATA PERCEPTION, INTERNATIONAL JOURNAL OF HEALTH SCIENCES ISSN 2550-6978 E-ISSN 2550-696X © 2022, HTTPS://DOI.ORG/10.53730/IJHS.V6NS5.9826, 18 JUNE 2022
[18] SRILATHA PULI, URBAN STREET CLEANLINESS, JOURNAL OF ALGEBRAIC STATISTICS VOLUME 13, NO. 3, 2022, P. 547-552, HTTPS://PUBLISHOA.COM, and ISSN: 1309-3452
[19] SRILATHA PULI, SELF-ANNIHILATION IDEATION DETECTION, and NEURO QUANTOLOGY | JUNE 2022 | VOLUME 20 | ISSUE 6 | PAGE 7229-7239 | DOI: 10.14704/NQ.2022.20.6.NQ22727
[20] SRILATHA PULI, CRIME ANALYSIS USING MACHINE LEARNING, YMER|| ISSN: 0044-0477, APRIL 2022
[21] SRILATHA PULI, N-GRAMS ASSISTED YOUTUBE SPAM COMMENT DETECTION, YMER || ISSN: 0044-0477, APRIL 2022
[22] SRILATHA PULI, ANALYSIS OF BRAND POPULARITY USING BIG DATA AND TWITTER, YMER|| ISSN: 0044-0477, APRIL 2022
[23] SRILATHA PULI, CYBER THREAT DETECTION BASED ON ARTIFICIAL NEURAL NETWORKS USING EVENT PROFILES, THE INTERNATIONAL JOURNAL OF ANALYTICAL AND EXPERIMENTAL MODAL ANALYSIS, ISSN NO: 0886-9367
[24] SRILATHA PULI, FACE MASK MONITORING SYSTEM, THE INTERNATIONAL JOURNAL OF ANALYTICAL AND EXPERIMENTAL MODAL ANALYSIS, ISSN NO: 0886-9367
[25] SRILATHA PULI, IOT BASED SMART DOOR LOCK SURVEILLANCE SYSTEM USING SECURITY SENSORS, ADVANCED SCIENCE LETTERS E-ISSN: 1936-7317
[26] SRILATHA PULI, SAFETY ALERTING SYSTEM FOR DROWSY DRIVER, 9TH INTERNATIONAL CONFERENCE ON INNOVATIONS IN ELECTRONICS & COMMUNICATION ENGINEERING (ICIECE-2021), PAGE – 40

[27] N. SWAPNA SUHASINI, SRILATHA PULI, BIG DATA ANALYTICS FOR MALWARE DETECTION IN A VIRTUALIZED FRAMEWORK, JOURNAL OF CRITICAL REVIEWS, ISSN:2394-5125 VOL.7, ISSUE 14, JULY – 2020

[28] SRILATHA PULI, BLOCK CHAIN BASED CERTIFICATE VALIDATION, INTERNATIONAL JOURNAL OF SCIENCE AND RESEARCH (IJSR), and ISSN: 2319-7064 SJIF (2022): 7.942, VOLUME 11 ISSUE 12, DECEMBER 2022, PAPER ID: SR221219113003, DOI: 10.21275/SR221219113003, WWW.IJSR.NET

[29] MRS. SRILATHA PULI, ENERGY EFFICIENT TEACHING-LEARNING-BASED OPTIMIZATION FOR THE DISCRETE ROUTING PROBLEM IN WIRELESS SENSOR NETWORK, INTERNATIONAL JOURNAL OF EARLY CHILDHOOD SPECIAL EDUCATION (INT-JECS) DOI: 10.48047/INTJECSE/V14I7.296 ISSN: 1308-5581 VOL 14, ISSUE 07 2022.

[30] MRS. SRILATHA PULI, A HYBRID BLOCK CHAIN-BASED IDENTITY AUTHENTICATION SCHEME FOR MULTI- WSN, INTERNATIONAL JOURNAL OF EARLY CHILDHOOD SPECIAL EDUCATION (INT-JECS) DOI: 10.48047/INTJECSE/V14I7.296 ISSN: 1308-5581 VOL 14, ISSUE 07 2022

[31] MRS. SRILATHA PULI, IMPLEMENTATION OF A SECURED WATERMARKING MECHANISM BASED ON CRYPTOGRAPHY AND BIT PAIRS MATCHING, INTERNATIONAL JOURNAL OF EARLY CHILDHOOD SPECIAL EDUCATION (INT-JECS) DOI: 10.48047/INTJECSE/V14I7.296 ISSN: 1308-5581 VOL 14, ISSUE 07 2022

[32] MRS. S.SUNITHA, MRS. SRILATHA PULI, MULTILEVEL DATA CONCEALING TECHNIQUE USING STEGANOGRAPHY AND VISUAL CRYPTOGRAPHY, INTERNATIONAL JOURNAL OF EARLY CHILDHOOD SPECIAL EDUCATION (INT-JECSE) DOI:10.48047/INTJECSE/V15I1.1 ISSN: 1308-5581 VOL 15, ISSUE 01 2023

[33] MRS. SRILATHA PULI, BLOOD BANK MANAGEMENT DONATION AND AUTOMATION, SPECIALUSIS UGDYMAS / SPECIAL EDUCATION 2022 1 (43), HTTPS://WWW.SUMC.LT/INDEX.PHP/SE/ARTICLE/VIEW/1995

[34] N. S. SUHASINI AND S. PULI, "BIG DATA ANALYTICS IN CLOUD COMPUTING," 2021 SIXTH INTERNATIONAL CONFERENCE ON IMAGE INFORMATION PROCESSING (ICIIP), SHIMLA, INDIA, 2021, PP. 320-325, DOI: 10.1109/ICIIP53038.2021.9702705.

[35] SURARAPU SUNITHA, BLOCKCHAIN-BASED ACCESS CONTROL SYSTEM FOR CLOUD STORAGE, YMER || ISSN: 0044-0477, APRIL 2022

[36] SURARAPU SUNITHA, ARTIFICIAL INTELLIGENCE SUPPORT FOR CLOUD COMPUTING INTRUSION, DEEP-CLOUD ISSUES, SOLID STAGE TECHNOLOGY, VOLUME:63, ISSUE:2S, PUBLICATION-2020

[37] SURARAPU SUNITHA, CRYPTOCURRENCY PRICE ANALYSIS USING ARTIFICIAL INTELLIGENCE, JOURNAL OF ALGEBRAIC STATISTICS VOLUME 13, NO. 3, 2022, P. 486-493 HTTPS://PUBLISHOA.COM ISSN: 1309-3452

[38] SURARAPU SUNITHA, AN EMPIRICAL STUDY ON SECURITY ISSUES AND MITIGATION TECHNIQUES IN OPPORTUNITIES NETWORKS, THINK INDIA JOURNAL, ISSN:0971-1260, VOL-22, ISSUE-41, DECEMBER-2019

[39] SURARAPU SUNITHA, A HYBRID BLOCK CHAIN-BASED IDENTITY AUTHENTICATION SCHEME FOR MULTI-WSN, INTERNATIONAL JOURNAL OF EARLY CHILDHOOD SPECIAL EDUCATION, ISSN: 1308-5581, VOL 14, ISSUE JULY-2022

[40].SUNITHASURARAPU, CRYPTOCURRENCY PRICE PREDICTION USING NEURAL NETWORKS, DEEP LEARNING AND MACHINE LEARNING , INTERNATIONAL JOURNAL FOR INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, ISSN:2456-5083, VOLUME 12, ISSUE 05 MAY 2023, PAGES:408-417

[41] SUNITHA SURARAPU, MULTILEVEL DATA CONCEALING TECHNIQUE USING STEGANOGRAPHY AND VISUAL CRYPTOGRAPHY, INTERNATIONAL JOURNAL OF EARLY CHILDHOOD SPECIAL EDUCATION, VOLUME 15, ISSN:1308-5581, IISUE JANUARY 2023