Journal of Theoretical and Applied Information Technology
15th February 2023. Vol.101. No 3
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A FRAMEWORK TO BUILD AND CLEAN MULTI-LANGUAGE TEXT CORPUS FOR EMOTION DETECTION USING MACHINE LEARNING

K ANUSHA[1] , D. VASUMATHI[2] , PRABHAT MITTAL[3]

[1*]Research Scholar, Dept. of Computer Science and Engineering,

JNTUH College of Engineering, JNT University, Hyderabad, 500085, India.

[2] Professor & HOD, Dept. of Computer Science and Engineering,

JNTUH College of Engineering, JNT University, Hyderabad, 500085, India.

E-mail: [1]anusha.kollu87@gmail.com, [2] rochan44@gmail.com , [3] profmittal@yahoo.co.in

ABSTRACT

In the recent times, extraction of emotions from the text corpuses have gained a huge popularity. The use of these extracted emotions is used for various purposes such as customer reviews for products, recommendations of books, movies or building opinion poll from the social media posts. A good number of research outcomes can be observed during the last few years. Nonetheless, these existing systems have significantly failed to solve particularly two major challenges. Firstly, the text corpus available in various sources are in diversified languages and during the extraction of the emotions these variations of language can be highly complete to extract the correct emotions due to the dependency on regional languages. Secondly, the increase of using emojis in the text corpuses have made the task of emotion extraction even challenging since, the use of emojis with text can sometimes reflect to a sarcasm and detection of the sarcasm can be highly complex. Henceforth, this work proposes an automated framework to build a sentiment extraction process with pre-processing of the text corpus for normalization of the text from emojis, sarcasm and multiple local language influences. The framework results into 10% decrease in time complexity and 80% improvement over the accuracy of emotion detection using proposed machine learning methods.

Keywords: *Topic Inference, Text Corpus Extraction, Emoji Replacement, Emoji Translation, Text Translation, Mutual Exclusion, Emotion Extraction*

1. INTRODUCTION

Emotion mining makes it harder to recognize public emotion, make commercial judgments, and forecast. Teachers overlook students' joy, sorrow, rage, fear, disdain, and astonishment. The suggested system recognizes emoticons (SEER). We classify user-provided content using an emotion-and-emoticon-based lexicon. Using a Bi-GRU network and attention mechanism, this project captures emotional vectors. Spreading emoticons online creates emotion vectors. This research chooses emoticons and weights from a tiny dataset. Emotion vectors reveal a text's mood. SEER improves EQ. The proposed technique boosts accuracy by 2.66 to 14.566, 5.38 to 4.488%, 4.9 to 2.68, and 3.17 [1].

Science emotions. AI prioritizes emotion identification over false-positives. Feelings hinder diagnosis. We study the emotional impact of online stories. Emotions can slow evolution and weaken or enhance connections. Text sentiment is determined using three features and two neural-network models. Contradiction (psychological instrument). Emotional growth is revealed through one-step, limited-step, and shortest-path transfers. Titles, content, and comments confirm methodology (long and short). Sometimes subjective expressions misrepresent rage. Sad, angry, wrath cycle. Objectivity fuels speculative joy. [2] We discuss HCI, social media, and public opinion.

NLP study identifies text emotions. Tweets, status updates, blogs, news pieces, and consumer evaluations might reveal emotions. This method uses emotional context and word embedding vectors. Sentimental language requires context and syntax. SENN uses pre-trained word representations to integrate semantic/syntactic and affective data. Second sub-network of SENN model employs CNN to extract emotional aspects and assess semantic correlations. Bidirectional Long-Short Term Memory (BiLSTM) analyses the environment to uncover semantic associations. Real-world evidence confirmed the theory. Use Ekman's six-emotion model. The proposed way is better.

Henceforth, after setting the context of the research, the rest of the paper is organized such that, in Section 2 the recent research reviews are analyzed, the proposed solutions and the furnished algorithms are discussed in the Section 3 and the obtained results are discussed in the Section 4 along with the comparative analysis and finally the research conclusion is presented in Section 5

2. RECENT RESEARCH REVIEWS

Acoustic background, content type, emotion display approach (acted vs. genuine), and other factors affect deep learning models' ability to recognize emotions in six emotion-laden speech corpora. IEMOCAP learns from speech and writing to recognize emotions. Voice-to-text uses emoticons alone. Speech-only models dominated emotional corpora. Cross-corpus research shows spoken, and written languages adapt. Single-corpus model is stronger [4].

HCI can't recognize speech emotions (HCI). As technology and knowledge of human emotions expand, it's vital to build accurate emotion detection systems for real-world applications to improve analytical capability and human-machine interfaces (HMI). MLMHFA and RNN decode a speaker's emotions (RNN). Word-process sounds. Open SMILE calculates MFCC. Time stamps assist RNNs self-focus. Emotion prediction uses multi-head attention. Combining MELD, CMU-MOSEI, and IEMOCAP enhances accuracy over using standalone models. Effective strategy [5].

Call centre operations, recommendation systems, and assistive technology use emotion recognition in user-generated material. Many systems and user interfaces are needed. Mixed Emotions Toolbox connects media and data. Video face identification and tracking, facial landmark localization, knowledge graph integration, audio emotion, age, and gender detection are being developed. This paper introduces and tests the Mixed Emotions Toolbox. [6] Smart TVs, call centres, brand image.

GMM-DNN detects human emotions in video using a deep neural network classifier and a cascaded Gaussian mixture model (GMM-DNN). GMM-DNN beats SVMs and sequential MLPs (MLP). Its 83.97% accuracy beats SVMs (80.33%) and SVM-based MLP (69.78%) Emotion recognition is best with hybrid classifiers. Human-like GMM-DNN results. Classifier [14-20] was tested using normal-volume and loud speech datasets. Signal overpowers noise [7].

Henceforth, after the detailed analysis of the existing systems, it is successful into detecting and separating the emojis, however the same detected emojis are not translated to sentiment scores[21-23]. In the preceding section the persisting research problems are furnished.

3. PROPOSED ALGORITHMS

Further, after the detailed discussion on the proposed methods using the mathematical modelling method, in this section the proposed algorithms are furnished. Firstly, the Topic Inference Based Text Corpus Extraction (TI-TCE) Algorithm is discussed.

Algorithm - I: Topic Inference Based Text Corpus Extraction (**TI-TCE**) Algorithm

Input: Text Dataset as DS1[], Topic as X

Output: Extracted Text as DS2[]

Process:

```
Step - 1.  Accept the topic for extraction as X
Step - 2.  Build the topic dictionary as X[t,R[]]
Step - 3.  For each topic in X[] as X[].t[i]
               a.  Inference the relation with other topics
               b.  R[] = X[].t[i] && X[].t[i+j]
Step - 4.  For each member in the intial dataset as DS1[j]
               a.  Build the new dataset as DS2[] = Filter for
                   R[]{DS1[]}
Step - 5.  Return DS2[]
```

Data mining methods, such as link and association analysis, visualization, and predictive analytics are all a part of the text analysis process. Other methods include retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, and visualization. The ultimate aim is to use natural language processing (NLP), various algorithms, and analytical techniques to transform text into data for study. Interpreting the data is a crucial part of this procedure. Secondly, the Dictionary Based Emoji Replacement (DER) Algorithm is discussed.

```
Algorithm - II: Dictionary Based Emoji Replacement (DER) Algorithm
Input: Topic Modelled Extracted Text as DS2[]
Output: Emoji Replaced Text Corpus as DS3[]
Process:
Step - 1.  Load the dataset as DS2[]
Step - 2.  Detect the emojis as K[]
Step - 3.  Build the text replacement as KT[]
Step - 4.  For each member of the dataset as DS2[i]
        a.  Build the replaced dataset, DS3[i] as DS2[i] Intersection KT[]
Step - 5.  Return DS3[]
```

Emoji are often misconstrued, as shown by studies. The receiver's interpretation of the emoji's design may play a role in the confusion, while differences in emoji display may also play a role. In the first place, there's the problem of how the emoji is understood in different cultural settings. It's possible that the recipient won't have the same mental image that the sender had when they chose a certain emoji.A smiley face, for instance, can be used to convey a demeaning, mocking, or even obnoxious attitude thanks to a system developed by the Chinese. This system relies on the fact that the emoji's orbicularis oculi (the muscle near the upper eye corner) remains immobile while the orbicularis oris (the one near the mouth) tightens, which is thought to be a sign of suppressing a smile. Thirdly, the Reduced Featuring Driven Text Translation (RFDT) Algorithm is discussed.

```
Algorithm - III: Reduced Featuring Driven Text Translation (RFDT) Algorithm
Input: Refined Text as DS3[]
Output: Translated Text as DS3[]
Process:
Step-1. Build the Unicode driven dictionary as D[]
Step-2. Load the dataset as DS3[]
Step-3. For each member in the data collection as DS3[i]
        a.  If DS3[i] is Not English,
        b.  Then, Build the translated text as X is DS3[i] = X
        c.  Else, Ignore
Step-4. Return DS3[]
```

Machine translation (MT) is the process of translating text automatically between two languages using a computer software. The truth is, however, that human intervention—in the form of pre- and post-editing—is usually necessary for machine translation. Commercial machine-translation tools can yield useful results with proper terminology work, preparation of the source text for machine translation (pre-editing), and reworking of the machine translation by a human translator (post-editing).

Finally, the Mutual Exclusion Based Emotion Extraction (ME-EE) Algorithm is discussed.

```
Algorithm - IV: Mutual Exclusion Based Emotion Extraction (ME-EE) Algorithm
Input: Refined Text as DS3[]
Output: Extracted Emotion as S[]
Process:
Step - 1.  Load the dataset as DS3[]
Step - 2.  For each member in the dataset as DS3[i]
        a.  Extract the text emotion as X[i]
        b.  Extract the emoji converted text emotion as Y[i]
        c.  Calculate the final emotion as S[i] = Mean{X[i],Y[i]}
Step - 3.  Return S[]
```

The proposed framework is furnished here

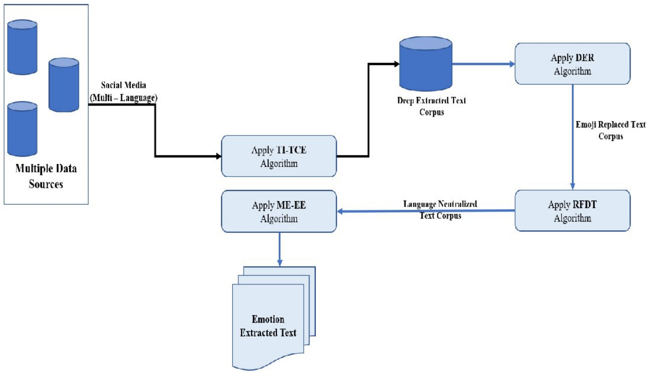

Figure 1: Proposed Framework

Further, in the next section of the work, the obtained results are discussed.

4. RESULTS AND DISCUSSIONS

After the analysis of the existing system and the proposed system, in this section of the work, the obtained results are discussed. This is the sentiment140 dataset. It contains 16,000 tweets extracted using the twitter api. The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment. The framework is tested on all the data items, however, only 20 for each dataset is furnished here.

The mean improvement for the 20 samples is 195% and for the complete dataset the mean improvement is nearly 320%. The overall mean time is 2.89 ns. The result is visualized graphically here.

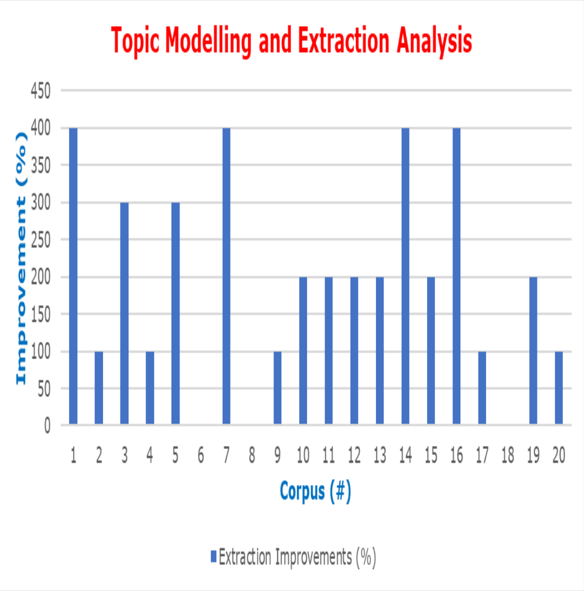

Figure 2: Topic Modelling and Text Extraction Accuracy

Thirdly, the emoji extraction process outcomes are furnished. The mean accuracy for the 20 samples is 92% and for the complete dataset the mean improvement is nearly 96%. The overall mean time is 2.84 ns. The result is visualized graphically here

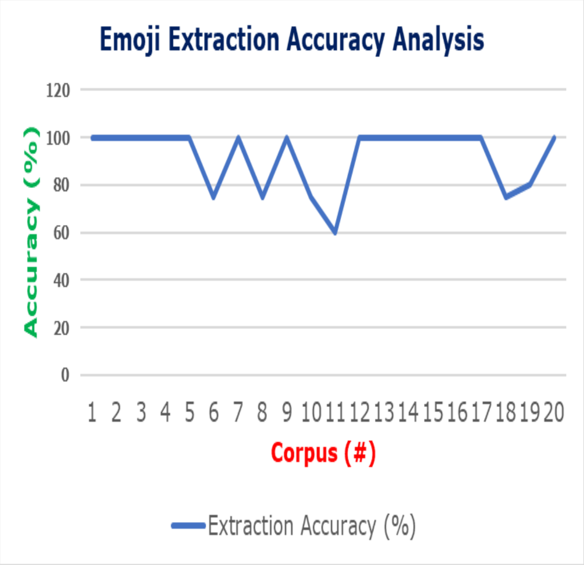

Figure 3: Emoji Extraction Accuracy Analysis

Fourthly, the emoji translation process outcomes demonstrate the mean accuracy for the 20 samples is 82.08% and for the complete dataset the mean improvement is nearly 85%. The overall mean time is 2.79 ns. The result is visualized graphically here.

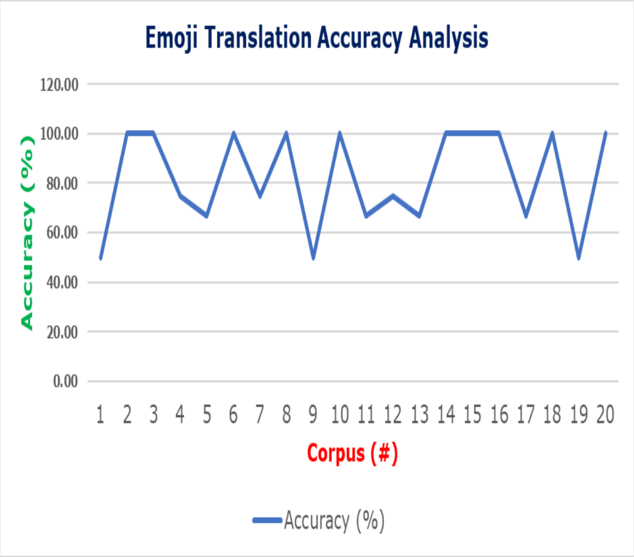

Figure 4: Emoji Translation Accuracy Analysis

Fifthly, the Sentiment extraction process outcomes demonstrate the overall mean time as 2.84 ns. The result is visualized graphically here.

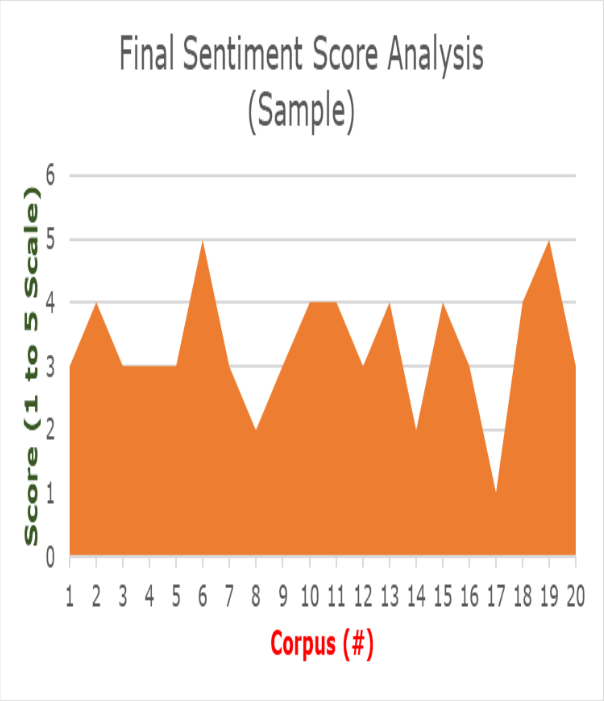

Figure 5: Sentiment Analysis Process samples

Finally, the overall time complexity is analyzed and the mean overall time for the complete process is 11.384 ns in the linear time complexity. The result is visualized graphically here

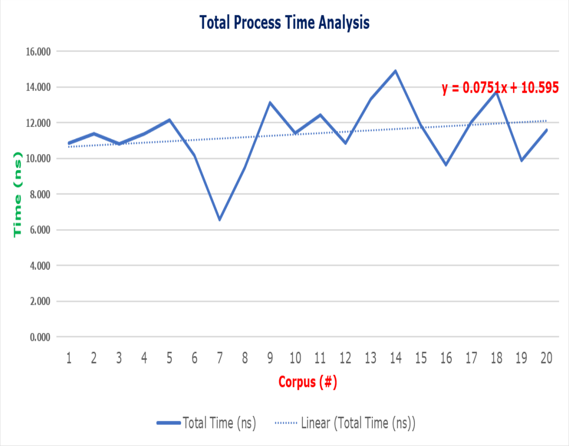

Figure 6: Total Process Time Analysis

Further, the obtained results are compared with the other parallel recent research outcomes in the next section of this work.

5. COMPARATIVE ANALYSIS

After the detailed discussions on the proposed algorithms and the outcomes from the algorithms, in this section of the work, the proposed work is compared with the other existing works as depicted below

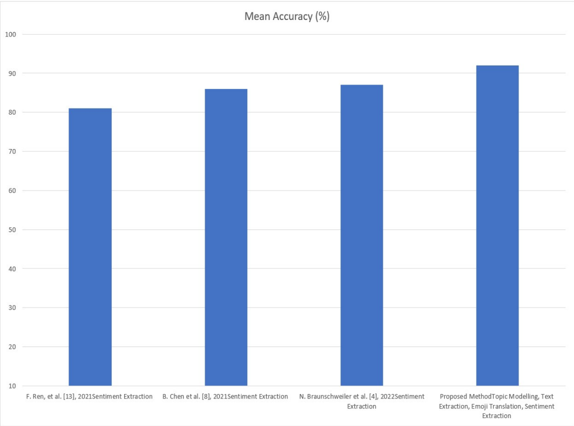

Figure 7: Expected Mean Accuracy of proposed system compared with existing methodologies.

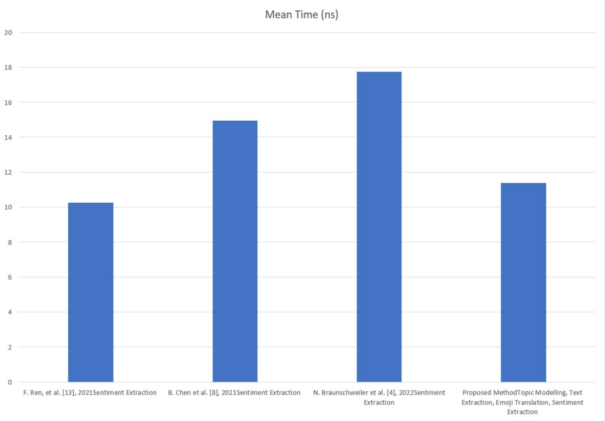

Figure 8: Expected Mean Time Analysis of proposed system compared with existing methodologies.

Thus, it is natural to realize that the proposed work has outperformed the parallel research outcomes by the means of features, accuracy and time complexity.

Henceforth, in the next section of the work, the research conclusion is furnished.

6. CONCLUSION

In recent years, there has been a significant surge in interest in the practice of extracting feelings from written corpora. The emotions that are collected from people's postings on social media are utilized for a variety of reasons, including customer evaluations of items, suggestions of literature and film, and the construction of opinion polls based on those posts. In the most recent few years, there has been a significant amount of study that has resulted in findings. Despite this, the currently available solutions have not been able to tackle notably two main problems to any substantial degree. To begin, the text corpus that is accessible from a variety of sources is written in a variety of languages. During the process of extracting emotions, these differences in language might be very difficult to extract the appropriate emotions because of the dependent on regional languages. Second, the proliferation of the use of emojis in text corpuses has made the task of emotion extraction even more difficult. This is because the combination of emojis and text can sometimes reflect sarcasm, and the detection of sarcasm can be a difficult and time-consuming process. Emojis have also become increasingly popular in recent years. This work proposes an automated framework to build a sentiment extraction process with pre-processing of the text corpus for the purpose of normalization of the text by removing emojis, sarcasm, and multiple local language influences. Henceforth, this work proposes an automated framework to build a sentiment extraction process. The framework leads to a reduction of 10% in the time complexity of emotion detection while leading to an increase of 80% in terms of accuracy when compared to the proposed machine learning algorithms.

REFERENCES:

[1] C. Liu, T. Liu, S. Yang and Y. Du, "Individual Emotion Recognition Approach Combined Gated Recurrent Unit With Emoticon Distribution Model," in IEEE Access, vol. 9, pp. 163542-163553, 2021.

[2] X. Wang, L. Kou, V. Sugumaran, X. Luo and H. Zhang, "Emotion Correlation Mining Through Deep Learning Models on Natural Language Text," in IEEE Transactions on Cybernetics, vol. 51, no. 9, pp. 4400-4413, Sept. 2021.

[3] E. Batbaatar, M. Li and K. H. Ryu, "Semantic-Emotion Neural Network for Emotion Recognition From Text," in IEEE Access, vol. 7, pp. 111866-111878, 2019.

[4] N. Braunschweiler, R. Doddipatla, S. Keizer and S. Stoyanchev, "Factors in Emotion Recognition With Deep Learning Models Using Speech and Text on Multiple Corpora," in IEEE Signal Processing Letters, vol. 29, pp. 722-726, 2022.

[5] N. -H. Ho, H. -J. Yang, S. -H. Kim and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," in IEEE Access, vol. 8, pp. 61672-61686, 2020.

[6] P. Buitelaar et al., "MixedEmotions: An Open-Source Toolbox for Multimodal Emotion Analysis," in IEEE Transactions on Multimedia, vol. 20, no. 9, pp. 2454-2465, Sept. 2018.

[7] I. Shahin, A. B. Nassif and S. Hamsa, "Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network," in IEEE Access, vol. 7, pp. 26777-26787, 2019.

[8] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu and D. Zhang, "Multimodal Emotion Recognition With Temporal and Semantic Consistency," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3592-3603, 2021.

[9] B. Xu, Y. Fu, Y. -G. Jiang, B. Li and L. Sigal, "Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization," in IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 255-270, 1 April-June 2018.

[10] N. Alswaidan and M. E. B. Menai, "Hybrid Feature Model for Emotion Recognition in Arabic Text," in IEEE Access, vol. 8, pp. 37843-37854, 2020.

[11] S. Hamsa, Y. Iraqi, I. Shahin and N. Werghi, "An Enhanced Emotion Recognition Algorithm Using Pitch Correlogram, Deep Sparse Matrix Representation and Random Forest Classifier," in IEEE Access, vol. 9, pp. 87995-88010, 2021.

[12] X. Kang, F. Ren and Y. Wu, "Exploring latent semantic information for textual emotion recognition in blog articles," in IEEE/CAA Journal of Automatica Sinica, vol. 5, no. 1, pp. 204-216, Jan. 2018.

[13] F. Ren and T. She, "Utilizing External Knowledge to Enhance Semantics in Emotion Detection in Conversation," in IEEE Access, vol. 9, pp. 154947-154956, 2021.

[14] X. Yang, S. Feng, D. Wang and Y. Zhang, "Image-Text Multimodal Emotion Classification via Multi-View Attentional Network," in IEEE Transactions on Multimedia, vol. 23, pp. 4014-4026, 2021.

Twitter Sentiment140 dataset, URL - https://www.kaggle.com/datasets/kazanova/sentiment140

[15] Swarajya lakshmi v papineni, A.Mallikarjuna Reddy, Sudeepti yarlagadda , Snigdha Yarlagadda, Haritha Akkineni "An Extensive Analytical Approach on Human Resources using Random Forest Algorithm" International Journal of Engineering Trends and Technology 69.5(2021):119-127.

[16] A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha, “Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)”, Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 4-Regular Issue, 2018.

[17] Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G., et al., (2019), “Generating cancelable fingerprint template using triangular structures”, Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers

5-6, pp. 1951-1955(5), doi: https://doi.org/10.1166/jctn.2019.7830.

[18] Mallikarjuna Reddy, A.,Venkata Krishna, V. and Sumalatha, L." Face recognition approaches: A survey" International Journal of Engineering and Technology (UAE), 4.6 Special Issue 6, volume number 7 , 117-121,2018.

[19] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, "Leveraging bio-maximum inverse rank method for iris and palm recognition", International Journal of Biometrics, 2022 Vol.14 No.3/4, pp.421 - 438, DOI: 10.1504/IJBM.2022.10048978.

[20] V. NavyaSree, Y. Surarchitha, A. M. Reddy, B. Devi Sree, A. Anuhya and H. Jabeen, "Predicting the Risk Factor of Kidney Disease using Meta Classifiers," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972392.

[21] B. H. Rao *et al.*, "MTESSERACT: An Application for Form Recognition in Courier Services," *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, 2022, pp. 848-853, doi: 10.1109/ICOSEC54921.2022.9952031.

[22] P. S. Silpa *et al.*, "Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm," *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, 2022, pp. 759-767, doi: 10.1109/ICOSEC54921.2022.9951883.

[23] B. S. Reddy, A. Mallikarjuna Reddy, M. H. D. S. Sradda, T. Mounika, S. Mounika and M. K, "A Comparative Study on Object Detection Using Retinanet," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972742.