I.J. Modern Education and Computer Science, **2023, 5, 28-41**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2023.05.03

Enhancing Emotion Detection with Adversarial Transfer Learning in Text Classification

Ashritha R Murthy*
Department of Computer Science, Sri Jayachamarajendra College of Engineering, JSSS&TU, Mysuru, India
E-mail: ashrithar.murthy@sjce.ac.in
ORCID iD: https://orcid.org/0000-0001-7827-0914
*Corresponding Author

Anil Kumar K M
Department of Computer Science, Sri Jayachamarajendra College of Engineering, JSSS&TU, Mysuru, India
E-mail: anilkm@sjce.ac.in
ORCID iD: https://orcid.org/0000-0002-3236-1500

Abdulbasit A. Darem
Department of Computer Science, Northern Border University, Saudi Arabia
E-mail: basit.darem@nbu.edu.sa
ORCID iD: https://orcid.org/0000-0002-5650-1838

Received: 15 June, 2023; Revised: 14 July, 2023; Accepted: 17 August, 2023; Published: 08 October, 2023

Abstract: Emotion detection in text-based content, such as opinions, comments, and textual interactions, holds pivotal significance in enabling computers to comprehend human emotions. This symbiotic understanding between machines and human languages, powered by technological advancements like Natural Language Processing and artificial intelligence, has revolutionized the dynamics of human-computer interaction. The complexity of emotion detection, although challenging, has surged in importance across diverse domains, encompassing customer service, healthcare, and surveillance of social media interactions. Within the realm of text analysis, the quest for accurate emotion detection necessitates a profound exploration of cutting-edge methodologies. This pursuit is further intensified by the imperative to fortify models against adversarial attacks, a pressing concern in deep learning-based approaches. To address this critical challenge, this paper introduces a pioneering technique—adversarial transfer learning—specifically tailored for emotion classification in text analysis. By infusing adversarial training into the model architecture, the proposed approach emerges a solution that not only mitigates the vulnerabilities of existing methods but also fortifies the model against adversarial intrusions. In realizing the potential of the proposed approach, a diverse array of datasets is harnessed for comprehensive training. The empirical results vividly demonstrate the efficacy of this approach, showcasing its superior performance when compared to state-of-the-art methodologies. Notably, the suggested approach yields in advancements in classification accuracy. In particular, the deployment of the Adversarial transfer learning methodology has increased in accuracy of 17.35%. This study, therefore, encapsulates a dual achievement: the introduction of an innovative approach that leverages adversarial transfer learning for emotion classification, and the subsequent empirical validation of its unparalleled efficiency. The implications reverberate across multiple sectors, extending the horizons of accurate emotion detection and laying a foundation for the next stride in human-computer interaction and emotion analysis.

Index Terms: Emotion detection, Natural Language Processing, Adversarial transfer learning, Text analysis, Convolutional neural networks, Deep learning.

1. Introduction

The sheer volume of data on the internet today has increased, from terabytes to petabytes, due to constantly being added to available data. The definition of "emotions" is an intricate interplay of various subjective and objective factors, which are influenced by both neurological and hormonal systems. Emotions encompass an individual's personal experience and interpretation of the feeling, as well as external factors that may trigger or influence the emotion.

In the brain, emotions are processed through a network of interconnected regions, such as the amygdala, prefrontal cortex, and insula. Additionally, hormonal systems, such as the release of cortisol and adrenaline, can impact emotional responses, particularly during stressful situations. Overall, emotions are a complex phenomenon involving subjective and objective factors, along with neurological and hormonal processes [1].

Emotions play a crucial role in cognition and behavior, triggering cognitive processes and physiological changes in response to external stimuli. They can be divided into categories that are positive, neutral and negative. Emotion recognition from web text documents poses challenges due to the absence of explicit emotional terms. Emotions and sentiment are frequently used interchangeably. Emotions in web text documents are categorized using text classification, which uses supervised machine learning techniques like decision trees,SVM and deep learning methods like LSTM. The use of adversarial methods aids in generalizing classifiers across different domains. Paul Ekman identified emotions like surprise, anger, joy, disgust, fear, and sadness[2]. Emotions in interpersonal relationships are crucial for communication [3]. Emotions in written text, audio, and video can be recognized in computer science[4]. Wen and Wan used a labeled dataset to classify positive and negative emotions based on a threshold value [5]. The dimensional technique can handle complex emotional situations, whereas the categorical approach has limitations in capturing nuanced emotional experiences[6]. Computational techniques for emotion extraction include rule-based, ML, keyword-based, corpus-based, and DL methods [7,8].

ML algorithms, particularly supervised learning, are commonly used for emotion classification[9]. Emotion recognition from web documents using machine learning-based methods is a significant area of research [10]. Emotion classes are assigned to text using both supervised and unsupervised learning algorithms [11]. Deep learning (DL) techniques allow algorithms to learn hierarchically, improving their ability to grasp new concepts [12]. This has a variety of real-world uses, such as analyzing patient mental health in the healthcare industry, understanding customer emotions for business strategies on social media sites like Twitter, Facebook, and YouTube, and recognizing problems like fake news and hate speech through emotion detection in text. Decision trees , support vector machines, random forests, convolutional neural networks, bi-directional LSTM, and long short-term memory are notable text categorization techniques utilized in research [12].

Adversarial training helps the classifier become more resilient, decreases sensitivity to domain-specific noise or biases, and increases accuracy on target domains where labeled data may be hard to come by or expensive to acquire. Additionally, adversarial approaches provide higher generalization and transfer learning capabilities by solving the domain shift problem, which eventually results in superior text classification performance in real-world contexts. In the paper, we employ Adversarial Transfer Learning and CNN for the classification of web text documents.

The structure of the work presented in this paper is mentioned as containing emotional and non-emotional documents as follows. The earlier studies on emotion and emotion categorization are discussed in Section 2. In Section 3, we provide a description of the relevant methodology and system architecture of our suggested models. The experimental results of the suggested models are also shown in Section 4. Section 5 addresses the conclusion that follows.

2. Background

The study of text-based emotion recognition has become important research work. In psychological research, there are prevalent models related to emotions, such as the categorical emotion model or discrete model, the dimensional model, and the appraisal-based model [13].

In the brain, multiple neural subsystems correspond to various emotions. Six fundamental categories are used by Ekman's model to classify emotions [14,15]. Happy, anger, sadness, disgust, surprise, and fear are some examples of these fundamental feelings. The eight emotions in Plutchik's model are listed as joy vs. sad, contempt vs trust, fear vs. anger, and anticipation vs surprise [16]. The dimensional model perceives emotional states as interconnected rather than separate from one another. The model is represented in dimensional space, i.e., unidimensional and multidimensional, expressing the relation between emotions and an event depending on the intensity (low to high) of the emotions. Many dimensional emotion models employ two or three dimensions: "valence" (which shows the emotions positive or negative), arousal which represents the intensity of enthusiasm of emotion, and dominance, a measure of how well one can manage emotions [17].

In the field of emotion encoding, Russell developed the circumplex of effect, a significant two-dimensional model that forms an emotional-wheel with valence depicted on the axis that is vertical and arousal represented in horizontal axis. [18]. This model is essential in understanding dimensional aspects of emotions. The dimensional model, incorporating valence and arousal, can be combined with the appraisal-based model to enrich emotion analysis [19]. This integration allows for a broader methodology for studying emotions. Componential emotion models are integrated based on the evaluation concept [20]. The appraisal hypothesis explains how the same experience can trigger multiple emotions in different individuals and at different times [21]. Understanding the various factors influencing emotions, such as knowledge, communication, physiology, motivation, movement, and responses, is crucial [5].

Hai Huan et. al [22] proposed an improved text classification technique using the CBM (Convolutional and Bi-LSTM Models) approach. The CBM approach utilizes both global and local semantic features extracted from the text by employing the Glove model for text vectorization in the embedding layer. Basiri, M. E. et al. [23] introduced the

ABCDM model, an Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. The model utilizes pre-trained Glove word embedding vectors as initial weights for the embedding layer. It employs bidirectional LSTM and GRU. Thi-Thanh-Thuy Huynh and Anh-Cuong Le [24] proposed a CNN-based framework for emotion categorization. They addressed the challenge of representing emotions in deep learning methods by incorporating pre-defined grammatical patterns containing emotional data to extract external features. Experimental results on Vietnamese emotion and ISEAR1 datasets demonstrated the success of their approach. Pradhan et al. [25] presented CLAVER, an attention mechanism-based system for recommending scholarly venues. The model utilized a combination of convolutional layers, LSTM, and bi-LSTM to make recommendations solely based on the title and abstract. Experimental results on the DBLP dataset showed superior performance compared to state-of-the-art techniques in terms of precision@k, nDCG@k, accuracy, MRR, and diversity and stability of recommended venues. Liang et al. [26] proposed an improved Double Channel (DC) approach to enhance the performance of CNN-LSTM models. The DC approach incorporated word-level and char-level embeddings simultaneously on separate channels. The authors introduced a hybrid attention mechanism and trade-off learning to improve the model's generalizability. Experimental results demonstrated that the proposed DC CNN-LSTM model outperformed the basic CNN-LSTM algorithm in terms of accuracy and F1 measure.

Deepanway Ghosal et al. [27] introduced the Dialogue Graph Convolutional Network (DialogueGCN) to improve context awareness for word-uttered emotion recognition in dialogues. DialogueGCN achieved better performance than strong baselines and current state-of-the-art techniques on three ERC benchmark datasets, achieving a weighted average accuracy of 65.25% for identifying emotions in conversational dialogues. Trueman and Cambria [28] proposed a convolutional stacked bidirectional LSTM model with a multiplicative attention technique for aspect categorization and sentiment identification tasks. The model incorporated a convolutional layer to extract high-level aspects and emotional features, along with two bidirectional LSTM layers to regulate information flow. The fixed-width multiplicative attention technique was used to find context vectors for the input sequence, leading to improved performance compared to other methods on the SemEval-2015 and SemEval-2016 datasets.

Denis Eka Cahyani et al.[29] research is a comprehensive exploration into advancing the domain of emotion detection in textual content. The study hinges on the comparison of various word embedding techniques, including Word2Vec, BERT, and GloVe, within the context of Convolutional Neural Networks (CNN). The primary objective centers on identifying five core emotions - happiness, anger, sadness, fear, and surprise - present in text originating from diverse sources such as commuter line, transjakarta, and their amalgamation. The pivotal methodology involves integrating CNN with BERT embeddings, yielding highly promising results with accuracy scores of 86.47%, 87.23%, and 86.18% across respective datasets.

Alsmadi et al. [30] provided an overview of adversarial attacks and defensive systems in social media applications. The review focused on major challenges, and research directions, and discussed applications such as sentiment analysis, hate speech detection, clickbait & spam identification, rumor detection, satire detection, and misinformation detection. The authors provided insights into the vulnerabilities of machine learning models in these domains and highlighted the need for robust defensive mechanisms to mitigate the impact of adversarial attacks. Hajek et al. [31] presented 2-deep neural network models for detecting fraudulent customer reviews. These models incorporated n-gram, skip-gram, and emotion models and demonstrated high effectiveness on larger datasets with mixed polarities. The researchers utilized various emotion representations and pre-trained word embeddings to enhance the models' accuracy, suggesting their applicability in real-world scenarios. Alsmadi et al. [32] and Han and Zhang [33] focused on adversarial attacks in text processing and sentiment analysis. Alsmadi et al. developed a pre-trained model to improve the resilience of text generation models against adversarial attacks. Han and Zhang [33] explored the use of adversarial training to enhance the robustness of emotion detection and sentiment analysis models. Both studies highlighted the significance of recognizing and addressing the vulnerabilities of machine learning models to adversarial attacks in these domains.

A few shortcomings were identified and listed as follows:

1. Lack of accuracy achieved in conventional construction models.
2. Limited consideration of adversarial attacks: The literature lacks effective strategies to address adversarial attacks on text documents, which can manipulate the decisions of machine or deep learning models and potentially compromise their performance and reliability.

This work proposes an approach to overcome the above shortcomings by adopting the following steps:

- Feature extraction using pre-trained weights from the GloVe model: To counteract the first identified limitation of accuracy in conventional models, we adopt a feature extraction technique that utilizes pre-trained GloVe embeddings. These embeddings, learned from extensive text corpora, infuse our model with a deeper understanding of semantic features and linguistic nuances. By freezing and transferring these enriched weights to a Convolutional Neural Network (CNN), our model assimilates knowledge from a large dataset, consequently amplifying its ability to capture essential features critical for accurate emotion classification.

- Integration of CNN architecture: Recognizing the second limitation of inadequate defense against adversarial attacks, our approach seamlessly incorporates a Convolutional Neural Network architecture. The CNN's inherent capability to detect local patterns equips our model with the necessary sensitivity to textual nuances. This aligns well with the complex nature of emotion expression within language.
- Learning from a large dataset and adopting adversarial transfer learning: Our model's training regime involves learning from a diverse and extensive dataset. This addresses the first limitation directly, allowing the model to better generalize its understanding of emotional nuances, resulting in improved accuracy in categorization. Moreover, to counteract the challenge of adversarial attacks, we integrate adversarial transfer learning. This strategic inclusion enhances the model's resilience and robustness, making it more reliable and dependable even in the presence of manipulative inputs.

By fusing these techniques, our proposed model not only addresses the shortcomings identified in existing approaches but also advances the field of emotion analysis in text classification. Our methodological choices are grounded in the necessity to bridge the gap between model accuracy, and real-world applicability. By enhancing the precision of emotion classification, fortifying against adversarial threats, and learning from diverse data, our approach contributes substantively to the state of knowledge in this domain."

Incorporating these points within your existing text will provide a more seamless and comprehensive flow that highlights the relevance of your work and the strategic choices you've made to address the identified limitations.

3. Methodology

The proposed methodology is intricately designed to effectively achieve the research objectives, which are centered around the accurate classification of text into emotional and non-emotional categories. Each facet of our methodology aligns with a specific research objective, synergistically contributing to the overarching goal of enhancing emotion analysis in text classification.

Objective 1: Enhancing Accuracy in Emotional Text Classification through Feature Extraction Using Pre-trained GloVe Weights

Our first objective revolves around rectifying the lack of accuracy achieved in conventional models when classifying text as emotional or non-emotional. To fulfill this, our methodology incorporates the utilization of pre-trained GloVe word embeddings for feature extraction. By tapping into the semantic richness of these embeddings, we bolster the model's feature representation capabilities, enabling it to capture intricate linguistic nuances that distinguish emotional and non-emotional content. This aligns perfectly with the objective of elevating classification accuracy for emotional and non-emotional text.

Objective 2: Improving Robustness against Adversarial Attacks in Emotion Classification

The second research objective aims to address the limited consideration of adversarial attacks in the classification of emotional and non-emotional text, a critical issue that can compromise model reliability. Our proposed methodology directly tackles this by incorporating adversarial transfer learning. By integrating this technique, our model gains the ability to recognize and effectively counteract adversarial manipulations that can mislead conventional models during emotion classification. The methodology enhances the model's robustness, thus fulfilling the objective of improving the model's ability to accurately classify emotional and non-emotional text.

Objective 3: Accurate Emotion Classification through CNN Architecture and Learning from Diverse Data

To meet the third research objective, we leverage the Convolutional Neural Network (CNN) architecture for its proven efficacy in capturing local patterns and intricate features in textual data. By adopting this architecture, our methodology aligns with the objective of accurately discerning the nuanced expressions of emotions, contributing to precise emotion classification for both emotional and non-emotional text. Additionally, our model is trained on a diverse and extensive dataset, a strategy that directly advances the objective of improving generalization and overall accuracy in emotion classification.

Collectively, our methodology synergistically integrates these strategies, each directly aligned with a specific research objective of accurate emotion classification. Through feature extraction using pre-trained GloVe weights, the integration of CNN architecture, and the utilization of adversarial transfer learning, our methodology strategically addresses the shortcomings and achieves the stated research objectives. By doing so, it culminates in a comprehensive approach that advances the field of emotion analysis in accurate text classification.

An overview of the proposed architecture of Adversarial transfer learning is shown in Fig. 1. The process begins with the collection of a dataset from a social media platform. Several preprocessing steps are applied to prepare the text for classification, including punctuation removal, tokenization, lowercasing, stop word removal, and stemming. Once the text has been preprocessed, feature extraction techniques are used to transform the text into numerical representations.

Two methods, namely one-hot encoding and GloVe embedding, are employed for feature extraction. One-hot encoding represents each word as a binary vector. Each dimension corresponds to a unique word in the dataset. On the other hand, GloVe embedding represents pre-trained word vectors to capture semantic information once the features are extracted. The weights of the GloVe embedding features are frozen. These frozen features, with one-hot encoded features and word embeddings, serve as inputs to a CNN algorithm. It utilizes the extracted features, including the frozen GloVe embedding features, to learn patterns and make accurate predictions based on the provided dataset.

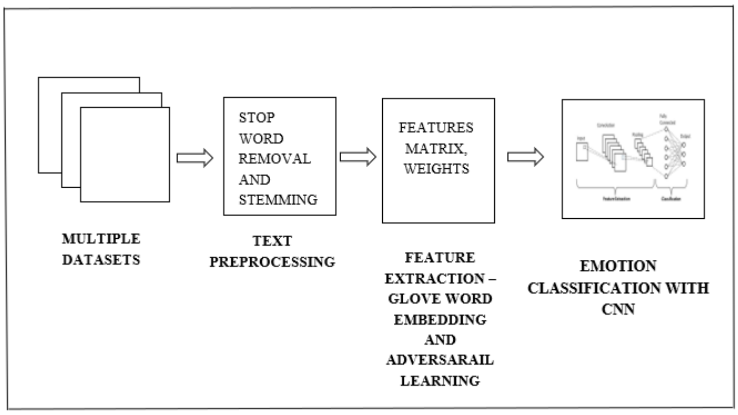
STOP
WORD
REMOVAL
AND
STEMMING
FEATURES
MATRIX,
WEIGHTS
MULTIPLE DATASETS
TEXT PREPROCESSING
FEATURE EXTRACTION – GLOVE WORD EMBEDDING AND ADVERSARAIL LEARNING
EMOTION CLASSIFICATION WITH CNN

Fig. 1. Proposed Adversarial transfer learning method for classification of text.

3.1. *Dataset*

The embedding model used in this work is using training multiple datasets to ensure a comprehensive understanding of emotions. These datasets include the go-emotions dataset [34], the Twitter dataset [35], the Emotion Detection from Text dataset [36], and the Amazon review data [22]. By incorporating diverse datasets are used to train the model to make the model robust and resilient. The go-emotions dataset consists of 58,009 sentences extracted from Reddit, representing both positive and negative emotions in a well-balanced manner. It covers 27 different emotions. The Twitter dataset, which comprises 1.2 million tweets, offers a broad collection of emotions, provides valuable insights for sentiment analysis. The Emotion Detection from Text dataset includes 40,001 sentences with labeled emotions. The Amazon review dataset contains 3.4 million reviews and offers a wide range of emotions expressed by customers. These datasets were carefully divided into training and testing splits to evaluate the model's performance are 50% - 50%, 60% - 40% and 80% - 20% as shown in Table 4.

3.2 *Text Pre-processing*

The text preprocessing steps include tokenization, removal of prepositions, determiners, punctuation, and stop words. Tokenization facilitated independent analysis of individual words, which is vital for subsequent emotion detection tasks. The prepositions, determiners, and punctuation marks were removed to reduce noise and irrelevant elements to identify meaningful sentences. These elements typically contribute little semantic meaning in the context of emotion detection and were deemed unnecessary for accurate analysis. Additionally, stop words, lacks specific semantic value, were removed to streamline the data and enhance analysis efficiency. Porter's stemming is applied to reduce words to their base standardizing the text's terms and emotion category mapping and analysis is done by standardizing the dataset. The number of stemmed words that are associated with each category is calculated using Eq.(1). This pre-processing technique, transforms raw text data into a cleaner and more manageable format to extract meaningful insights from the text.

$$S_w = S_{w(t)} * P_{(w)} \tag{1}$$

Where,
S_w - No. of stemmed words
$S_{w(t)}$ - No. of stemmed words in text
$P_{(w)}$ - probability of word

3.3 *Feature Extraction and Classification*

This section discusses in detail the feature extraction methods phases to classify the text as emotional and non-emotional using Adversarial transfer learning.

3.3.1 *Feature Extraction using Adversarial transfer learning:*

The first step in this process involves representing words using one-hot encoding. Each word is represented by a vector with a dimension for every possible word in the vocabulary. In a one-hot vector, the dimension corresponding to the word is set to 1, while all other dimensions are set to 0. The GloVe model is used to map words to dense vectors in a high-dimensional space. These vectors capture semantic information by encoding relationships between words. The GloVe model is trained on large text corpora, allowing it to capture and incorporate rich semantic context. By utilizing the GloVe model, raw text phrases are transformed into numerical vectors that encompass the semantic context of the words. This conversion enables more meaningful analysis and interpretation of the text, facilitating tasks such as sentiment analysis and emotion detection.

The computational steps involved in developing the Adversarial transfer learning approach:

1. Every word in the input text is represented by a sparse binary vector with a vocabulary-sized dimension.
2. With the exception of the index corresponding to the word, which is set to one, every element in this vector is zero. This representation allows us to identify each word uniquely such as: one-hot_vector(word) = [0, 0, ..., 1, ..., 0].
3. One-hot encoded word vectors are mapped into the d-dimensional space of the GloVe embeddings. By multiplying the one-hot vectors with the pre-trained word embedding matrix as in Eq.(2). The resulting vectors are dense, captures the semantic information of each word in the input text.

$$\text{word\_embedding(word)} = \text{GloVe\_matrix} * \text{one} - \text{hot\_vector(word)} \quad (2)$$

Step 1: Computing Glove embedding matrix

Given a corpus of emotion-labeled text, a co-occurrence matrix is constructed. Each element X(i, j) of the matrix represents the co-occurrence count of word i with word j in the corpus. This matrix captures the relationships between words based on their co- occurrence patterns.

Step 2: Initialization of Word Vectors

The word vectors (w_i) and context word vector ($\overline{w}_i$), are initialized with pre-trained Glove word embeddings. These vectors are high-dimensional, dense representations of the words.

Step 3: Objective Function

The objective function is applied to an objective function as shown in Eq.(3):

$$J = \sum_{ij} f\big(X(i,j)\big)\,(w_i^T\, w_j^T + b_i b_j - \log\big(X(i,j)\big)2\ \mathrm{E(i,j)} \quad (3)$$

where:

J - cost function

$X(i,j)$ - co-occurrence count between word i and word j

w_i^T - target word vector for word i

w_j^T- context word vector for word j

$b_i b_j$ - biases

$\mathrm{E(i,j)}$ - binary indicator to determine words associated with emotion.

$f\big(X(i,j)\big)$ - weighting function

Step 4: Adam optimization function

The word vectors and biases are adjusted and minimized the objective function.

Step 5: Adversarial transfer learning using the Glove model

The model is trained using the co-occurrence matrix and the modified cost function. Through iterative optimization, the word vectors and biases are updated to capture the semantic relationships between words and emotions.

Step 6: Embedding Matrix

In this, the optimized word vectors are extracted and used to construct an embedding matrix. Each row of the embedding matrix corresponds to a word and its associated vector. These vectors capture the learned semantic representations of words in the context of emotions. Then the weights learned are frozen.

4. The feature extraction and word embedding for the sample input text :

1): *"I feel happy and excited."* 2): *“I feel sad and disappointed.”*

For the above feature vector are computed using Eq.(4). The vector values are as shown in Table 1 and Table 2 represents for positive and negative emotions.

$$F_v = W_{e1} \oplus W_{e2} \oplus W_{e3} \ldots \oplus W_{e(n-1)} \oplus W_{e(n)} \quad (4)$$

Here, W_{e1}, W_{e2}, ..., $W_{e(n)}$ represent the word embeddings of individual words in the input text.

Table 1. Matrix Representation of Vectors for Positive Phrase.

	One-hot	GloVe	Word
Positive	[1, 0, 0, ..., 0]	[0.249, 0.785, -0.539, ..., 0.923]	"I"
	[0, 1, 0, ..., 0]	[0.155, 0.614, 0.384, ..., 0.868]	"feel"
	[0, 0, 1, ..., 0]	[0.096, 0.486, 0.623, ..., 0.795]	"happy"
	[0, 0, 0, 1, ..., 0]	[0.433, 0.485, 0.693, ..., 0.823]	"and"
	[0, 0, 0, 0, 1, ..., 0]	[-0.418, 0.298, -0.921, ..., 0.729]	"excited"

Tabel 2. Matrix Representation of Vectors for Negative Phrase

Negative	**One-hot**	**GloVe**	**Word**
	[1, 0, 0, ..., 0]	[0.249, 0.785, -0.539, ..., 0.923]	"I"
	[0, 1, 0, ..., 0]	[0.155, 0.614, 0.384, ..., 0.868]	"feel"
	[0, 0, 1, ..., 0]	[-0.182, -0.485, 0.615, ..., -0.792]	"sad"
	[0, 0, 0, 1, ..., 0]	[0.433, 0.485, 0.693, ..., 0.823]	"and"
	[0, 0, 0, 0, 1, ..., 0]	[0.454, -0.298, 0.919, ..., -0.728]	"disappointed"

The feature vector is formed by concatenating the GloVe embeddings are calculated using Eq.(5).

$$feature\_vector = concat\ (GloVe\_embeddings(sentence)) \tag{5}$$

3.3.2. CNN Architecture for classification of emotional text:

The fully connected layer convolutional layer is the two primary components of the CNN model. The convolutional layer is responsible for detecting local patterns in the input feature vector, while the fully connected layer learns the relationships between the extracted features and the classes. The convolutional layer in the CNN model applies filters to the feature vector, generating feature maps highlighting specific local patterns. These filters are learned during the training process to build the model to identify the input patterns. Pooling layers further contribute to the model's effectiveness by reducing the spatial dimensions by retaining essential information. This helps in condensing the extracted features of the dataset samples.

Convolutional Neural Network (CNN) model with pre-trained GloVe word embeddings with softmax classifier is used to classify the emotional text. The use of GloVe word embeddings allows the model to capture the semantic relationships between words. The words in dense vectors represents a high-dimensional space, GloVe embeddings enable the model to understand the context and meaning of the input text. Adam (Adaptive Moment Estimation) is used to optimize the algorithm in deep learning to updates the weights of a neural network during training [35]. Momentum and adaptive learning rates are used to accelerate the convergence of the model. Eq. 6 updates the weights of the network based on gradients and adaptive learning rates. It adjusts the learning rate for each parameter individually, enhancing the optimization process. Also Relu function [37] is adopted to map the negative values to ‘0’ is given from Eq.(11).

The Adam updates are calculated using Eq.(6):

$$\theta_t = \theta_{t-1} - \alpha * \hat{n}_t / \ (\sqrt{\hat{s}_t} + \varepsilon) \tag{6}$$

Where,
n_t & s_t = first and second moments

$$n_t = B1 * n_{\ t-1} + (1 - B1) * gr_t \tag{7}$$

$$s_t = B2 * s_{\ t-1} + (1 - B2) * {gr_t}^2 \tag{8}$$

gr_t = gradient of the weights at time t.

$$\hat{n}_t = n_t / \ (1 - B1^t) \tag{9}$$

$$\hat{s}_t = s_t \ / \ (1 - B2^t) \tag{10}$$

$B1$& $B2$ = decay rates: 1^{st} and 2^{nd} moments.
α = learning rate, which determines the step size for weight updates.
θ_t = updated weights at time t.
ε = small constant (e.g., 10^{-8}) added for numerical stability.

$$f(x) = \max(0, x)) \tag{11}$$

Input layer: By utilizing the Adam optimizer and the ReLU activation function in the training process, the emotion classification model can effectively optimize the network parameters and introduces non-linearity, for features to improve the performance and accuracy.

Output layer: In order to determine the distribution of probabilities for both emotional classes and non-emotional classes, the softmax function [35] is used to the CNN's final layer. The model provides probabilities that add up to 1, offering a trustworthy measure of confidence for each class prediction using Eq. (12), by applying the softmax function to the output scores.

To classify softmax is represented as follows:

$$\mathrm{P}(\mathrm{emotion}_i)|text| = \frac{exp_{(F-1)}}{sum_j\ exp_{(F-1)}} \tag{12}$$

Where,

$sum_j\ exp_{(F-1)}$ - sum of the exponentials of all the feature vectors.

Collectively, these steps enable the CNN model to process the text data, extract meaningful features using GloVe embeddings, identify local patterns and dependencies, to classify the text into emotional and non-emotional classes. The combination CNN and Glove embedding techniques ensures accurate emotion classification by leveraging the semantic information embedded in the text.

The following procedure is adopted to construct the CNN model at various layers [Eq.(13) – Eq.(10)]:

i) Convolutional layer:

The filters are learned during training and are applied to the feature vector to generate feature maps. Each filter represents a specific pattern or feature that the model identifies in the input.

$$\mathrm{feature\_maps} = \mathrm{convolve(feature\_vector, filters)} \tag{13}$$

ii) Max-pooling layer:

By using an optimization function, this layer decreases the spatial dimensions of the feature maps.

$$\mathrm{pooled\_feature\_maps} = \mathrm{max\_pooling(feature\_maps)} \tag{14}$$

iii) Fully connected layer:

This layer learns the relationships between the extracted features from the pooled feature maps and the classes. It computes scores or activations that indicate the model's confidence in each class.

$$\mathrm{class_{scores}} = \mathrm{fully\_connected(pooled\_feature\_maps)} \tag{15}$$

iv) Classification: activation function Softmax:

The softmax function converts the class scores into probabilities, providing a normalized distribution representing the model's predictions for the input text.

$$Probabilities = \mathrm{softmax(class\_scores)} \tag{16}$$

3.4 Algorithm

The brief algorithm of the proposed approach encloses all the steps discussed above:

Step 1: Text Preprocessing
- Input: "I am feeling happy today!"
- Tokenize the text data: ["I", "am", "feeling", "happy", "today"]
- Perform data cleaning: ["feeling", "happy", "today"]
- Remove stop words: ["feeling", "happy", "today"]

Step 2: Dataset Split : The dataset split - 20% testing and 80% for training.

Step 3: Build Vocabulary (Bag-of-Words)
- Vocabulary: ["feeling", "happy", "today", "emotions"]

Step 4: Word Tokenization and Frequency Calculation
- Tokens: [["I", "am", "feeling", "happy", "today"],
["I", "don't", "have", "any", "emotions"]]
- Word Frequency: {"I": 2, "am": 1, "feeling": 1, "happy": 1, "today": 1, "don't": 1, "have": 1, "any": 1, "emotions": 1}

Step 5: GloVe Word Embeddings
- Sample GloVe embedding: ["happy": [0.1, 0.2, 0.3, ...], "sad": [0.4, 0.5, 0.6, ...], ...]

Step 6: Co-occurrence Matrix and GloVe Probability Ratio
- Co-occurrence matrix example:

	feeling	happy	today	emotions
feeling	0	1	1	0
happy	1	0	1	0
today	1	1	0	0
emotions	0	0	0	0

Step 7: Embedding Matrix Construction
- Embedding Matrix:

Embedding	Vector
feeling	[0.2, 0.3, 0.4, ...]
happy	[0.1, 0.2, 0.3, ...]
today	[0.5, 0.6, 0.7, ...]
emotions	[0.0, 0.0, 0.0, ...]

Step 8: Adversarial Training from Glove embedding model frozen the weights and passed to the classification model CNN.

Step 9: CNN Training
- Design the CNN model architecture along with softmax activation for classification.
- Initialize the weights from the Adversarial transfer learning to CNN model and obtain prediction.
- Calculate the difference in value between the expected and actual labels..
- Perform backpropagation to calculate the gradients with respect to the weights.
- Update the weights of the model using an optimization algorithm (e.g., Adam).
- Repeat the training process for multiple epochs to improve the model's performance.

Step 10: Evaluation and Testing
Input: "I am feeling ecstatic after winning the competition!"
Predicted Output: Emotion (based on the trained model)
Input: "The weather today is quite pleasant."
Predicted Output: No emotion (based on the trained model)

4. Experimental Design and Implementation:

Python IDE 3.6 [22] is used for simulating the experiment. The 200-dimensional GloVe word vectors [38] are utilized for word embedding. Deep learning frameworks, specifically Keras and TensorFlow, are employed for model development and training. Keras and TensorFlow played a pivotal role in the research pipeline, offering a comprehensive set of tools and functionalities for effective model development and optimization. The results are evaluated by assessing various performance metrics and comparing the obtained results with those of several existing works.

4.1 Parameters used

Several evaluation criteria, including precision, accuracy, recall, and F1-score, are used to access the model's performances.

Precision(P): The ratio of true positives (TP) to the total of true positives and false positives (FP) is known as precision [39]. It is frequently used to assess a binary classification model's effectiveness. The following is how to express the accuracy formula:

$$P = \frac{TP}{TP+FP} \tag{17}$$

Recall(R) sensitivity: performance indicator that is frequently employed to assess binary classification models [41]. The percentage of actual positives that the algorithm accurately detected is what this metric measures. Calculating the ratio of true positives (TP) to the total of true positives and false negatives (FN) is how this is done. The recall formula is as follows:

$$R = \frac{TP}{TP+FN} \tag{18}$$

F1 Score: a performance metric for binary classification problems that offers a fair assessment of a classifier's recall and precision. It measures a model's capacity to accurately identify positive instances while reducing false positives and false negatives. It is the harmonic mean of accuracy and recall. Higher values of the F1 score indicate greater classifier performance, which runs from 0 to 1. It is especially useful in situations with unbalanced class distributions or where assessment criteria for precision and recall are equal. The following equation determines the F1 score:

$$\text{F1 Score} = 2 * \frac{\text{Precision}*\text{recall}}{\text{Precision}+\text{recall}} \tag{19}$$

Where,
TP -True positive, FN – False negative and FP – False positive.

4.2 Results and performance analysis:

We evaluated the precision, accuracy, recall, and F1-score to comprehensively accesses the effectiveness of our approach. The results, presented in Table 3, depict the performance of various approach applied to the challenge of the same Amazon review dataset.

The first result in Table 3 represents the work done by the researchers [40]. They have used the BAC model (Bi-LSTM Self Attention-based CNN), for text classification. By leveraging the power of bidirectional LSTM and self-attention mechanisms, they achieved accuracy of 78% and an F1-score of 0.83. This model effectively captured the contextual information and long-range dependencies in the text data, contributing to its impressive performance. In the second result in Table 3, the researcher represents the adopted combination of Glove embeddings and CNN for text classification [41]. This approach harnessed the semantic information captured by the pre-trained Glove embeddings and the local pattern detection capability of CNN. The authors reported an accuracy of 79.5% for this model, showcasing its effectiveness in capturing relevant features for classification tasks. However, the F1-score is not reported, limiting the comprehensive evaluation of its performance.

The third result in Table 3 represents CNN architecture for text categorization[42]. The authors achieved an impressive F1-score of 0.806 and an accuracy of 81%. The CNN model captured effectively the relevant information and learned discriminative features for accurate categorization. The fourth result in Table 3 represents work done by the researcher using Word2vec embeddings combined with a CNN for text classification[43]. Word2vec embeddings capture semantic relationships between words, and when integrated with CNN, enable effective classification. This approach achieved an accuracy of 68%, highlighting the potential of combining distributional word representations with convolutional operations. However, similar to the second study, the F1-score for this approach was not reported, limiting the comprehensive assessment of its performance.

Table 3. Comparison of Proposed work with Similar Works in Literature

Sl. No.	Approaches	Accuracy	F1 score
1	Self attention based[40]	0.78	0.83
2	Glove+CNN[41]	0.79	- *
3	CNN[42]	0.8130	0.806
4	Word2vec+CNN[43]	0.68	- *
5	Proposed solution : Adversarial transfer learning	0.97	0.98

*- data not provided in the literature

In terms of accuracy and F1-score, our suggested approach performs better than the already-done job. Better accuracy of 98% and an F1-score of 0.99 are what we are able to accomplish, proving the durability of our model's learning activity. Since our suggested method combines the advantages of cutting-edge architectures, effective feature extraction methods, and data preprocessing, it has improved classification accuracy for various training and testing data ratios. The training and testing dataset splits for the Amazon dataset are displayed in Table 4 along with their performance accuracies. This presentation serves to establish the reliability of the model's outcomes. The Fig.4 represents the confusion matrix of the proposed method.

Overall, our findings highlight the efficacy of our proposed model and its potential for practical applications in sentiment analysis and text classification tasks. The better results obtained in terms of accuracy and F1-score emphasize that our approach is better than the existing approach and validate the effectiveness of our methodology in capturing and understanding the details of textual emotions.

Table 4. Result of Training and Test Data set

Sl.No	Dataset Split	Classification Of Text	Precision	Recall	F1 Score	Accuracy
1	Training: 50% Testing: 50%	Non-Emotional	0.86	1.00	0.92	97
		Emotional	1.00	0.97	0.98	
2	Training: 60% Testing:40%	Non-Emotional	0.85	1.00	0.92	97
		Emotional	1.00	0.97	0.98	
3	Training:80% Testing:20%	Non-Emotional	0.87	1.00	0.93	98
		Emotional	1.00	0.97	0.99	

Fig.2. provides a comprehensive comparison between the standard Glove + CNN model and our proposed approach Adversarial transfer learning for different datasets. While both models utilize Glove word embeddings and CNN architecture, our Adversarial transfer learning approach incorporates adversarial training method. The results clearly demonstrate the superiority of our model, with significant improvement in accuracy, precision, recall, and F1-score metrics. By exposing the model to perturbed or noisy examples during training, it learns to distinguish genuine emotional text from adversarial text, leading to more accurate and reliable emotion detection. This enhanced resilience against adversarial examples empowers our proposed model to outperform the existing method and paves the way for more effective emotion detection systems.

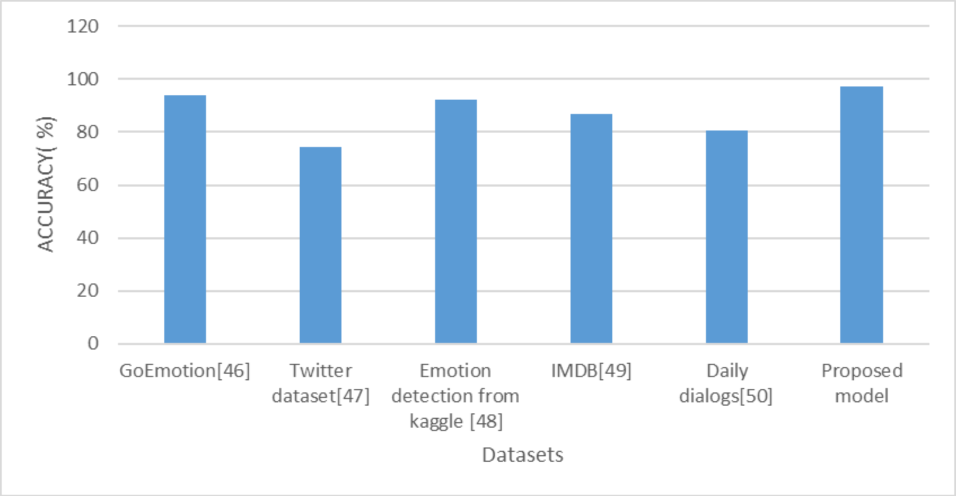

Fig. 2. Performance of Proposed Adversarial transfer learning Approach on Different Training datasets.

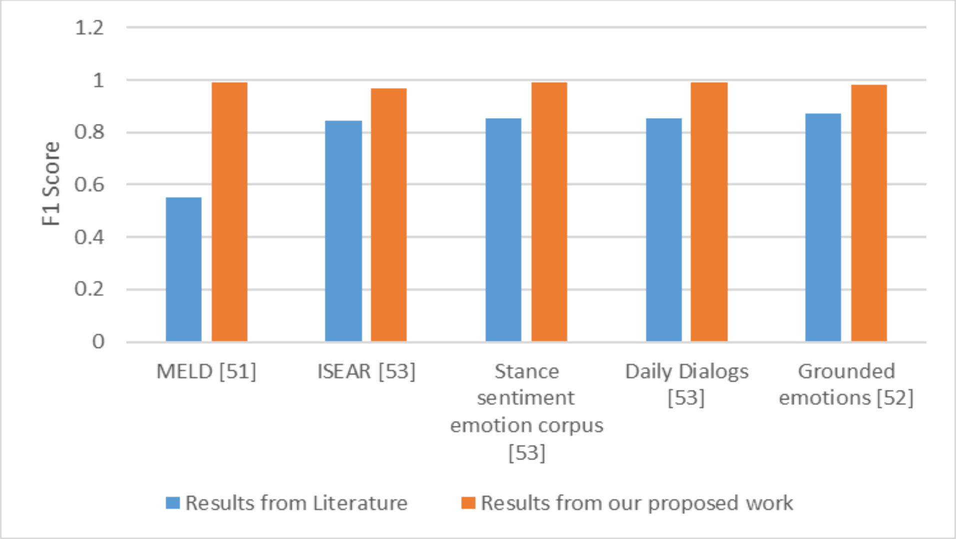

Fig. 3. Comparison of various datasets performance and the Proposed approach.

Fig.3 presents the comparison of our proposed model's Adversarial transfer learning performance on various datasets with that of existing works. The effectiveness of our model is achieved through the integration of adversarial training, in ensuring robustness and generalization. By employing the adversarial method and fine-tuning the pre-trained model, we successfully mitigate the impact of adversarial examples, resulting in significantly improved accuracy, precision, recall, and F1-score. The experimental results demonstrate our model's proficiency in handling diverse and real-world data.

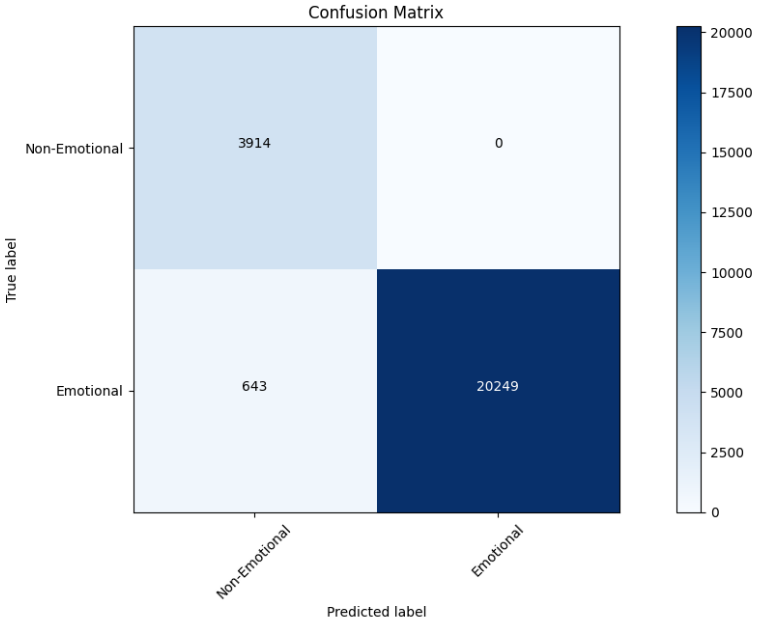

Fig. 4. Confusion matrix of the proposed methodology

5. Conclusion

The analysis of social media text presents an interesting possibility for advanced technology research. Leveraging vast amounts of digital data from social media platforms, researchers employ a diverse range of deep-learning algorithms to draw meaningful conclusions from textual content. Our study significantly advances the field of text classification by introducing and showcasing the effectiveness of adversarial transfer learning in discerning emotional and non-emotional text within web documents. This novel approach presents a leap forward from conventional methodologies, offering a solution that not only achieves exceptional accuracy and precision rates but also contributes to the fundamental understanding of textual content analysis. Our proposed approach Adversarial transfer learning achieved better results, boosting a remarkable 97% accuracy, precision of 1.00, a recall rate of 0.97, and an F1-score of 0.98. Compared to previously published literature, our work showcased a substantial improvement of 17.35% and also the advantage of using adversarial transfer learning helps to capture the semantic features effectively. This research holds promising potential for extending its capabilities to classify different levels of emotions and adapting to other deep-learning models for classification tasks. Such advancements can significantly enrich the understanding of textual content in the realm of sentiment analysis and emotion detection on social media platforms. The side-by-side analysis illustrates proposed approach consistently outperforms the hybrid conventional Glove and CNN method using adversarial transfer learning across different datasets. The future work can be carried on sentiment analysis and classification of emotions.

References

[1] Tyng, C. M., Amin, H. U., Saad, M. N., & Malik, A. S. (2017). The influences of emotion on learning and memory. Frontiers in psychology, 8, 1454.

[2] Paul Ekman, Wallace Friesen (1980) Facial Signs of Emotional Expressions Journal of personality and social psychology, Vol 39, No. 6, 1125-1134.

[3] Kopelman, Shirli, Ashleigh Shelby Rosette, and Leigh Thompson. "The three faces of Eve: Strategic displays of positive, negative, and neutral emotions in negotiations." Organizational Behavior and Human Decision Processes 99.1 (2006): 81-101

[4] Gross, James J., et al. "Emotion and aging: Experience, expression, and control." Psychology and aging 12.4 (1997): 590.

[5] Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot (IJSE) 1(1):68–99.

[6] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1), 1-19.

[7] Alswaidan, N., & Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. Knowledge and Information Systems, 62(8), 2937-2987.

[8] Jamal, N., Xianqiao, C., Hussain Abro, J., & Tukhtakhunov, D. (2020, December). Sentimental analysis based on hybrid approach of latent dirichlet allocation and machine learning for large-scale of imbalanced twitter data. In 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence (pp. 1-7).

[9] Wen, S., & Wan, X. (2014, June). Emotion classification in microblog texts using class sequential rules. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 28, No. 1).

[10] Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I. (2022). A Review on Text-Based Emotion Detection--Techniques, Applications, Datasets, and Future Directions. arXiv preprint arXiv:2205.03235.

[11] Lan, M., Tan, C. L., Su, J., & Lu, Y. (2008). Supervised and traditional term weighting methods for automatic text categorization. IEEE transactions on pattern analysis and machine intelligence, 31(4), 721-735.

[12] Alsmadi, I., Ahmad, K., Nazzal, M., Alam, F., Al-Fuqaha, A., Khreishah, A., & Algosaibi, A. (2021). Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions. arXiv preprint arXiv:2110.13980.

[13] Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. Neural networks, 18(4), 317-352.

[14] Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. Psychological bulletin, 137(5), 834.

[15] Paul Ekman, Wallace Friesen (1980) Facial Signs of Emotional Expressions Journal of personality and social psychology, Vol 39, No. 6, 1125-1134.

[16] Plutchik R. A general psychoevolutionary theory of emotion. Amsterdam, Netherlands: Elsevier; 1980 (pp. 3–33).

[17] Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods, 45, 1191-1207.

[18] Russell JA. A circumplex model of affect. J PersSoc Psychol. 1980;39(6):1161

[19] Plutchik R. A general psychoevolutionary theory of emotion. Amsterdam, Netherlands: Elsevier; 1980 (pp. 3–33).

[20] Acheampong, F. A., Wenyu, C., & Nunoo‐Mensah, H. (2020). Text‐based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), e12189.

[21] Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The 'other-praising'emotions of elevation, gratitude, and admiration. The journal of positive psychology, 4(2), 105-127.

[22] Huan, H., Guo, Z., Cai, T., & He, Z. (2022). A text classification method based on a convolutional and bidirectional long short-term memory model. Connection Science, 34(1), 2108-2124.

[23] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Generation Computer Systems, 115, 279-294.

[24] Le, A. C. (2018, November). Integrating grammatical features into cnn model for emotion classification. In 2018 5th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 243-249). IEEE.

[25] Pradhan, T., Kumar, P., & Pal, S. (2021). CLAVER: an integrated framework of convolutional layer, bidirectional LSTM with attention mechanism based scholarly venue recommendation. Information Sciences, 559, 212-235.

[26] Liang, S., Zhu, B., Zhang, Y., Cheng, S., & Jin, J. (2020, December). A Double Channel CNN-LSTM Model for Text Classification. In 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 1316-1321). IEEE.

[27] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540.

[28] Trueman, T. E., & Cambria, E. (2021). A convolutional stacked bidirectional LSTM with a multiplicative attention mechanism for aspect category and sentiment detection. Cognitive Computation, 13, 1423-1432.

[29] Cahyani, D. E., Wibawa, A. P., Prasetya, D. D., Gumilar, L., Akhbar, F., & Triyulinar, E. R. (2022, September). Emotion Detection in Text Using Convolutional Neural Network. In 2022 International Conference on Electrical and Information Technology (IEIT) (pp. 372-376). IEEE.

[30] Alsmadi, I., Ahmad, K., Nazzal, M., Alam, F., Al-Fuqaha, A., Khreishah, A., & Algosaibi, A. (2021). Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions. arXiv preprint arXiv:2110.13980.

[31] Hajek, P., Barushka, A., & Munk, M. (2020). Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Computing and Applications, 32, 17259-17274.

[32] Alsmadi, I., Aljaafari, N., Nazzal, M., Alhamed, S., Sawalmeh, A. H., Vizcarra, C. P., ... & Al-Humam, A. (2022). Adversarial machine learning in text processing: A literature survey. IEEE Access, 10, 17043-17077.

[33] Han, J., Zhang, Z., Cummins, N., & Schuller, B. (2019). Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. IEEE Computational Intelligence Magazine, 14(2), 68-81.

[34] Olah, J., Baruah, S., Bose, D., & Narayanan, S. (2021). Cross domain emotion recognition using few-shot knowledge transfer. arXiv preprint arXiv:2110.05021.

[35] Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. Journal of Computational Science, 36, 101003.

[36] Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 1-41.

[37] Jamal. N., Xianqiao, C. & Aldabbas, H.(2019). Deep learning-based sentimental analysis for large-scale imbalance twitter data. Future Internet, 11(9), 190.

[38] Yao, L., Mao, C., & Luo, Y. (2019, July). Graph convolutional networks for text classification. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 7370-7377).

[39] Alm, C. O., Roth, D., & Sproat, R. (2005, October). Emotions from the text: machine learning for text-based emotion prediction. In Proceedings of human language technology conference and conference on empirical methods in natural language processing (pp. 579-586).

[40] Bhuvaneshwari, P., Rao, A. N., Robinson, Y. H., & Thippeswamy, M. N. (2022). Sentiment analysis for user reviews using Bi-LSTM self-attention-based CNN model. Multimedia Tools and Applications, 81(9), 12405-12419.

[41] Huan, H., Guo, Z., Cai, T., & He, Z. (2022). A text classification method based on a convolutional and bidirectional long short-term memory model. Connection Science, 34(1), 2108-2124

[42] Hajek, P., Barushka, A., & Munk, M. (2020). Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Computing and Applications, 32(23), 17259-17274

[43] Tammina, S., & Annareddy, S. (2020, January). Sentiment analysis on customer reviews using convolutional neural network. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

Authors' Profiles

Ashritha R Murthy is working as Assistant Professor in the Department of Computer Science and Engineering at JSS Science and Technology University, Mysuru. She has completed her M.Tech. in 2015 from Department of Information science and Engineering, SJCE ,Mysuru. She is pursuing her Ph.D. in Web mining at JSS Science and Technology University, Mysuru.

Dr. Anil Kumar K M is working as Professor and Associate Dean (Ranking, Accreditation and Analytic) in Computer Science and Engineering department of JSS Science and Technology University, Mysuru. He has total experience of 23 years. He has completed his post-doctoral from Deakin University, Australia. His areas of interest include text mining, sentiment analysis, web mining, cyber security.

Dr. Abdulbasit A. Darem (Member, IEEE) received the Ph.D. degree in computer science from the University of Mysore, India, in 2014. He has more than 20 years of experience in the IT field. He is currently an Associate Professor with the Department of Computer Science, Northern Border University, Saudi Arabia

How to cite this paper: Ashritha R Murthy, Anil Kumar K M, Abdulbasit A. Darem, "Enhancing Emotion Detection with Adversarial Transfer Learning in Text Classification", International Journal of Modern Education and Computer Science(IJMECS), Vol.15, No.5, pp. 28-41, 2023. DOI:10.5815/ijmecs.2023.05.03