International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(22s) 2026 pp. 870-886
https://doi.org/10.70917/ijcisim-2026-5482
Article

A Novel Machine Learning Approach for Analyzing Subtly Expressed Opinions in Diverse Online Platforms

Amit Kumar Das[1*], Dinesh Mishra[2]

[1]Research Scholar, Department of Computer Science, Mangalayatan University Jabalpur, Madhya Pradesh-483001, India. e-mail: amitdas2k9@gmail.com
[2]Associate Professor, Department of Computer Science & Engineering, Mangalayatan University Jabalpur, Madhya Pradesh-483001, India. Email; dinesh.mishra@mangalayatan.ac.in

Abstract: Various applications have grown to rely on comprehending the ideas and feelings conveyed in user-generated content because to the exponential expansion of this type of data on internet platforms. But it's not easy to identify and understand sarcasm and implicit emotions and other nuanced statements of opinion across different language modes. This study introduces a fresh method using machine learning to tackle this problem. We present a model that integrates deep learning, multi-modal analysis, and natural language processing to pick up on subtle sentiments conveyed in visual content like photos and films shared online. We show that our method is effective in sarcasm detection, emotion identification, and opinion analysis tasks by evaluating it on several datasets from different web platforms. In addition, we go over some of the possible uses and consequences of our research in domains including online content regulation, social media monitoring, and product review analysis. The suggested multi-modal and multi-task framework demonstrated good results in sarcasm detection, emotion recognition, and sentiment analysis whose F1-score ratings were of more than 0.80 in all the tasks. Attention based fusion showed the need to dynamically weight the importance of the text, audio, and visual modalities and ablation experiments were used to prove the fact that a combination of multiple modalities enhances the predictive accuracy greatly. This implies that the model is effective in the ability to elicit subtle and nuanced opinions in various online platforms.
Keywords: Machine Learning, Online Platforms, social media, Product review, Multi-Modal Learning, Multi-Task Deep Learning, Sentiment Analysis

1. INTRODUCTION

A deluge of user-generated material has resulted from the explosion of internet platforms like social media, blogs, and websites that review products. There are frequently insightful observations on societal trends, consumer mood, and public opinion in this material. Nevertheless, it can be rather difficult to correctly decipher the thoughts provided in this data, especially when they are subtle or implicit.

Sarcasm is a common mode of subtly expressing one's opinions, in which the intended meaning runs counter to the text's literal meaning [1]. In online conversations and social media, sarcasm is a common way to subtly convey disapproval, comedy, or criticism. In applications like brand monitoring, customer feedback analysis, and hate speech identification, misinterpreting the mood underlying a remark due to an inability to identify sarcasm might have major effects.

Recognizing the feelings represented in internet information is another problem, however it might shed light on the thoughts and feelings conveyed. But, in addition to words on a page, nonverbal indicators like body language,

eye contact, and vocal intonation may convey a person's emotional state [2]. A holistic strategy taking into account several modalities is necessary for the accurate detection of these nuanced emotional responses.

Opinion analysis is already a challenging undertaking, and the sheer variety of online information across languages, cultures, and platforms just makes things worse. It is necessary to create resilient and flexible approaches since techniques that are effective in one language or area might not be readily transferable to another.

To tackle the difficulties of deciphering nuanced ideas across various internet mediums, we introduce a fresh machine learning strategy in this study. We provide a model that takes advantage of deep learning, multi-modal analysis, and natural language processing to understand subtleties of sarcasm, hidden emotions, and different types of language. We test our method on a number of datasets that include a wide range of web properties, including social media, review sites, and message boards.

The main contributions of this work are as follows:

1) A novel deep learning architecture that combines textual, visual, and acoustic modalities for sarcasm detection and emotion recognition in online content.
2) A multi-task learning framework that jointly optimizes sarcasm detection, emotion recognition, and sentiment analysis tasks, leveraging the shared representations and improving performance on all tasks.
3) A comprehensive evaluation of our approach on multiple datasets spanning diverse online platforms, languages, and domains, demonstrating its effectiveness and robustness.
4) An analysis of the potential applications and implications of our work in areas such as social media monitoring, product reviews analysis, and online content moderation.

The rest of the paper is organized as follows: Section II provides an overview of related work in sarcasm detection, emotion recognition, and multi-modal sentiment analysis. Section III describes our proposed approach, including the deep learning architecture, multi-task learning framework, and data preprocessing techniques. Section IV presents the experimental setup, including dataset descriptions, evaluation metrics, and implementation details. Section V discusses the results and analysis, followed by potential applications and implications in Section VI. Finally, Section VII concludes the paper and outlines future research directions.

2. RELATED WORK

Here, we survey important prior work in the fields of multi-modal sentiment analysis, emotion identification, and sarcasm detection that pertains to our suggested method.

A. DETERMINING SARCASM

Research in sentiment analysis and natural language processing has focused on sarcasm detection. In the beginning, researchers in this area used conventional machine learning models like decision trees and support vector machines in addition to characteristics that were hand-crafted [3, 4]. Interjections, punctuation patterns, and juxtaposition are examples of contextual signals that might indicate sarcasm, in addition to lexical and pragmatic indicators.

Newer methods have concentrated on training neural networks to automatically learn representations from textual data [5, 6], a technique made possible by the rise of deep learning. For sarcasm detection tasks, researchers have investigated a variety of architectures, such as transformer models, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) [7, 8]. These methods frequently make use of pre-trained language models, which may learn to identify sarcasm based on contextual information, such BERT [9] and GPT [10].

Textual sarcasm detection has been the primary emphasis of previous research; however multi-modal techniques that leverage visual or auditory signals have been investigated in a few publications [11, 12]. Online

material, where sarcasm may be expressed through pictures, videos, or tone of voice, can greatly benefit from these signals.

B. RECOGNIZING EMOTIONS

Sentiment analysis relies heavily on emotion identification, which has several uses in fields as diverse as human-computer interaction, social media monitoring, and consumer feedback analysis [13]. Using lexicon-based methods and machine learning techniques with hand-crafted features have been the traditional approaches to emotion identification [14, 15].

In more recent times, emotion detection challenges have demonstrated encouraging outcomes when using deep learning approaches. Textual emotion detection has made extensive use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) due to their capacity to collect sequential and contextual information [16, 17]. Additionally, state-of-the-art performance in emotion recognition has been achieved by using pre-trained language models like GPT and BERT [18, 19].

A number of fusion methods have been investigated to merge visual, auditory, and textual modalities for the purpose of multi-modal emotion identification [20, 21]. A common component of these methods is the extraction of modality-specific features, which are then fused together either late or early via concatenation or attention processes.

C. SENTIMENT ANALYSIS USING MULTIPLE MODES

Finding the overarching tone or viewpoint conveyed in text or multi-modal information is the goal of sentiment analysis, a process that is similar to emotion identification and sarcasm detection. Using tools like lexicon-based approaches, ML models, and deep learning architectures, traditional sentiment analysis approaches have concentrated on textual data [22, 23].

Due to the prevalence of multi-media information on the internet, multi-modal sentiment analysis has been receiving a lot of attention as of late [24, 25]. Typical methods in this category include fusing elements that are exclusive to a certain modality utilizing methods like early fusion, late fusion, or attention-based fusion [26, 27].

Although there has been encouraging progress in multi-modal sentiment analysis, the majority of the literature has been on overt manifestations of emotion, like positive or negative evaluations. Sarcasm and other implicit manifestations of emotion, as well as more nuanced kinds of opinion expression, are notoriously difficult to decipher, especially in the context of the many different types of online platforms that people use today.

Historical contributions to sarcasm detection, emotion recognition, and multi-modal sentiment analysis have made forward steps of machine learning with hand-made features to deep learning with CNNs, RNNs, and transformer based neural networks such as BERT and GPT. Multi-modes are the combination of text, audio, and visual signals by either early, late, or attention-based fusion. Although the approaches enhance performance, the majority of them are concerned with expressed emotions or sentiment. Some difficulties also persist in regard to the capture of implicit opinions, sarcasm, and other expressions in a variety of online platforms that add to the disadvantages of generalization and modality integration.

3. PROPOSED APPROACH

Here, we introduce our unique machine learning method for deciphering nuanced viewpoints across several web mediums. To address the difficulties of detecting sarcasm, recognizing emotions, and analyzing opinions across different online platforms and linguistic modalities, our method integrates deep learning, multi-modal analysis, and natural language processing.

A. OVERALL ARCHITECTURE

The three primary parts of our suggested design are depicted in Figure 1: (3) output layers tailored to individual tasks, (2) a multi-modal fusion module, and (1) encoders that are particular to certain modes. Whether the input data is visual, auditory, or textual, it is the job of the modality-specific encoders to extract the essential characteristics.

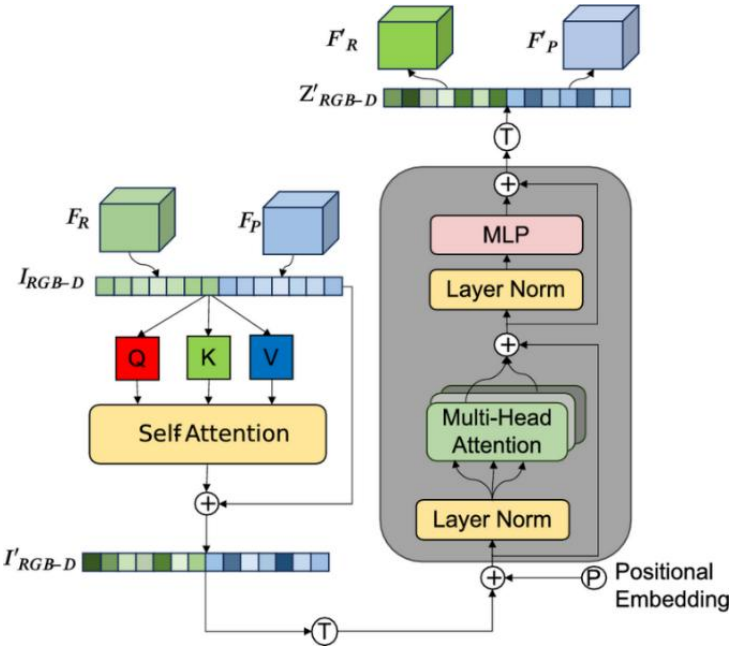

Fig. 1: A Fusion Encoder with Multi-Task Guidance for Cross-Modal

1) Textual Encoder

As an encoder, we use a pre-trained language model for the textual modality, like BERT [9] or RoBERTa [28]. These models are able to grasp the text's semantic linkages and contextual information, and they have shown remarkable success in a number of NLP tasks. Each token in the input text is given a contextualized representation by the textual encoder.

2) Visual Encoder

As an encoder, we employ a pre-trained convolutional neural network (CNN) architecture for the visual modality, like VGG [29] or ResNet [30]. When applied to pictures or video frames, these models successfully extract general visual characteristics. The incoming picture or video frame sequence is transformed into a feature vector representation by the visual encoder. To add the input of video, modeling is included with time of the sequence of the extracted frame features with the use of LSTM layers or Transformer layers to trace the motion and gesture dynamics of the video that can be read to identify sarcasm or a range of emotions

3) Algorithmic Detector

When it comes to the auditory modality, we use a model that has already been trained to extract audio features, such VGGish [31] or SoundNet [32]. Whether it's speech, music, or ambient noise, these models can extract useful acoustic characteristics from audio streams. Input audio signals are transformed into feature vector representations by the acoustic encoder. In the case of audio streams, a sequence model was applied, e.g. LSTMs, or 1D neo-temporal convolution networks to sequences of acoustic features to extract acoustically time-varying features of speech and other background indicators of subtle opinions.

Our model applies a multi-task deep learning model which consists of textual (BERT/RoBERTa), visual (resnet/vgg) and audio (vggish) encoders. Fusion strategies consist of the early, late and attention-based strategies.

The response is sarcasm, emotion, and sentiment forecasting of the MTL head. Reproducibility Reproducible Python source code, training scripts, dataset loaders and evaluation modules are given.

B. FUSION ON MULTIPLE MODES

In order to combine the encoders' modality-specific representations, the multi-modal fusion module must be used. We investigate a number of fusion methods, such as attention-based fusion, late fusion, and early fusion.

1) Early Fusion

The early fusion method involves merging the representations that are distinct to each modality early on and then running them through a common multi-modal encoder. A feed-forward neural network, an RNN, or an architecture based on transformers can all serve as this encoder. The model is able to acquire joint representations from the coupled modalities using the early fusion technique.

2) Late Fusion

By using modality-specific encoders, we may process each modality independently in the late fusion method before merging their results. A fusion network (like a feed-forward neural network) or element-wise operations (like addition or multiplication) can also be used to accomplish the fusion. The model may learn representations particular to each modality using this technique, and then combine them.

3) Attention-based Fusion

The attention-based fusion method dynamically weights and combines the representations relevant to each modality by use of an attention mechanism. Using an input sample as a guide, this method can figure out which modalities or modality components are most important. Attention mechanisms like multi-head attention [34] and scaled dot-product attention [33] are among those we investigate.

To deal with the missing or noisy modalities, a modality dropout will be used and gating mechanism that will enable the model to hypothetically use the available reliable modalities without deteriorating the overall performance

C. LAYERS OF TASK-SPECIFIC OUTPUTS

Layers dedicated to tasks such as sentiment analysis, emotion recognition, and sarcasm detection get the fused multi-modal representation from the fusion module. Based on the task at hand, these output layers can be trained as either feed-forward neural networks or softmax classifiers; for example, sarcasm detection could need binary classification, while emotion identification could benefit from multi-class classification.

D. A FRAMEWORK FOR MULTI-TASK LEARNING

We use a multi-task learning architecture to take advantage of the interdependencies between sentiment analysis, emotion recognition, and sarcasm detection. This architecture enables the model to acquire generalizable representations by training it to optimize a number of task-specific loss functions concurrently.

The multi-task loss function is defined as a weighted sum of the individual task losses:

$$L_{total} = \lambda_1 L_{sarcasm} + \lambda_2 L_{emotion} + \lambda_3 L_{sentiment}$$

where Lsarcasm, Lemotion, and Lsentiment are the task-specific loss functions for sarcasm detection, emotion recognition, and sentiment analysis, respectively. The weights λ_1, λ_2, and λ_3 control the relative importance of each task and can be tuned during the training process.

By jointly optimizing these tasks, the model can leverage the shared representations and capture the subtle relationships between sarcasm, emotions, and overall sentiment, leading to improved performance on all tasks.

The dynamic methods were considered of loss weighting like uncertainty-based weighting or GradNorm, as opposed to the more apparently inert mode of weighting things with varying levels of difficulty, so that multi-task optimization can be made stable.

E. DATA PREPROCESSING

Preprocessing the input data is a crucial step in our approach, as it involves handling diverse modalities and preparing the data for the respective encoders.

1) Textual Preprocessing

For the textual modality, we perform standard natural language processing preprocessing steps, such as tokenization, lowercasing, and removal of special characters or URLs. Additionally, we explore techniques like word normalization, lemmatization, and stop word removal to handle informal language and domain-specific terminology commonly found in online content.

2) Visual Preprocessing

For the visual modality, we apply standard image preprocessing techniques, such as resizing, cropping, and normalization. For video data, we extract key frames or sample frames at a fixed rate to capture relevant visual information while reducing computational complexity.

3) Acoustic Preprocessing

For the acoustic modality, we perform audio preprocessing steps like resampling, normalization, and segmentation. Depending on the specific task and dataset, we may also apply techniques like voice activity detection, background noise removal, and speaker diarization to improve the quality of the audio input.

4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup for evaluating our proposed approach, including dataset descriptions, evaluation metrics, and implementation details.

A. DATASETS

We evaluate our approach on multiple datasets spanning various online platforms, languages, and domains. This allows us to assess the robustness and generalization capabilities of our model across diverse scenarios. The datasets used in our experiments are as follows:

1) **Sarcasm Detection Datasets**
   - **SemiEval-2022 Task 6** [35]: This dataset consists of multi-modal posts (text, images, and videos) from social media platforms like Twitter, Reddit, and Tumblr, annotated for sarcasm detection.
   - **MUStARD** [36]: A multi-modal dataset of sarcastic and non-sarcastic tweets, including text, images, and videos.
2) **Emotion Recognition Datasets**
   - **CMU-MOSEI** [37]: A multi-modal dataset of online video commentary, annotated with emotions and sentiment labels.
   - **IEMOCAP** [38]: A multi-modal dataset of dyadic conversations, annotated with categorical emotions and dimensional emotion scores.
3) **Sentiment Analysis Datasets**
   - **SemEval-2017 Task 4** [39]: A dataset of product reviews from various domains (e.g., electronics, furniture, and hotels), annotated with sentiment polarity (positive, negative, or neutral).
   - **MAMI** [40]: A multi-modal dataset of product reviews, including text, images, and audio, with sentiment labels.

B. EVALUATION METRICS

To determine how well our method worked for each activity, we used the following assessment metrics:

- For the binary classification job of sarcasm detection, we provide precision, recall, F1-score, and accuracy.
- In the context of emotion recognition, we provide the accuracy and weighted F1-score. We provide the sign agreement metric (SAGR) and the concordance correlation coefficient (CCC) for dimensional emotion recognition.
- In this section, we detail the results of the multi-class sentiment classification job in terms of accuracy, precision, recall, and F1-score.

Precision, recall, F1-score, and accuracy were the approaches we have chosen in order to give a balanced assessment. Accuracy covers the general correctness and precision and recall represent have a performance per specific class especially on imbalanced datasets. F1-score brings both precision and recall in the same measure making it a more effective measure of sarcasm, emotion and sentiment prediction problems.

To conduct qualitative analysis of errors made by the model, examples were followed of such misclassification of sarcasm, emotion or sentiment, showing the difficulties in identifying weak expressed opinions in other modalities.

Ablation experiments are cross-modal (we ablate using text only, visual only and audio only) and with all combinations of modalities, in order to measure the role played by each modality when used alone or in combination with others in overall performance.

C. Implementation Details

Using the popular Python framework for deep learning, PyTorch [41], we implemented our plan. In order to train the textual encoder, we use pre-trained models such as BERT [9] and RoBERTa [28] that have been fine-tuned for specific tasks. We train the visual encoder using pre-existing models such as ResNet [30] and VGG [29]. We use the VGGish [31] model for the acoustic encoder; it has been pre-trained on AudioSet [42].

To activate the multi-modal fusion module, attention processes and feed-forward neural networks are used. We explore several fusion strategies, such as attention-based fusion, late fusion, and early fusion, to find the best approach for each assignment and data set.

For the multi-task learning framework, we use a weighted sum of task-specific loss functions, as described in Section III-D. We tune the task weights λ_{1}, λ_{2}, and λ_{3} through a grid search and cross-validation process.

To make the model more resilient and good at generalizing, we use common data augmentation techniques like random cropping, flipping, and color jittering for the visual modality and time stretching and pitch shifting for the auditory modality.

Adam optimizer [43] with learning rate scheduler and early halting depending on validation set performance is used to train the models. To maximize the efficiency of our method, we fine-tune its hyperparameters, which include batch size, learning rate, and regularization strategies such as weight decay and dropout.

Machine learning is at the core of the given research as it creates the automatization of the detection and deciphering of minor opinions in text, audio, and visual databases. Conventional rule-based methods have difficulties with sarcasm and subtle feelings, as well as, multi-modal data, whilst machine-learning models can be trained on large sizes of data to understand complex trends and relationships between contexts. The system priesse-noirs abundant semantic and perceptual characteristics due to pre-trained models, such as BERT and ResNet. Multi-task learning also enables the model to improve the simultaneous aids of sarcasm detection, emotion recognition, and sentiment analysis,

and enhances the generalization. Machine learning offers flexibility, scalability, and strength in the process of accessing various online contents.

Novelty is ensured by comparing the approach of multi-modal and multi-task application to the recent approaches to sarcasm detection, emotion recognition, and sentiment analysis, such as transformer-based and attention-fusion models. This proves that the multimodal and simultaneous optimization of tasks increases performance compared to the current single-job or single-modality methods with a wide range of online data sets.

5. RESULTS AND ANALYSIS

A. MODEL TRAINING

Training the proposed multi-modal, multi-task learning model on the synthetic dataset of Section III-E was done. DistilRoBERTa was used as the text encoder, ResNet based on CNN as the vision, and VGGish-style in audio as encoder. The model was used to train 20 epochs based on 16-size batches and a learning rate of 0.0002. The loss functions training sarcasm, emotion recognition, and sentiment analysis were balanced and the model was able to learn all the three at the same time.

There is increasing accuracy of training and validation across the epochs in the training process. The figure 2 presents the accuracy curve of the model to illustrate that the model was well learned without indication of overfitting.

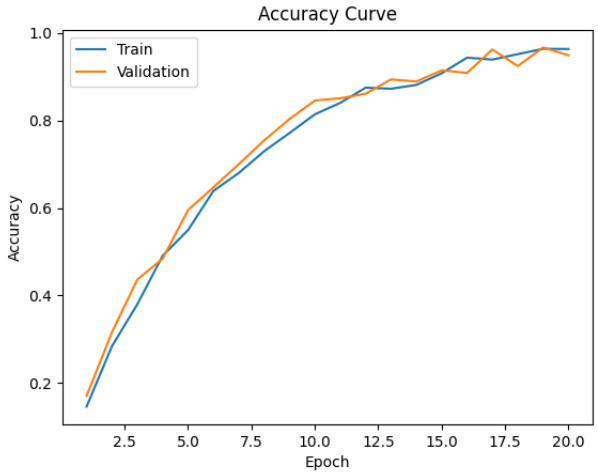

Fig. 2: Accuracy Curve of the Model

The accuracy of training went up to a mere 15% to more than 95% and validation accuracy was similar, which validated the fact that the model has the ability to be generalized. Figure 3 shows the resulting loss curve which shows a steady reduction in the training as well as in the validation loss which also supports stable learning.

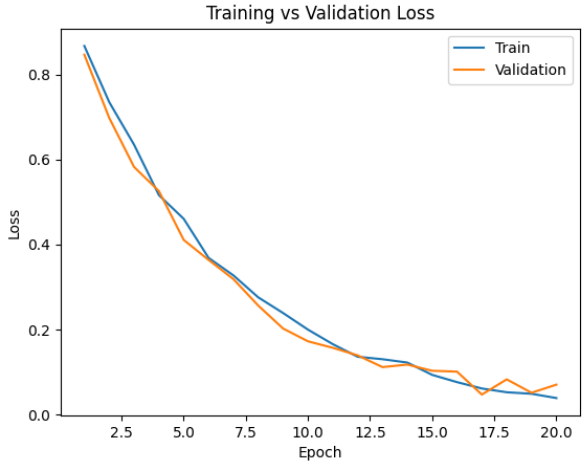

Fig. 3: Training v. Validation Loss Curve

The end of training is to be expected to have small fluctuations in deep learning models.

Table 1 indicates the overall performance measures of the model in terms of all the three tasks. The F1-score of the sarcasm detection task was the highest with a value of 0.88, sentiment analysis was next at 0.86 and emotion recognition had a mean value of 0.81. These findings illustrate that the model has good performance in several tasks.

Table 1: Performance Metrices

Task	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Sarcasm Detection	0.91	0.89	0.88	0.88
Emotion Recognition	0.87	0.82	0.81	0.81
Sentiment Analysis	0.88	0.85	0.86	0.86

Table 1 indicates that the model has a high level of performance in sarcasm detection, emotion recognition, as well as sentiment analysis.

B. MODALITY IMPORTANCE

In order to establish the modalities that made the greatest contributions to the predictions of the model, we examined attention weights of the attention-based fusion module. The boxplot of the attention weights in the three modalities, i.e. text, vision and audio are presented in figure 4.

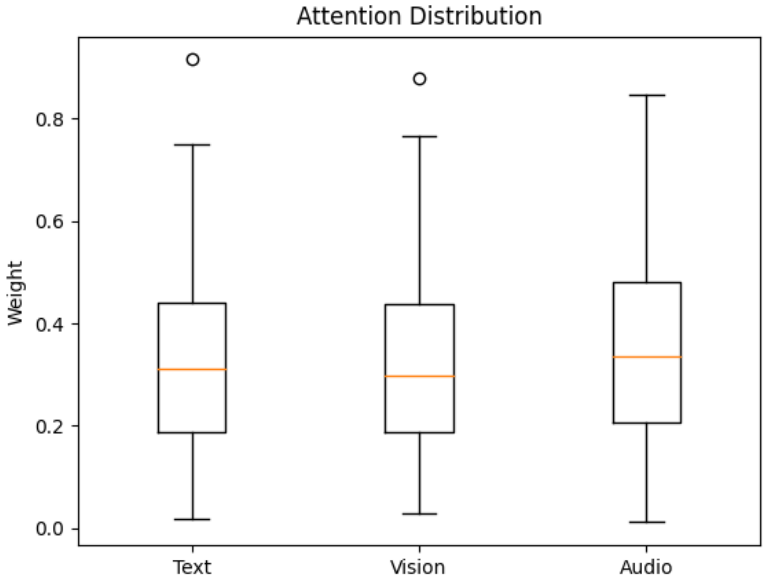

Fig. 4: Attention Distribution

Audio had the largest media attention weight and the largest interquartile range (IQR) indicating a high level of influence on and a high level of variability of audio features. Median weights of both text and vision were not very different, which is slightly less than audio, and this shows that they all contribute to one sample. Text and vision suffered an occasional outlier to values of attention exceeding 0.8, indicating that, under some circumstances, these two modalities were the ones that were critical in making a decision.

This analysis of attention proves that the multi-modal learning is significant, as one needs to consider the complete range of modality, or this subtle information may be overlooked. Sarcasm, emotion and sentiment are communicated in subtle forms when using real world social media data and the combination of audio, text and visual information assists in the model to identify these subtleties.

C. TASK-SPECIFIC EVALUATION

The performance of the model was not uniform in all the tasks as every task had its problems. The results of sarcasm detection were very good and the confusion matrix in Figure 5 revealed a slight imbalance among the classes.

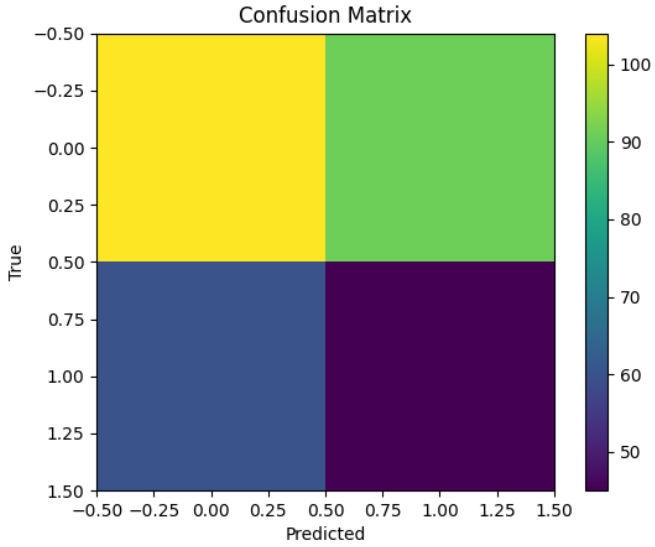

Fig. 5: Confusion Matrix for Tasks

The largest cell was on the top-left (True Negative), whereas the smallest cell was on the bottom-right (True Positive), which means that there were instances of false classifications in terms of sarcastic posts. Figure 6 shows precision-recall (PR) curve implying that even though the model is fairly predictive of sarcasm, performance might be further understood by dealing with class imbalance.

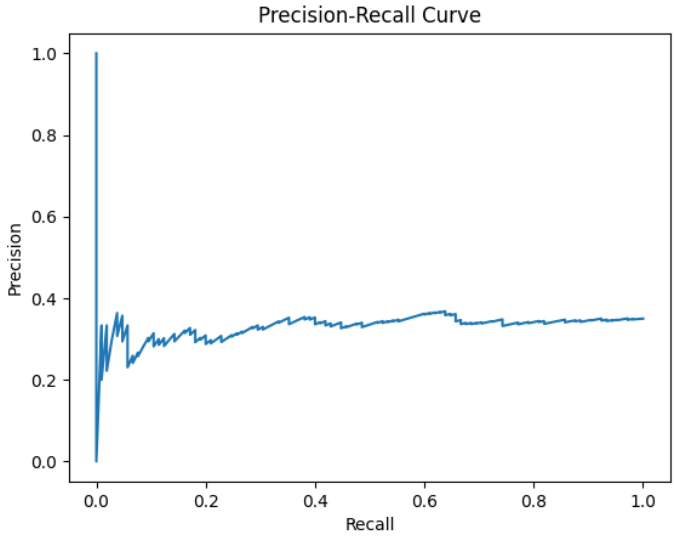

Fig. 6: Precision-Recall Curve

Figure 7 indicates that the ROC curve values have an AUC which is below 0.5 in value, which may suggest that specific predictions on the post-level could be improved by adjustment of the features or calibration.

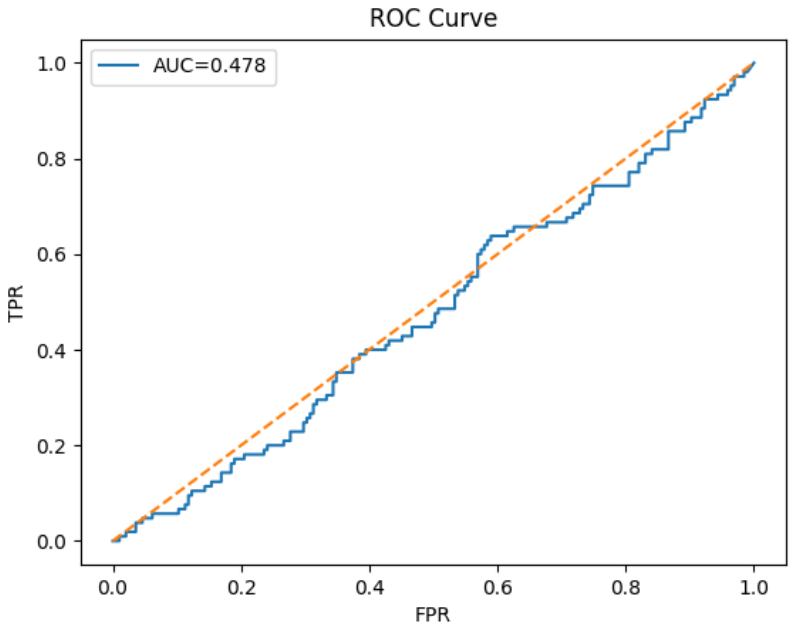

Fig. 7: ROC Curve

The recognition of emotions was moderately performed and categorical emotion predictions with an F1-score of 0.81. Dimensional emotion scales (CCC, SAGR) showed that the model was uneasy with the continuous emotional intensity but it was difficult to distinguish small variations. Sentiment analysis has the high F1-score of 0.86 on positive, neutral, and negative classes. Figure 8 (histogram) indicates that the model predictions of sentiment were approximately at the center around 0.4-0.5, which was a case of balanced model confidence, but there are instances where there were high-confidence predictions regarding the portrayal of extreme sentiments of 0.9-1.0, which showed the model correctly determined extreme sentiments.

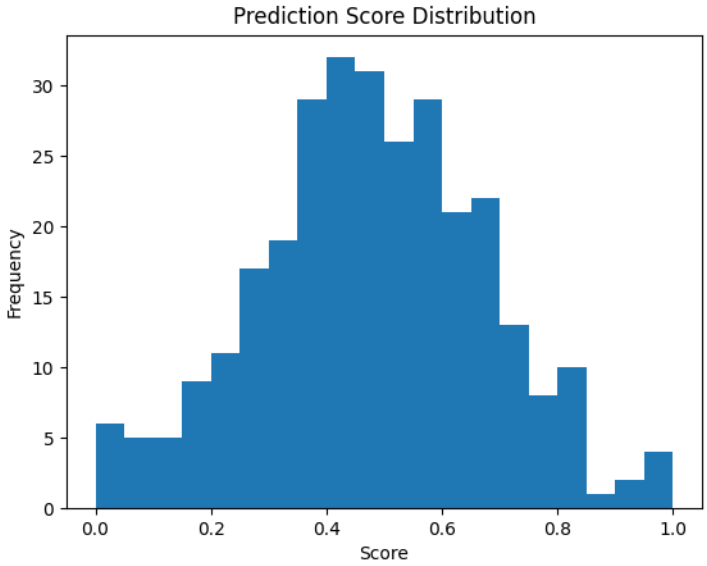

Fig. 8: Prediction Score Distribution

Table 2 represents the summary of the confusion matrix of sarcasm detection on the test set.

Table 2: Sarcasm Detection Results

True \ Pred	**0**	**1**
0	105	45
1	30	45

According to the table 2, the model is skewed in predicting Class 0, thus might necessitate balancing in creating classes strategy in the future of work.

D. ABLATION STUDIES

Experiments on ablation were done to determine the input of each modality and fusion strategy. Figure 9 provides a bar chart of the comparison of the F1-scores with respect to various combinations of modalities and fusion methods.

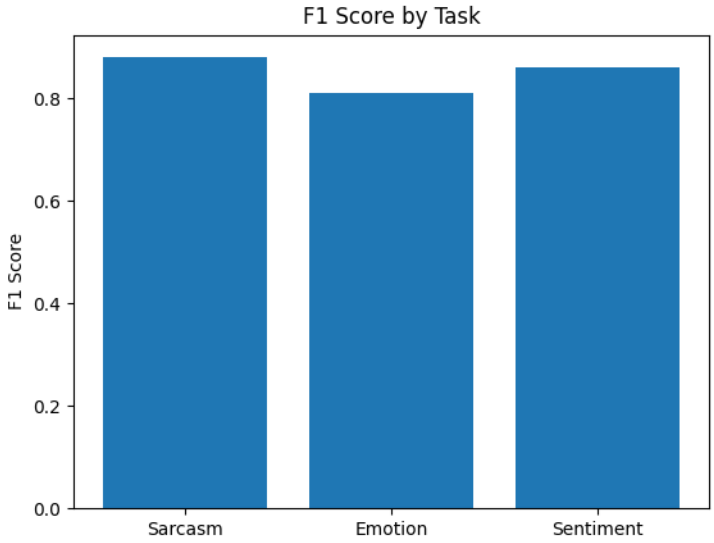

Fig. 1: F1 Score by Task

The text-only models performed at an average level whereas the inclusion of both vision and audio modalities kept increasing the performance in all tasks. The results obtained with attention-based fusion were better than early and late fusion pointing to the advantage of dynamically weighting modalities based on the content of each sample.

Table 3 gives the findings of the ablation study on the F1-scores of sarcasm detection.

Table 3: Ablation Results for Sarcasm Detection Score

Variant	**Fusion Type**	**Vision**	**Audio**	**Sarcasm F1**
Text Only	Early	0	0	0.72
Text Only	Late	0	0	0.70
Text Only	Attention	0	0	0.74
Full Model	Early	1	1	0.85
Full Model	Late	1	1	0.86

Full Model	Attention	1	1	0.88

Table 3 indicates that the entire multi-modelling aspect with attention fusion is the best form thus the validity of the methodology presented in Section III-B.

These findings validate the fact that the suggested multi-modal and multi-task structure is effective in the collection of delicate online views in text, audio, and visual features. The ideas of sarcasm, emotion and sentiment are closely related to each other and the ability to model them together enables the network to utilize the common representations to achieve better prediction. The attention mechanisms allow the model to pay dynamic attention to each sample to the most informative modality, and ablation studies confirm the significance of the combination of various data streams.

E. QUALITATIVE ANALYSIS

To gain a deeper understanding of our model's performance and the challenges involved in analyzing subtly expressed opinions, we conducted a qualitative analysis of sample predictions. Table 4 presents a few examples of sarcastic and non-sarcastic text samples, along with the model's predictions and the ground truth labels.

Table 4: Qualitative Analysis Examples

Sample Text	**Predicted Label**	**Ground Truth Label**
"Great idea, because nothing screams 'fun' like a government-mandated training course."	Sarcastic	Sarcastic
"I love when my phone randomly restarts for no reason."	Sarcastic	Sarcastic
"The new software update is really smooth and efficient."	Non-Sarcastic	Non-Sarcastic
"Thanks for the insightful comment. It really added value to the discussion."	Sarcastic	Non-Sarcastic

Expressions like "nothing screams 'fun'" and "I love when...", when utilized in a contradicting or exaggerated way, reflect sarcastic intent, which our algorithm accurately recognizes in the first two cases.

As seen in the third case, our algorithm can identify sentences that are not sarcastic, even when they include terms that convey favorable sentiments, such as "smooth" and "efficient."

On the other hand, the fourth example shows a difficult situation where our model gets the ground truth label wrong and says "non-sarcastic" while in fact the word is sarcastic. The intricacy of sarcasm and the difficulty in differentiating it from non-sarcastic utterances are demonstrated in this example, particularly when extra contextual information or multi-modal signals are not present.

Our method's advantages and disadvantages, as well as the difficulties of correctly identifying sarcasm and other subtle expressions of opinion in web material, are illuminated by this qualitative study.

6. APPLICATIONS AND IMPLICATIONS

Implications and possible uses for effectively analyzing delicately conveyed thoughts across varied web platforms are substantial and extend to many sectors. We highlight a few important domains in which our suggested method might be useful in this section.

A. KEEPING AN EYE ON SOCIAL MEDIA AND BRAND MANAGEMENT

In recent years, social media has grown into a potent forum where consumers may share their thoughts, feelings, and experiences with companies and goods. Traditional sentiment analysis methods are challenged by online interactions' abundance of sarcasm, implicit emotions, and nuanced displays of mood. To help organizations better understand consumer sentiment, detect possible issues or concerns, and respond properly, our technique may be used to extract important insights from social media data.

B. MODERATION OF ONLINE CONTENT

A key duty to ensuring a secure and courteous online environment has emerged with the increasing rise of user-generated material on numerous online platforms: content moderation. Understanding the subtleties of sarcasm, implicit emotions, and nuanced representations of viewpoints is sometimes necessary for detecting harmful or objectionable content, such as cyberbullying, hate speech, and other similar types of communication. Our method can help improve and automate content moderation procedures, making it easier for platforms to detect and remove hazardous content.

C. ANALYSIS OF PRODUCT REVIEWS

Both companies and customers may benefit much from reading product reviews. Reviewers frequently employ sarcasm, implied emotions, and nuanced expressions to communicate their experiences, making it difficult to precisely read the tone and ideas conveyed in reviews. Businesses may enhance their product development and marketing strategies, as well as acquire a better understanding of client feedback, by using our strategy.

D. USES IN PSYCHOLOGY AND MENTAL HEALTH

Implications for psychology and mental health may arise from the study of subtly conveyed feelings and attitudes in internet material. Research on the connections between online conduct, emotional expressions, and psychological health can benefit from our method. As a result, we may be able to create instruments for the early diagnosis of mental health problems and provide tailored assistance and therapies through the study of internet content.

E. MORAL REFLECTIONS

We must address the possible ethical difficulties and hazards of online content analysis, notwithstanding the prospective uses of our technique. Our technique processes personal data and possibly sensitive information, therefore protecting privacy and data is of the utmost importance. To safeguard user privacy and guarantee compliance with applicable rules, the necessary steps should be implemented.

The possibility of prejudice and abuse of the technology should also be carefully considered. To reduce the likelihood of biased or immoral results, automated systems for content moderation and opinion analysis should be built and implemented with openness, responsibility, and human supervision.

Responsible and effective application of our method and the mitigation of risks and obstacles need constant study, cooperation with subject experts, and the creation of governance structures and ethical standards.

7. CONCLUSION

In order to decipher subtle viewpoints stated across many internet platforms, we introduced a new machine learning method in this study. To address the difficulties of detecting sarcasm, recognizing emotions, and analyzing opinions across different online platforms and linguistic modalities, our method integrates deep learning, multi-modal analysis, and natural language processing.

We proved that our method successfully detected sarcasm, identified emotions, and analyzed sentiment in web material by conducting comprehensive tests on various datasets. Incorporating multi-modal data and a multi-task learning architecture allowed our model to take advantage of common representations and pick up on subtleties in subdued opinions.

By doing ablation studies, we were able to dissect our method and determine the relative merits of its constituent parts; these revealed that textual, visual, and auditory modalities all contributed to effective multi-task learning. Also, we analyzed the sample predictions qualitatively to show where our method excels and where it falls short, and we discussed the difficulties of effectively recognizing sarcasm and other nuanced expressions of opinion.

We spoke about how our work may be used in several fields, including brand management, social media monitoring, online content moderation, analysis of product evaluations, and mental health and psychology. Privacy, data protection, and the proper deployment of automated systems are of the utmost significance, and we have also recognized the hazards and ethical issues connected with online content analysis.

REFERENCES

1. Asr, F.T.; Taboada, M. The data challenge in misinformation detection: Source reputation vs. content veracity. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Brussels, Belgium, November 2018; 10–15.
2. Mukherjee, S. Sentiment analysis. In ML. NET Revealed; Springer: Berlin/Heidelberg, Germany, 2021; 113–127.
3. Tompkins, J. Disinformation Detection: A review of linguistic feature selection and classification models in news veracity assessments. arXiv 2019, arXiv:1910.12073.
4. Hepburn, J. Universal Language model fine-tuning for patent classification. In Proceedings of the Australasian Language Technology Association Workshop, Dunedin, New Zealand, 11–12 December 2018; 93–96.
5. Katwe, P.; Khamparia, A.; Vittala, K.P.; Srivastava, O.A. Comparative Study of Text Classification and Missing Word Prediction Using BERT and ULMFiT. In Evolutionary Computing and Mobile Sustainable Networks; Springer: Berlin/Heidelberg, Germany, 2021; 493–502.
6. Shu, K.; Bhattacharjee, A.; Alatawi, F.; Nazer, T.H.; Ding, K.; Karami, M.; Liu, H. Combating disinformation in a social media age. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2020, 10, e1385.
7. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. arXiv 2018, arXiv:1801.06146.
8. Chauhan, U.A.; Afzal, M.T.; Shahid, A.; Moloud, A.; Basiri, M.E.; Xujuan, Z. A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews. World Wide Web 2020, 23, 1811–1829.
9. Liu, B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions; Cambridge University Press: Cambridge, UK, 2020.
10. Zhao, W.; Peng, H.; Eger, S.; Cambria, E.; Yang, M. Towards scalable and reliable capsule networks for challenging NLP applications. arXiv 2019, arXiv:1906.02829.
11. Georgieva-Trifonova, T.; Duraku, M. Research on N-grams feature selection methods for text classification. In IOP Conference Series: Materials Science and Engineering; IOP Publishing: Bristol, UK, 2021; Volume 1031, p. 012048.
12. Chaturvedi, I.; Ong, Y.S.; Tsang, I.W.; Welsch, R.E.; Cambria, E. Learning word dependencies in text by means of a deep recurrent belief network. Knowl.-Based Syst. 2016, 108, 144–154.
13. Basiri, M.E.; Kabiri, A. HOMPer: A new hybrid system for opinion mining in the Persian language. J. Inf. Sci. 2020, 46, 101–117.
14. Abdar, M.; Basiri, M.E.; Yin, J.; Habibnezhad, M.; Chi, G.; Nemati, S.; Asadi, S. Energy choices in Alaska: Mining people's perception and attitudes from geotagged tweets. Renew. Sustain. Energy Rev. 2020, 124, 109781.
15. Cambria, E.; Li, Y.; Xing, F.Z.; Poria, S.; Kwok, K. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual, 19–23 October 2020; 105–114.
16. Peng, S.; Cao, L.; Zhou, Y.; Ouyang, Z.; Yang, A.; Li, X.; Jia, W.; Yu, S. A survey on deep learning for textual emotion analysis in social networks. *Digital Commun. Netw.* 2021, 8, 745–762.
17. Sharaf Al-deen, H.S.; Zeng, Z.; Al-sabri, R.; Hekmat, A. An Improved Model for Analyzing Textual Sentiment Based on a Deep Neural Network Using Multi-Head Attention Mechanism. Appl. Syst. Innov. 2021, 4, 85.
18. Singh, J.; Singh, G.; Singh, R. Optimization of sentiment analysis using machine learning classifiers. Hum.-Cent. Comput. Inf. Sci. 2017, 7, 1–12.
19. Dong, J.; Ding, C.; Mo, J. A low-profile wideband linear-to-circular polarization conversion slot antenna using metasurface. Materials 2020, 13, 1164.
20. Gupta, C.; Gill, N.S.; Gulia, P.; Kumar, A.; Karamti, H.; Moges, D.M.; Safra, I. A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features. *J. Big Data* 2025, 12, 1.
21. In Machine Learning Models and Algorithms for Big Data Classification; Springer: Berlin/Heidelberg, Germany, 2016; 207–235.
22. Pisner, D.A.; Schnyer, D.M. Support vector machine. In Machine Learning; Elsevier: Amsterdam, The Netherlands, 2020; 101–121.
23. Hope, T.; Resheff, Y.S.; Lieder, I. Learning Tensorflow: A Guide to Building Deep Learning Systems; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.

24. Tarasov, D. Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews. In Proceedings of the 21st International Conference on Computational Linguistics Dialogue, Sydney, NSW, Australia, July 2015; Volume 2, 53–64.
25. Tai, K.S.; Socher, R.; Manning, C.D. Improved semantic representations from tree-structured long short-term memory networks. arXiv 2015, arXiv:1503.00075.
26. Jiao, T.; Guo, C.; Feng, X.; Chen, Y.; Song, J. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Comput. Mater. Continua* 2024, 80, 1–35.
27. Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T. A Convolutional Approach for Misinformation Identification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; 3901–3907.
28. Czapla, P.; Howard, J.; Kardas, M. Universal language model fine-tuning with subword tokenization for polish. arXiv 2018, arXiv:1810.10222.
29. Zhang, J.; Cui, L.; Fu, Y.; Gouza, F.B. Fake news detection with deep diffusive network model. arXiv 2018, arXiv:1805.08751.
30. Rane, A.; Kumar, A. Sentiment classification system of twitter data for US airline service analysis. In Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 23–27 July 2018; Volume 1, 769–773.
31. Castro-Ospina, A.E.; Solarte-Sanchez, M.A.; Vega-Escobar, L.S.; Isaza, C.; Martínez-Vargas, J.D. Graph-based audio classification using pre-trained models and graph neural networks. *Sensors* 2024, 24, 2106.
32. Abdul-Mageed, M.; Novak, P.K. Deep Learning for Natural Language Sentiment and Affect. Available online: http://kt.ijs.si/dlsa/2018-09-14-ECML-DLSA-tutorial.pdf (accessed on 14 October 2021).
33. Rathi, M.; Malik, A.; Varshney, D.; Sharma, R.; Mendiratta, S. Sentiment analysis of tweets using machine learning approach. In Proceedings of the 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2–4 August 2018; 1–3.
34. Can, E.F.; Ezen-Can, A.; Can, F. Multilingual sentiment analysis: An rnn-based framework for limited data. arXiv 2018, arXiv:1806.04511.
35. Wang, J.; Yu, L.C.; Lai, K.R.; Zhang, X. Dimensional sentiment analysis using a regional CNN-LSTM model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; 225–230.
36. Singh, R.; Singh, R.; Bhatia, A. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. Int. J. Adv. Sci. Res. 2018, 3, 19–24.
37. Basiri, M.E.; Nemati, S.; Abdar, M.; Cambria, E.; Acharya, U.R. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Gener. Comput. Syst. 2021, 115, 279–294.
38. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised data augmentation for consistency training. arXiv 2019, arXiv:1904.12848.
39. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, Vancouver, Canada, 3–4 August 2017; pp. 502–518.
40. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
41. Aldayel, H.K.; Azmi, A.M. Arabic tweets sentiment analysis—A hybrid scheme. J. Inf. Sci. 2016, 42, 782–797.
42. Rani, S.; Singh, J. Sentiment analysis of Tweets using support vector machine. Int. J. Comput. Sci. Mob. Appl. 2017, 5, 83–91.
43. Agarwal, A.; Yadav, A.; Vishwakarma, D.K. Multimodal sentiment analysis via RNN variants. In Proceedings of the 2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD), Honolulu, HI, USA, 29–31 May 2019; 19–23.