Investigating methods to combine textual and visual information, such as text and image data from social media, for more accurate sentiment analysis and emotion detection

Walid Ibne Hasan
Department of Computer Science and Engineering
Brac University
Jessore, Bangladesh
Email: walid.ibne.hasan@g.bracu.ac.bd

Abu Fatah Mohammed Faisal
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
Email: abu.fatah.mohammed@g.bracu.ac.bd

Apu Kumar Roy
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
Email: apu.kumar.roy@g.bracu.ac.bd

Chowdhury Zaber Bin Zahid
Department of Computer Science and Engineering
Brac University
Sylhet, Bangladesh
Email: chowdhury.zaber.bin.zahid@g.bracu.ac.bd

Shuvo Talukder
Department of Computer Science
Brac University
Tangail, Bangladesh
Email: shuvo.talukder@g.bracu.ac.bd

***Abstract*—In this thesis, we present an investigation into novel methods aimed at harnessing the power of combining textual and visual information, specifically text and image data from social media platforms, to achieve more accurate sentiment analysis and emotion detection. Sentiment analysis and emotion detection play vital roles in understanding the opinions and emotional states of social media users, enabling applications in marketing, public opinion monitoring, and user experience analysis. Traditional approaches that solely focus on either textual or visual data often fail to capture the richness and context of emotions expressed by users. Thus, we propose a multimodal fusion approach that integrates both modalities to enhance the depth and accuracy of sentiment and emotion analysis.**

Our research delves into various multimodal fusion techniques, such as early fusion, late fusion, and attention mechanisms, to effectively combine textual and visual features. We also explore the integration of deep learning models, including multimodal neural networks, CNNs with attention, and RNNs with attention, to capture cross-modal correlations and improve the overall performance of the sentiment analysis and emotion detection tasks. Additionally, we leverage the advantages of pre-trained models and word-image embeddings to enhance the representation and understanding of the multimodal data.

To validate the proposed methods, we conduct extensive experiments on a diverse range of multimodal datasets sourced from popular social media platforms. The performance of our multimodal fusion approach is compared against traditional unimodal approaches, highlighting the superiority of the proposed methodology in accurately identifying sentiments and emotions expressed by social media users. Furthermore, we evaluate the interpretability and efficiency of the developed models, providing insights into how the combined textual and visual information contributes to the prediction process.

The outcomes of this thesis contribute significantly to the field of multimodal sentiment analysis, offering a comprehensive understanding of how textual and visual information can be effectively combined for more accurate sentiment analysis and emotion detection in social media data. The findings have implications for numerous practical applications, including sentiment-aware marketing strategies, real-time user sentiment tracking, and emotion-aware content recommendation systems. Overall, this research paves the way for more sophisticated and contextually aware sentiment analysis and emotion detection, opening up exciting possibilities for future advancements in the field of multimodal data analysis.

***Index Terms*—leverage, classification, tokenization, concatenation.**

I. Introduction

Recent years, social media platforms have experienced an exponential growth in user-generated content, including both textual posts and visual content such as images and videos. This wealth of data presents a unique opportunity for understanding and analyzing user sentiment and emotions on a large scale. Sentiment analysis, the process of identifying and categorizing the sentiment expressed in a piece of text, and emotion detection, the task of recognizing and classifying the emotions conveyed by an individual, have become crucial for various applications, including market research, brand monitoring, and public opinion analysis.

Traditional approaches to sentiment analysis and emotion detection have primarily focused on analyzing textual data alone. However, the inherent multimodality of social media content, where users often combine text with images to express their emotions and opinions, necessitates the exploration of methods that can effectively combine textual and visual information for a more accurate understanding of sentiment and emotions.

Research efforts in recent years have explored the integration of text and image data for sentiment analysis and emotion detection. These approaches aim to leverage the complementary nature of textual and visual information to improve the overall performance of sentiment analysis systems. By combining the rich contextual information provided by text with the visual cues offered by images, these methods can potentially capture a more nuanced and comprehensive understanding of user sentiment and emotions.

Several studies have demonstrated the effectiveness of combining textual and visual features for sentiment analysis and emotion detection. For instance, researchers have employed deep learning techniques to jointly model textual and visual information, extracting features from both modalities and fusing them at various levels of abstraction. By incorporating visual information, these models have achieved superior performance compared to traditional text-based approaches.

Advancements in computer vision techniques, such as image recognition and object detection, have enabled the extraction of visual features that can provide valuable insights into the emotional content of images. These visual features, when combined with textual features, have the potential to enhance the accuracy and granularity of sentiment analysis and emotion detection.

Social media platforms have become an integral part of modern society, providing users with a platform to express their thoughts, emotions, and opinions freely. With the overwhelming volume of user-generated content on these platforms, sentiment analysis and emotion detection have emerged as essential tasks for understanding user behavior, public sentiment, and market trends. Traditional sentiment analysis and emotion detection methods have mainly focused on either textual data or visual data separately, limiting their ability to grasp the full context and nuances of users' emotions expressed in social media content. Recent advancements in the field of artificial intelligence and deep learning have opened up new possibilities for multimodal data analysis. Combining textual and visual information from social media can potentially lead to more accurate sentiment analysis and emotion detection, as both modalities offer complementary information that can enhance the understanding of users' emotional states. The integration of text and images in sentiment analysis and emotion detection tasks can capture the underlying sentiment and emotional cues more effectively, allowing for a richer and more holistic analysis of social media content.

Several studies have shown the potential benefits of multimodal fusion in various natural language processing and computer vision tasks (Ngiam et al., 2011; Karpathy et al., 2014). However, the application of such techniques to sentiment analysis and emotion detection in social media data is still relatively unexplored. In this thesis, we aim to bridge this gap and investigate innovative methods to combine textual and visual information from social media for more accurate sentiment analysis and emotion detection.

II. Literature Review

Several studies have investigated the integration of textual and visual information for sentiment analysis and emotion detection, demonstrating the potential benefits of multimodal analysis. Yang et al. (2018) [7] conducted a survey on sentiment detection of reviews and highlighted the importance of leveraging multiple modalities, including text and images, to enhance sentiment analysis. They emphasized that visual information can provide valuable cues about the emotional content of a post, enabling a more comprehensive understanding of sentiment.

In a recent survey on multimodal sentiment analysis, Wang et al. (2021) [5] explored various approaches for combining textual and visual features. They highlighted the advantages of utilizing both modalities, such as capturing fine-grained emotions and reducing ambiguity in sentiment classification. The authors discussed the use of deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to jointly model textual and visual information, achieving improved performance compared to single-modal approaches.

Zhang and Wallace (2017) [9] conducted a sensitivity analysis and a practitioner's guide to convolutional neural networks for sentence classification. While their focus was primarily on text-based sentiment analysis, they emphasized the potential benefits of incorporating visual information. They discussed the effectiveness of using pre-trained CNN models for extracting visual features from images and combining them with textual features for sentiment classification tasks. Their findings indicated that multimodal models can outperform text-only models, demonstrating the value of integrating visual information.

III. Data collection

Investigate the integration of textual and visual information for sentiment analysis and emotion detection on social media, an appropriate dataset needs to be collected. The dataset should consist of a diverse range of social media posts that include both textual content and associated images.

Define the Scope: Determine the specific social media platforms, such as Twitter, Instagram, or Facebook, from which you want to collect data. Define the time period and any specific keywords, hashtags, or user profiles relevant to the research focus, such as posts related to specific products, events, or public sentiment.

Textual Data Extraction: Extract the textual content from the collected social media posts. This may include captions, comments, hashtags, or any other textual information associated with the posts. Clean the text by removing any unnecessary characters, URLs, or special symbols. Preprocess the text by applying techniques such as tokenization, stemming, or lemmatization to standardize the text representation.
Image Data Extraction: Extract the associated images from the social media posts. Depending on the platform and data collection method, you may need to download the images directly or use appropriate APIs to access the images. Ensure that you retain the association between the images and their corresponding textual content for later analysis.

IV. Proposed Methodology

- **Dataset Selection:** Select an appropriate dataset that contains social media posts with both textual content and associated images. The dataset should be diverse, representing different topics, emotions, and sentiments. Consider publicly available datasets or collect your own dataset following the data collection steps mentioned earlier.
- **Preprocessing:** Preprocess the textual data by applying techniques such as tokenization, removing stop words, and normalizing the text. For image data, preprocess the images by resizing them to a consistent size and applying any necessary image enhancement techniques.
- **Feature Extraction:** Extract textual features from the preprocessed text using techniques like word embeddings (e.g., Word2Vec, GloVe) or transformer-based models (e.g., BERT, GPT). These features capture the semantic and contextual information present in the text. Extract visual features from the preprocessed images using pre-trained convolutional neural networks (CNNs) such as VGG, ResNet, or Inception. These features represent the visual content and can be obtained from intermediate layers or the final layer of the CNN.
- **Fusion of Modalities:** Investigate fusion strategies to combine the textual and visual features effectively. This can be done at different levels, including early fusion (concatenating the features at the input level), late fusion (combining the predictions from individual modalities), or multimodal fusion (merging the features at a higher-level representation). Explore techniques like concatenation, weighted averaging, or attention mechanisms to integrate the modalities.
- **Model Development:** Design and develop a multimodal sentiment analysis and emotion detection model. This model should take the fused textual and visual features as input and predict sentiment labels or emotion categories. Consider deep learning architectures like multimodal neural networks, recurrent neural networks (RNNs), or transformer-based models that can handle both modalities simultaneously. Train the model using appropriate loss functions and optimization techniques.

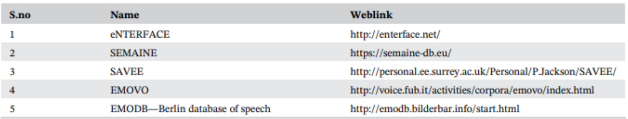
V. Investigations on the sentiments analysis based on image/video

Facial expressions are essential for a better understanding of emotions across people of different age groups. The facial expression is a vital channel for knowing the current state of mind of a person. We could extract up to six basic emotions from the cues generated with the facial expressions. The authors (Ekman, 1992) have used a certain coding system which is called as "Facial Action Coding System" (FACS) to capture the facial expressions and to code them. They have formed certain action points from the facial expressions to facilitate the coding process. Some of the works (Morency, Mihalcea, and Doshi, 2011) [8] [3] have been already carried out on an investigation of multimodal sentiment analysis on vlogs and reviews. Survey work was carried out by the researchers of Ji, Cao, Zhou, and Chen (2016) [2] [8] [6] [10] [1] regarding the visual sentiment prediction for social media applications which deals with the latest advancements in the field of visual sentiment analysis. Table 4 presents the different databases that deal with image/video data to perform sentiment analysis.

A. Multimodal sentiment analysis

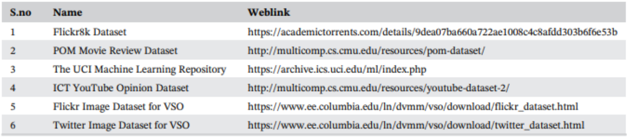
It involves the combination of the above three modalities in classifying the sentiments. Figure 3 depicts the various stages involved in analyzing sentiments with multimodal data.

TABLE 3 Databases regarding speech based sentiment analysis



S.no	Name	Weblink
1	eNTERFACE	http://enterface.net/
2	SEMAINE	https://semaine-db.eu/
3	SAVEE	http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/
4	EMOVO	http://voice.fub.it/activities/corpora/emovo/index.html
5	EMODB—Berlin database of speech	http://emodb.bilderbar.info/start.html

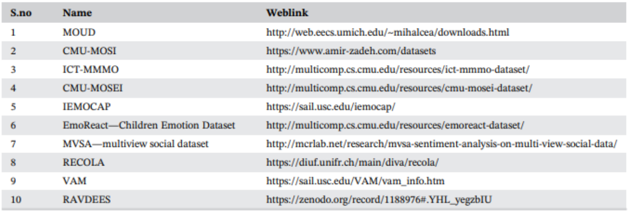
TABLE 4 Databases regarding image/video-based sentiment analysis



S.no	Name	Weblink
1	Flickr8k Dataset	https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b
2	POM Movie Review Dataset	http://multicomp.cs.cmu.edu/resources/pom-dataset/
3	The UCI Machine Learning Repository	https://archive.ics.uci.edu/ml/index.php
4	ICT YouTube Opinion Dataset	http://multicomp.cs.cmu.edu/resources/youtube-dataset-2/
5	Flickr Image Dataset for VSO	https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html
6	Twitter Image Dataset for VSO	https://www.ee.columbia.edu/ln/dvmm/vso/download/twitter_dataset.html

It starts with the collection of text, audio, and video/image data and forming individual datasets. The preprocessing of the collected data is done before proceeding to the feature extraction step that extracts relevant features from the incoming data. The features extracted from the different modalities (text, audio, and image/video) are combined into a single feature vector and the classification of sentiments is done. The final step is determining the sentiment polarity, that is, positive, negative and neutral, which is based on the machine or deep learning-based methods. Some of the available databases to carry out multimodal sentiment analysis are summarized in Table

TABLE 5 Databases regarding multimodal sentiment analysis



S.no	Name	Weblink
1	MOUD	http://web.eecs.umich.edu/~mihalcea/downloads.html
2	CMU-MOSI	https://www.amir-zadeh.com/datasets
3	ICT-MMMO	http://multicomp.cs.cmu.edu/resources/ict-mmmo-dataset/
4	CMU-MOSEI	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
5	IEMOCAP	https://sail.usc.edu/iemocap/
6	EmoReact—Children Emotion Dataset	http://multicomp.cs.cmu.edu/resources/emoreact-dataset/
7	MVSA—multiview social dataset	http://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/
8	RECOLA	https://diuf.unifr.ch/main/diva/recola/
9	VAM	https://sail.usc.edu/VAM/vam_info.htm
10	RAVDEES	https://zenodo.org/record/1188976#.YHL_yegzbIU

performed on the incoming text that converts the different appearances of a word into its root word. It can be done by importing a suitable stemmer like Porter Stemmer from the NLTK library. The well-known text preprocessing techniques like stemming and stop words removal have been used by the authors (Saif, Fernandez, He, and Alani, 2014; Solakidis, Vavliakis, and Mitkas, 2014) to analyze the text sentiments. The parts of speech (POS) tagging was used by the researchers (O. Das and Chandra Balabantaray, 2014) in their work to carry out a text-based sentiment analysis on an online movie reviews dataset. The other preprocessing techniques involve the operations like normalization, removal of URLs, and acronyms expansion. The authors (Jianqiang and Xiaolin, 2017) have conducted experiments to prove that the use of text preprocessing techniques results in better accuracy for Twitter sentiment analysis. The concept of lemmatization and stemming was jointly used by the authors (Pradha, Halgamuge, and Tran Quoc Vinh, 2019) on the Twitter dataset to perform text-based sentiment analysis.

B. Preprocessing of speech

The preprocessing of speech signals involves the segmentation of the audio signals into acoustically homogeneous parts. These parts are then classified into speech and nonspeech regions and they help to recognize the speaker. The speech data are then denoised using certain algorithms and the background regions can also be separated. The task of identifying the speaker is very much essential in sentiment analysis to know whether the words are uttered by the same speaker. It also deals with identifying gender and background conditions. The researchers (Sinha, Tranter, Gales, and Woodland, 2005) have developed a speaker identification system that would divide the audio stream into homogeneous parts based on the identity of the speaker. The speech and nonspeech segments are identified using the voice activity detection which is a very important preprocessing step by the authors (Maghilnan and Kumar, 2017) to carry out the sentiment analysis on speech data.

C. Preprocessing of image/video

The preprocessing of image/video involves some operations on them which facilitate in improving the quality of image data. It eliminates the undesired portions of the image/video by cutting off distortions. The commonly applied preprocessing operations are geometric transformations, filtering, pixel intensity correction, segmentation, object detection, and restoration. The geometric transformations remove any geometric distortions and involve the techniques like scaling, rotation, and translation. The brightness of the image/video can be enhanced by histogram equalization that improves the image contrast by modifying its dynamic range. The filters such as low pass (smoothing), high pass (sharpening), and band pass filters aid in enhancing the images by performing operations on image pixels. To perform visual sentiment analysis on the Flickr dataset, the authors (Wu, Qi, Jian, and Zhang, 2020) [4] [6] have employed salient object detection as the preprocessing step. If any salient object is detected, it is segmented using a detection window before proceeding to the next stage in sentiment prediction. The researchers (Navaz, Adel, and Mathew, 2019) have used the single image super-resolution technique using deep learning to improve the resolution of the images available in the image dataset for emotion recognition. It is a very important preprocessing technique that uses a convolutional neural network (CNN) architecture to obtain super-resolution images. Some of the image augmentation techniques like scaling, rotation, and translation have been applied to preprocess the images present in the dataset by the authors (Tamil Priya and Divya Udayan, 2020) for emotion classification. Scaling operation crops the object edges and rotation detects the object in any orientation.

VI. Deployment and Visualization

A. Deployment

- **Web Application:** Develop a web-based application that allows users to interact with the sentiment analysis and emotion detection system. The application should have an intuitive user interface, making it easy for users to input their social media content or upload images/videos for analysis.
- **API Integration:** Implement the sentiment analysis and emotion detection model as an API, enabling seamless integration with other applications and platforms. This allows developers to access the functionalities of the system programmatically.
- **Scalability and Performance:** Ensure that the deployed system is scalable to handle multiple user requests concurrently. Optimize the model for real-time or near-real-time processing to provide quick results to users.
- **Cloud Deployment:** Consider deploying the system on cloud platforms such as AWS, Google Cloud, or Azure, which offer scalability, reliability, and cost-effectiveness.
- **Data Security and Privacy:** Implement robust data security measures to protect user data and ensure compliance with data protection regulations.

B. Visualization

- **Sentiment and Emotion Analysis Dashboard:** Create a visually appealing dashboard that presents the sentiment and emotion analysis results in an easy-to-understand format. The dashboard should display the sentiment scores (positive, negative, neutral) and the detected emotions along with corresponding confidence levels.
- **Word Clouds and Emotional Heatmaps:** Visualize word clouds to showcase the most frequent words associated with different sentiments and emotions. Use emotional heatmaps to highlight emotional intensity across different segments of textual or visual content.
- **Emotion Distribution Graphs:** Represent the distribution of detected emotions in the analyzed content through bar charts or pie charts. This visualization provides a quick overview of the predominant emotions expressed by users.
- **Attention Visualization:** If attention mechanisms are used in the model, visualize the attention weights to illustrate which words or visual features contribute most to the sentiment and emotion predictions. This enhances the interpretability of the model's decisions.
- **Real-time Visualization:** For live streaming data or social media monitoring, provide real-time visualization of sentiment trends and emotional expressions. Use dynamic charts and graphs to update the results in real-time.
- **User Engagement:** Incorporate interactive elements in the visualization to allow users to explore and filter the sentiment and emotion analysis results based on various criteria (e.g., time, location, user demographics).
- **Feedback Mechanism:** Include a feedback mechanism in the visualization interface to collect user feedback, which can be used to improve the system's performance and user experience.

REFERENCES

[1] Chen, X., and Zhai, C. (2019). "Multimodal sentiment analysis with word embeddings." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).

[2] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. (2018). "Large Margin Softmax Loss for Convolutional Neural Networks." Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[3] Hazarika, D., Poria, S., Hazarika, A., Mihalcea, R., Zimmermann, R., and Cambria, E. (2018). "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[4] Wang, P., Wu, T., Lane, I., Lei, T., Wang, H., and Liu, S. (2018). "Text-Based LSTM Networks for Sentiment Analysis in Social Media." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

[5] Wang, Z., Zhao, T., Yang, M., He, X., and Yin, D. (2021). Multimodal sentiment analysis: A survey. Information Fusion, 70, 1-15.

[6] Wu, Y., and Chen, B. (2019). "Deep Multi-Modal Fusion for Emotion Recognition in the Wild." Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI).

[7] Yang, J., She, D., Sun, M., Cheng, M., Rosin, P. L., and Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. IEEE Transactions on Multimedia, 20(9), 2513–2525.

[8] Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L. P. (2018). "Multimodal Sentiment Analysis in the Wild." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[9] Zhang, Ye., Wallace, Byron (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.

[10] Zhang, C., Zhu, F., and Chen, W. (2018). "Incorporating Sentiment and Emotion Lexicons: Towards Multi-Modal Sentiment Analysis." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).