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 study, we tried to use some methods to detecting emotion and analysing sentiment based on the textual and visual information for social media. The study was challenging for us as we have to use the textual and visual data from a huge data set and then try to implement it using some methods or algorithms based on the accuracy, our analysis etc. The purpose of this study is to emphasize on the result of detecting emotion and sentiment from the textual and visual data by using different algorithms and methods.**

***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.

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) [2] 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) [1] 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) [3] 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. Propose 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.

REFERENCES

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

[2] 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.

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