Affective Computing Based Human Emotion Analysis Using Fusion of Image and Text Features

Contents

List of Figures

List of Tables

1 Affective Computing Based Human Emotion Analysis Using Fusion of Image and Text Features

Avantika, Aishvarya, Ayush, Aman, Saksham, Shailendra
Computer Science and Engineering Department,
Thapar Institute of Engineering & Technology, Patiala, India

CONTENTS

Abstract

In today's rapidly evolving social media landscape, multimodal content has given rise to affective computing, which offers a more complete and organic knowledge of human emotions and emotional responses. Affective computing research has progressed from traditional unimodal analysis to more complicated kinds of multimodal analysis as more and more multimodal posts, videos, GIFs etc. are shared online (e.g., on YouTube, Facebook, Twitter) for meme analysis, product reviews, movie reviews, political opinions, and more. Affective computing is the way to program machines to recognise human emotions and sentiments. This paper focuses on the multimodal meme analysis, an aspect of affective computing , which blends visuals with text that has been superimposed for identifying hate speech emotions in the posts.. Memes are famous for quickly transferring concepts, information, and content from the physical world to the digital one using a variety of media types, including videos, GIFs and other kinds of multimodal posts. The aim of this study is to analyze affective computing in terms to get more understanding in identification and examining the hate speech or offensive content in multimodal memes. As multimodal memes often come in a multimodal manner, it becomes challenging to accurately classify the right emotions behind them. We reviewed the most recent studies in this area in depth, evaluating and synthesizing the suggested strategies. We evaluated these strategies' advantages and shortcomings critically. Even though there are few studies and datasets particularly created for hate speech detection and sentiment analysis in memes, we noticed an evolution in the applied approaches that produced better results despite the gaps in the literature and the scarcity of available datasets. It provides a thorough analysis of the multimodal meme identification and suggests the possible directions for future research.

1.1 INTRODUCTION

Multimodal memes, an aspect of affective computing resemble cultural inventions that spread swiftly through the internet. They are used in online communities to convey thoughts, feelings, and opinions and can take the shape of images, text, or videos. Because memes are based on experiences or knowledge that we all share, they are frequently humorous, snarky, or relatable. Memes are the challenging task to use for affective classification problems as usually they are analyzed for their textual content only [60, 53, 4] without taking into account the meanings of each, the written and visual modes and how they work together [33]. They encourage conversation, laughter, and interpersonal connections. Memes are short and funny, making them easy to share. They now play a significant role in internet culture, affecting the way that society views and discusses many topics from distant areas. However, some people use memes to spread hate, which could have a bad impact on people's emotion and should be kept secret from the public to preserve social harmony and peace. Fostering a secure and welcoming online community requires precisely analyzing and comprehending the sentiment communicated in memes. Facebook introduced the "hateful meme challenge" as part of its initiative to control the dissemination of offensive memes [33] with 10K+ offensive memes. This prompted the creation of numerous multimodal deep-learning algorithms for categorizing nasty memes [15, 40].

It is essential to keep in mind that the context and perception of memes might differ from emotion to emotion, and that humor is a personal experience. Figure 1 displays one of the example from the Hateful meme dataset. It's probable that some people will find amusement in this particular meme , but others will find it insulting because of the connotations it conveys about the girl and the comparisons it draws between her and other things. When a person, particularly a female, is referred to as a "broken sandwich maker," her value is diminished to that of a simple item, and damaging stereotypes are further spread. Memes that objectify or demean people on the basis of their gender, appearance, or any other attribute may contribute to a culture of negativity and discrimination on the internet. It is of the utmost importance to disseminate memes and other forms of internet material that encourage inclusiveness, respect, and compassion. As such memes not only spread hatred among people but also affects the mental health of the people which sometimes could result into sucidal activities also. Everyone will benefit from a friendlier and more pleasurable online environment if we promote humor that is both good and helpful and encourage its use.

1.1.1 MOTIVATION

In recent years, memes have emerged as a powerful cultural phenomenon, playing a significant role in shaping digital communication and online interactions. The widespread popularity of memes calls for a deeper understanding of their influence on society. Memes have become an integral part of modern digital culture and communication. By studying memes, you engage with a relevant and dynamic subject

Figure 1.1 Example of Hateful Meme

matter that is widely consumed and shared online. Memes reflect societal trends, humor, and shared experiences. Analyzing memes allows you to delve into popular culture, exploring how humor and ideas are transmitted and understood in the digital era. Researching memes provides an opportunity to bridge various disciplines such as communication, sociology, psychology, linguistics, and media studies. It allows you to examine the social, psychological, linguistic, and cultural dimensions of meme creation, dissemination, and impact. While memes are pervasive, in-depth academic research on meme analysis is still relatively limited. Writing a chapter on meme analysis allows you to contribute to this growing field, bringing new insights, theories, and methodologies to the table. Meme analysis has practical applications in areas such as marketing, advertising, social media strategies, and political campaigns. Understanding the mechanisms of meme virality and impact can provide valuable insights for practitioners and professionals in these domains. Studying memes challenges you to think critically and creatively about internet culture, humor, and communication. It encourages you to develop analytical skills and explore innovative research methods to uncover the underlying dynamics of meme production and consumption. By investigating the dynamics of memes, we seek to uncover the underlying social, psychological, and linguistic factors driving their widespread adoption and engagement. Understanding the mechanisms behind memes can enhance our comprehension of online communication, cultural trends, and the evolving nature of internet culture. Additionally, researching memes can be fun and engaging, as you get to analyze funny and relatable content. Furthermore, the findings can have practical applications in fields like marketing, advertising, and social media strategies. Overall, writing a research paper on analyzing the human emotions of hatred through memes allows you to explore a unique and relevant topic, gain valuable insights, and contribute to the growing body of knowledge in this field.

1.1.2 CONTRIBUTION

This chapter aims to develop a comprehensive taxonomy and classification system for memes. It focuses on analyzing the sentiment expressed in memes. By examining the affective computing aspect of meme analysis that incorporates fusion of text and image features, it aims to understand the emotions, attitudes, and opinions conveyed through memes. The main contribution of the paper can be summarized as follows:

Introduces and explores the application of attention mechanisms in meme analysis.
Examines non-attention mechanisms in meme analysis, exploring alternative methods for feature extraction, representation, and classification that do not rely on attention mechanisms.
Provides a comprehensive comparison of different approaches and their effectiveness in meme analysis.
Offers a detailed explanation of data representation techniques for both text and image data in meme analysis.
Explores various methods for data representation.
Examines the use of fusion techniques in meme analysis, which combine different modalities such as text and image to improve classification accuracy and obtain deeper understanding.
It discusses several applications for Meme Classification,
In addition, we have studied the datasets for memes that are critical to the classification of memes as well as the comparison of different approaches in the field.

These contributions are helpful in performing a better approach for meme classification, as it emphasizes the significance of efficient data representation methods and classification models.The structure of the paper is shown in Figure 2.

.

1.2 DEEP LEARNING-BASED METHODS

Deep learning is a type of machine learning that uses artificial neural networks to learn and make predictions from data. It's like teaching a computer brain to understand and solve problems. These networks are made up of many layers of connected nodes called neurons. They learn by adjusting their internal settings based on lots of example data. For example, they can learn to recognize images or understand language. Deep learning is very good at handling complicated tasks like identifying objects in pictures or understanding speech. It can automatically find important features in the data without needing humans to tell it what to look for. The power of deep learning comes from its ability to learn in layers. The first layer might learn basic shapes, like edges or corners, while later layers learn more complex patterns. This allows the network to understand and recognize things in the data. There are different types of deep learning networks. For images, we use Convolutional Neural Networks (CNNs), for sequences of data like language, we use Recurrent Neural

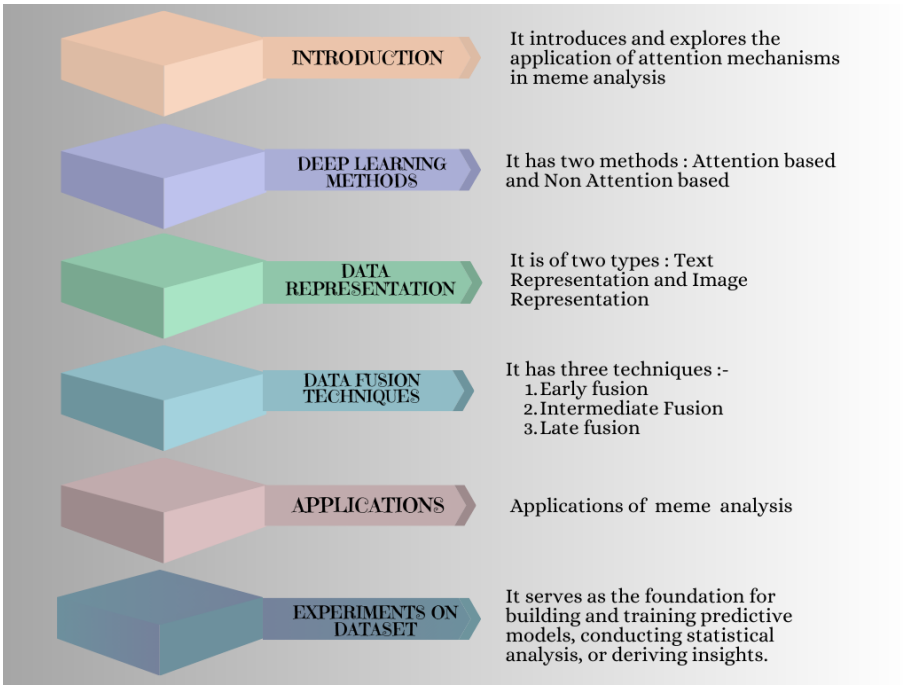

Figure 1.2 Structure of the chapter

Networks (RNNs), and for tasks involving both language and images, we use Transformers. Although deep learning has been very successful, it also has challenges. It needs lots of labeled data to learn effectively, and sometimes it can learn too much from the training data and not generalize well to new examples. Understanding how the network makes decisions can also be difficult. Researchers are working on solutions to these challenges, and deep learning continues to make incredible progress. It has revolutionized many fields and is helping us solve complex problems and build intelligent systems that can learn and make predictions from data. Deep learning (DL) is a type of machine learning that takes inspiration from how our brains process information. It doesn't rely on explicit rules created by humans; instead, it uses a vast amount of data to learn patterns and associate inputs with specific labels. DL is built using layers of algorithms known as artificial neural networks (ANNs). Each layer of the network interprets the data differently, helping the system understand and make predictions based on the input it receives [38, 84]. Deep learning (DL) can automatically learn and generate sets of features for various tasks, which is different from traditional machine learning (ML) methods that require manual feature engineering [65]. Deep learning allows us to learn and classify things in one go. It has become really popular in machine learning because of the big data boom and how the field has grown and changed [49, 25]. This section further examines the Deep Learn-

ing models for meme analyses from two aspects : Attention based and Non-attention mechanisms.

1.2.1 ATTENTION BASED

The attention mechanism, like other neural networks based techniques, tries to mimic how the human brain and vision process information. Human vision only concentrates on the specific areas of the image; it does not analyze the complete image at once. This results in a ”high-resolution” perception of the concentrated areas of the human vision space and a ”low-resolution” perception of the surroundings. In other words, it minimizes the irrelevant sections and gives them lower weights, giving the relevant parts a higher weight. Instead of analyzing the full vision space, this enables the brain to precisely and efficiently process and concentrate on the most crucial elements. Researchers created the attention mechanism as a result of this aspect of human vision. Created in 2014 for applications involving natural language processing [8], it has now been extensively used for a variety of tasks [22], particularly those involving computer vision [51]. It has been suggested that it could improve systems that are mostly reliant on CNN [42].

It has also been applied in conjunction with graph neural networks [79] and recurrent neural network models [3]. Giving various pieces of information varying weights is the primary concept behind the attention mechanism. The unique characteristics of memes, combining textual and visual elements, make them an intriguing subject for attention-based techniques. In a study by [48], a novel approach is introduced that bridges the gap between recurrent and recursive neural network architectures, which are prominent types of deep neural networks. The proposed method, known as the neural tree indexer (NTI), is a tree-structured model that operates independently of syntactic parsing. It learns to represent both a premise and a hypothesis, and then leverages an attention mechanism to combine these representations. As a result, the NTI achieves enhanced precision in various inference and classification tasks, introducing a new level of effectiveness in these domains. In their study, [80] introduce a model called the hierarchical attention network (HAN) designed specifically for document classification tasks. This model takes into account the hierarchical structure of documents, recognizing that words and sentences possess varying levels of informativeness. To address this, the HAN incorporates two levels of attention mechanisms at the word and sentence levels. This enables the model to assign varying degrees of attention to different content elements, effectively constructing a comprehensive representation of the document. By leveraging contextual information, the HAN identifies relevant sequences within the document. The model utilizes a Gated Recurrent Unit (GRU) as the encoder, and experimental results demonstrate that the proposed approach outperforms previous methods on six different datasets. In their study, authors [82] present an enhanced model that introduces a hierarchical iterative attention approach for aspect-based sentiment analysis. They approach the task of aspect-based sentiment analysis by formulating it as a machine comprehension problem. By employing their proposed model, they were able to achieve superior performance compared to the baseline hierarchical attention network. The research

[39] presents an approach for discerning keywords that distinguish between positive and negative sentences, employing a weakly supervised learning method based on a convolutional neural network (CNN). The CNN model is trained on sentence matrices, and subsequently, a word attention mechanism is incorporated. This mechanism employs class activation maps (CAM) to identify words that contribute significantly, utilizing the weights obtained from the fully connected layer at the end of the CNN. Notably, this method yields both sentence-level and word-level polarity scores, leveraging only weak labels, i.e., the polarity of the sentence from the dataset.

Furthermore, Lin et al. [41] in their study presented a novel method for extracting sentence embedding by incorporating a self-attention mechanism. This technique generates sentence embeddings depicted as 2-D matrices, where each row corresponds to a particular sentence segment. The sentence is initially processed using a recurrent neural network (RNN). Due to the inclusion of a penalization term, multiple attention values are subsequently learnt for each RNN state, allowing each attention vector to focus on distinct portions of the sentence. This method achieves exceptional precision in sentiment classification and textual entailment tasks. By leveraging attention-based mechanisms, these studies highlight the potential of such techniques in meme analysis. Attention allows for a more fine-grained examination of relevant visual and textual components within memes, offering a means to capture their unique attributes and contribute to their effective interpretation. These attention-based models not only aid in meme classification but also provide valuable insights into the mechanisms underlying humor, sentiment, and social dynamics in meme culture. As research in attention-based techniques progresses, there is a growing potential to further advance the field of meme analysis. The ability to precisely identify and weigh key elements in memes holds promise for improving our understanding of internet culture, humor, and the broader social implications of meme communication. Continued exploration of attention mechanisms in meme analysis is essential in unlocking new avenues of research and enhancing our comprehension of the intricate interplay between text, images, and sentiment in this digital medium.

1.2.2 NON-ATTENTION BASED

Deep learning techniques that lack explicit attention mechanisms in their architecture or processing are commonly known as non-attention-based approaches. These methodologies do not employ attention mechanisms in a direct manner; rather, they depend on alternative approaches to collect significant data and generate predictions. Feature extraction is another common operation, and deep learning models carry it out by using the deep layers as a set of feature extractors.

The researchers of [69] presents a variety of deep learning models combinations used to extract textual and visual data for the purpose of identifying offensive memes. GloVe is first utilised to get word vector representations for the textual features. To generate word embeddings, GloVe is an unsupervised learning system that has been taught to collect global word co-occurrence information. Long Short-Term Memory (LSTM) [83], Bidirectional LSTM (BiLSTM) [13], and Convolutional Neural Network (CNN) are then used separately to extract textual features. Multimodal ap-

proaches are ideal for detecting offensive memes, and deep learning feature extractors are widely believed to be able to provide features that are representative of the whole. Contrary to expectations, this was not the case in this study. The authors make the point that human moderation is still necessary.

Sabat et al. in their study [61] acquired textual and visual features independently. Initially, the text is extracted from the image through the application of the Tesseract Optical Character Recognition (OCR) technique. Next, the process involves extracting both visual and textual features. The model employed for extracting features from text and generating a 768-dimensional feature vector is known as Bidirectional Encoder Representations from Transformers (BERT) [16]. The authors utilize the VGG-16 [67] model to extract visual features, resulting in a feature vector with a dimensionality of 4096. Subsequently, the two feature vectors are combined through early fusion. The feature vector, consisting of 4864 dimensions, is ultimately utilized as input to a fully connected Neural Network. This network is responsible for generating a score that indicates the degree of hate speech present in the meme.

The findings obtained by [74] exhibit disparities in comparison to previous studies. In order to identify instances of hate speech within memes, the authors propose a methodology that utilises a single model known as VGCN-BERT [81] to extract textual attributes, alongside three CNNs namely ResNet50, ResNet152, and VGG-16, to extract visual features. In this iteration, the text embedding and each image embedding are combined in pairs through early fusion, resulting in the creation of distinct multimodal models. The findings derived from an analysis of a dataset comprising 1600 memes pertaining to Italian political matters demonstrated that the VGCN-BERT + ResNet50 multimodal model outperformed unimodal techniques, exhibiting an AUCROC value of 81.69. The authors acknowledge the presence of certain issues in their study, including the use of abbreviated language in certain memes and the potential subjectivity introduced by the annotators of the dataset. However, it is worth noting that despite these limitations, the multimodal approach employed in the study demonstrated superior performance compared to the unimodal technique.

These approaches highlight different strategies for meme classification without explicitly incorporating attention mechanisms. They explore the utilization of deep learning architectures, convolutional neural networks, recurrent neural networks, ensemble models, feature engineering, and traditional machine learning algorithms. Each approach provides valuable insights into non-attention-based mechanisms and their effectiveness in capturing relevant features and improving classification performance in the context of meme analysis.

1.3 DATA REPRESENTATION

Data representation is of paramount importance in meme analysis using deep learning techniques. Meme analysis involves understanding and interpreting the content, context, and sentiment conveyed by memes, which are typically image-based or multimedia artifacts. Effective data representation facilitates the extraction of meaningful features from memes and enables deep learning models to comprehend and analyze them accurately. Choosing an appropriate representation that captures the relevant

Table 1.1
Comparison between different representation techniques

Representation Technique	Context Sensitive Embedding	ML Based	Transformer Based	Pretrained	Applications
Word2Vec	No	Yes	No	Yes	[32, 66]
GloVe	NO	Yes	No	Yes	[63, 1, 24]
ELMo	Yes	Yes	No	Yes	[70, 35]
FastText	No	Yes	No	Yes	[71, 36]
BERT	Yes	Yes	Yes	Yes	[70, 28, 10]
CNN	No	Yes	No	Yes	[2, 7]
GAN	No	Yes	No	No	[85]
Transfer learning	NO	Yes	Yes	Yes	[63, 71]
Transformer-based Models	Yes	Yes	Yes	Yes	[58, 75]
Autoencoders	No	Yes	No	No	[52, 27, 26]

features, multimodal interactions, semantics, and promotes generalization is crucial for developing accurate and robust deep learning models for meme analysis tasks. The research paper [30], used separate word or image representations as input for the classification models. When these representations provide a deeper understanding of the words or images, it is expected that the predictive performance will improve. Consequently, it is crucial to leverage superior representation techniques as they directly impact the overall performance of the model. There are different techniques for text and image representation. For text data representation we have: Word2Vec, GloVe, ELMo, fastText, and BERT technique, and for image data representation we have three subcategories, Raw pixel values, Fixed-size vectors, Convolutional Neural Networks (CNNs), Transfer learning, Data augmentation, Autoencoders.

1.3.1 TEXT DATA REPRESENTATION TECHNIQUES

Text data representation in deep learning refers to the process of transforming raw textual data into a numerical format that can be effectively processed by deep learning models. Text data representation plays a crucial role in meme analysis using deep learning techniques. Effective representation methods capture the semantic and contextual information in meme text, enabling accurate understanding and classification. Techniques like word embeddings, such as Word2Vec or GloVe, convert text into dense vector representations that capture word semantics. Additionally, more advanced methods like BERT or Transformer models can capture contextual information through pre-training on large corpora. These representations enable deep learning models to effectively learn from and analyze meme text.

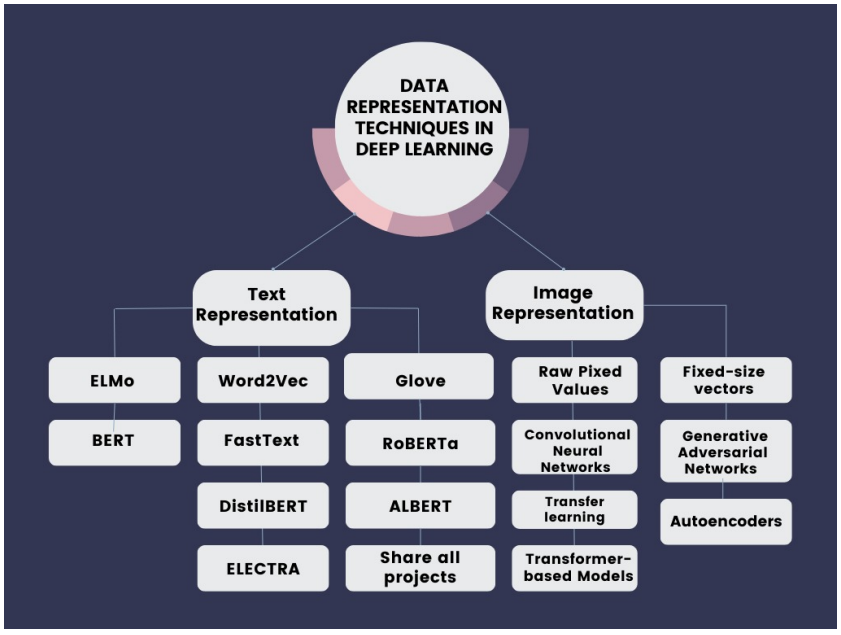

Figure 1.3 Types of Data Representation

1.3.1.1 Word2Vec

Word2Vec [45] is a method for recreating word linguistic contexts. The technique utilizes a neural network comprising two layers. We use a large collection of words as input, and it creates a space with lots of different directions. Each unique word in the collection gets its own space in this system. The words in the collection are organized so that words with similar meanings or usage tend to be grouped. Word2Vec is a fast way to learn word meanings from plain text. It has two methods: the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model. As mentioned by [77], to use machine learning for NLP tasks, it needs to convert text into vectors.There are two ways to do this: one-hot representation and word embeddings. In one-hot representation, a long vector is used, and its length is equal to the size of the dictionary used in the text. Relying solely on word vectors in one-hot representations can make it challenging to establish the relationship between words. Another method called distributed representation, which has shown the best performance in deep learning, involves assigning a fixed-length vector to each word. These vectors are distributed to create a vector space [46]. A word vector is a representation of a word as a set of numbers that capture its meaning and important features. These word vectors are created using unsupervised neural network models trained on large amounts of text data. Word2Vec is a specific type of neural network that helps process the text before it is used by deep learning algorithms. It helps to convert words into meaningful numerical representations that can be used for various language tasks [64].

1.3.1.2 GloVe

GloVe [54] is an unsupervised ML technique that stands for Global Vector for Word Representation. Stanford has made GloVe to create word embeddings. They did this by looking at how often words appear together in a big group of text. The result is that words are put into groups based on how they are related. This helps us see interesting patterns in how words are connected. The gloVe is a technique that helps us understand and represent words as vectors using unsupervised learning algorithms. This technique looks at how often words appear together in a text and creates a special matrix based on this information. This matrix helps us train the model and extract important word features. Although collecting this information from a large text can take time, it's a one-time process. After that, training becomes faster because we only need to work with a smaller set of entries in the matrix. Overall, GloVe helps us create meaningful word representations efficiently. GloVe models [54, 18] have been proven to perform better in many tasks related to understanding language compared to models based on word2vec.

1.3.1.3 ELMo

The acronym ELMo [55] stands for Embeddings from Language Model. This method is used to represent words as their corresponding sequence of vectors. To create word-level representations, the inputs consist of tokens at the character level, which are then used in a bi-directional LSTM. ELMo is a model used to convert words into numbers. ELMO embedding is a popular and widely used method in Natural Language Processing (NLP) research. ELMO embedding captures the complexity of words, including their syntax and meaning, and also considers how these meanings change in different contexts. It is known as a deep contextualized word representation and has been helpful in NLP research. Paper [29] uses ELMO, Text summarizing means taking a long piece of writing and creating a short version that still contains the important information. There's a way to do this automatically using a special tool called an Automatic Text Summarizer. This tool analyzes the big chunk of text and generates shorter summaries with valuable information. There are two types of Automatic Text Summarizers: 1) Extractive Text Summarizer, and 2) Abstractive Text Summarizer. The article focuses on the first type, the Extractive Text Summarizer. In this approach, the tool picks out the most relevant sentences from the original text to create the summary. To improve the Extractive Text Summarizer, the article suggests using something called "Elmo embedding." This is a type of contextual embedding that other researchers have used for the second type, the Abstractive Text Summarizer, but this article explores using it for the Extractive Text Summarizer instead.

1.3.1.4 FastText

FastText is a data representation technique used in deep learning. It represents words as continuous vectors, capturing their semantic meaning and relationships. It extends traditional word embeddings by considering subword information, such as character n-grams. This enables FastText to handle out-of-vocabulary words and capture mor-

phological information. By learning embeddings for both words and subwords, FastText provides richer representations that enhance the performance of deep learning models on various natural language processing tasks. The Facebook research team created FastText [31] as a library. There are two ways this method is helpful. The first way is that it helps us learn the meaning of words smartly and quickly. The second way is that it helps us organize sentences into different categories. This method can work with both supervised and unsupervised representations of words and sentences. Social media uses this for providing you with better ads related to your interest or search history. Facebook uses it according to your timeline updation and shows you the ads respectively.

1.3.1.5 BERT

Bidirectional Encoder Representation from Transformers (BERT) [16] is based on the transformer architecture. BERT makes use of a Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. BERT is designed to understand the context and meaning of words in a sentence by considering both the preceding and following words. It utilizes a transformer-based neural network architecture that enables it to capture complex relationships and dependencies between words. The BERT architecture is built upon the transformer model. BERT comes in two versions: BERT base and BERT large. BERT base consists of 12 layers of transformer blocks, 12 attention heads, and 110 million parameters. On the other hand, BERT large has 24 transformer layers, 16 attention heads, and 340 million parameters. BERT underwent pre-training on two natural language challenges. The first task is called Masked Language Modeling (MLM), where it learns about the connections between words. The second task is Next Sentence Prediction (NSP), which helps BERT understand how sentences are related to each other.

1.3.1.6 RoBERTa

Robustly Optimized BERT [42] proposed an improved version for training BERT models, namely RoBERTa. RoBERTa has made some changes to improve its performance. First, it is trained using a larger amount of data and for a longer time using bigger groups of data called batches. Second, it no longer focuses on predicting the next sentence. Third, it is trained on longer sequences of words. Lastly, the way words are hidden or masked during training is adjusted to change over time. It represents text data by learning contextualized word embeddings. It captures the meaning and relationships between words by considering the surrounding context, enabling a more accurate and nuanced understanding of a text.

1.3.1.7 ALBERT (A Lite BERT)

ALBERT (A Lite BERT) is a model that represents data by using transformer-based architectures. It improves upon BERT by reducing memory requirements and training time. ALBERT achieves this by sharing parameters across layers, making it more efficient while maintaining high performance in various natural language processing

tasks. [37] Improving the model performance is not always possible due to GPU/TPU memory limitations and longer training times. To mitigate the issue, the authors reduced two parameters to lower the memory consumption and to increase the training speed of BERT. Several studies demonstrate that ALBERT performs better than BERT when tested on GLUE, RACE, and SQuAD benchmarks.

1.3.1.8 ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

[14] BERT changes the input by replacing certain words with a placeholder called MASK. It then trains a model to predict the original words. The models replaced the words with different options that make sense, which greatly improves how well the model performs. It represents data by training a generator and discriminator together. The generator creates fake data samples, and the discriminator tries to distinguish them from real data. This process helps the model learn meaningful representations of the data. ELECTRA has shown impressive performance in various natural language processing tasks, including text classification and named entity recognition.

1.3.1.9 DistilBERT (Distilled BERT)

[62] DistilBERT is a smaller, faster version of the BERT model, which is a widely used language representation model. It is pre-trained and fine-tuned to perform well on various tasks. DistilBERT is specifically designed to be used in environments with limited resources. It achieves comparable performance to the original BERT model while using fewer resources. During its pre-training phase, knowledge distillation was used, reducing the size of the general BERT model by 40%. Despite this reduction, DistilBERT still maintains a strong language understanding capability of 97%. As a result, DistilBERT is not only faster and smaller but also more efficient to pre-train.

1.3.2 IMAGE DATA REPRESENTATION TECHNIQUES

Effective image data representation captures relevant visual features and semantic information, enabling accurate meme classification, generation, and understanding. Various techniques have been employed, including convolutional neural networks (CNNs) for feature extraction and encoding, as well as graph-based approaches to model image-text relationships. Proper representation facilitates tasks such as meme classification, sentiment analysis, and meme generation, improving overall performance.

1.3.2.1 Raw pixel values

In this approach, each pixel in the image is considered as a feature. The pixel values are directly used as inputs to the deep learning model. Raw pixel values in deep learning refer to the direct use of pixel intensity values from an image as input for training deep neural networks. In this approach, each pixel's grayscale value or the

RGB values of each pixel in an image are used as input features without any preprocessing or transformation. Using raw pixel values as input in deep learning models allows the network to learn directly from the visual information present in the images. The model can capture patterns and features by examining the pixel values and their spatial arrangement. While using raw pixel values can be a straightforward approach, it may have limitations in capturing higher-level structural information or being sensitive to variations in lighting conditions. Therefore, additional preprocessing or feature extraction techniques are often employed to enhance the representation of the image data.

1.3.2.2 Fixed-size vectors

In deep learning, fixed-size vectors refer to a method of representing data, such as images, using vectors of a fixed length. This approach allows for efficient processing and analysis of complex information. Images can be represented as fixed-size vectors by extracting features from different layers of a pre-trained convolutional neural network (CNN). These features capture important visual patterns and can be fed into the deep learning model. When it comes to image representation, fixed-size vectors are used to capture the important features and patterns within an image. Instead of considering each pixel, the image is transformed into a vector of a predefined length. Each element of the vector represents a specific aspect or characteristic of the image. By representing images as fixed-size vectors, deep learning models can process and analyze them more effectively. These vectors act as condensed representations that capture the essential information of an image in a compact form. For instance, let's consider a 100x100 pixel image. Each pixel contains information about the color intensity and location. Instead of feeding each pixel as a separate input to the deep learning model, we can transform the image into a fixed-size vector of, say, 1000 elements. Each element in the vector could represent a certain region or feature of the image. This fixed-size vector representation allows the deep learning model to focus on the most relevant aspects of the image, making the computation more efficient. It also helps in reducing the dimensionality of the data, which can be beneficial for training and inference processes. Using fixed-size vectors in deep learning image representation simplifies the complexity of images, making them more manageable for the neural network to process. It provides a concise yet meaningful representation of the image data, enabling the model to make accurate predictions or classifications. While the concept of fixed-size vectors is widely used in deep learning, it's important to note that some different techniques and architectures can be employed based on the specific task and domain of the application.

1.3.2.3 Convolutional Neural Networks (CNNs)

CNNs are specialized deep learning architectures designed for image processing tasks. They consist of multiple convolutional layers, pooling layers, and fully connected layers. CNNs learn hierarchical representations of images, capturing local patterns and global structures [34]. Convolutional Neural Networks (CNNs) are a

type of neural network that processes information in a specific way. They have different layers like convolution, fully connected, relevance weights, and pooling. CNNs are known for being easier to train and requiring less training data compared to other neural network approaches. They also have fewer parameters and connections, which makes them efficient [77]. CNN is effective in sentiment analysis or opinion-mining tasks. It has achieved excellent performance when working with text data [59].

1.3.2.4 Generative Adversarial Networks (GANs)

GANs are generative models that learn to generate new samples by training a generator network to produce realistic data, while simultaneously training a discriminator network to distinguish between real and generated samples [25]. GAN (Generative Adversarial Network) image representation in deep learning for memes involves using a specific type of neural network model to generate or transform memes. GANs consist of two components: a generator network and a discriminator network. The generator network generates new meme images based on random input, while the discriminator network tries to distinguish between real and generated memes. In simple terms, GANs learn to create realistic and meaningful meme images by training the generator and discriminator networks together. The generator network learns to produce convincing memes that resemble real ones, while the discriminator network learns to differentiate between real and generated memes. As training progresses, the generator network gets better at creating memes that fool the discriminator. The goal of GAN image representation for memes is to generate new and creative meme variations. By training the GAN on a large dataset of existing memes, the generator network learns to capture the visual style, humor, and characteristic elements of memes. This allows it to generate new memes that have similar attributes. The use of GANs in meme generation or transformation can provide a means for automating meme creation or exploring different variations of existing memes. It can be a useful tool for meme creators, designers, or researchers interested in understanding and manipulating visual humor. It's important to note that the field of GAN-based meme generation is evolving, and there may be more recent research papers and techniques available. Exploring the latest literature and research papers in the field would provide more detailed and up-to-date information.

1.3.2.5 Transfer learning

This technique involves using a pre-trained CNN on a large dataset (e.g., ImageNet) and then fine-tuning it on a specific image classification task. By leveraging the knowledge learned from the large dataset, transfer learning enables the effective representation of image data with limited training data. Transfer learning in deep learning is a technique where knowledge gained from pre-training a model on one task is transferred and applied to a different but related task. In the context of image representation for memes, transfer learning can be used to leverage pre-trained models on large-scale image datasets and apply their learned representations to meme classification tasks. When it comes to memes, they often contain unique visual and contextual

elements that make them challenging to classify accurately. However, by using transfer learning, utilize the knowledge learned from pre-training on general image recognition tasks and adapt it to meme-specific classification. Paper [57] presents MemeSem is a smart computer program that can understand the emotions behind memes, those funny pictures or videos you see online. It uses two special types of learning to do this. The first is called VGG19, which helps it understand the visual parts of the memes. The second is called BERT, which helps it understand the words in the memes. These two learnings work together to make predictions about the feelings expressed in the memes. To see how well MemeSem works, the researchers compared it to other similar programs. They used a dataset of 10,115 memes with three sentiment classes: positive, negative, and neutral emotions. MemeSem did better than the other programs. On average, it outperformed the other programs by 10.69% when compared to models that only looked at images, and by 3.41% when compared to models that only read the text. Transfer learning allows the model to learn general features from the pre-training task, such as recognizing shapes, edges, and textures, which apply to meme classification. By fine-tuning the pre-trained model on the meme dataset, the model can learn to capture the specific visual and contextual characteristics of memes, leading to improved classification accuracy. By leveraging transfer learning, the model benefits from the knowledge gained during pre-training, reducing the need for large amounts of labeled meme data. This approach saves computational resources and time while still achieving competitive performance on meme classification tasks. In conclusion, transfer learning in deep learning for meme image representation enables the utilization of pre-trained models' knowledge to improve the accuracy of meme classification. It allows the model to capture both general visual features and specific meme characteristics, leading to more effective and efficient classification.

1.3.2.6 Transformer-based Models

Originally designed for natural language processing, Transformer-based models have been adapted for image data representation. They utilize self-attention mechanisms to capture global relationships between image pixels, allowing them to learn rich spatial representations [72]. Transformer-based models are advanced deep-learning models that are effective in representing images in memes. They can capture complex relationships between visual and textual information. One notable paper in this field is "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks" by [43]. The paper introduces ViLBERT, a Transformer-based model that learns joint representations of images and text. ViLBERT understands the visual context of memes and connects them with relevant textual content, leading to improved meme analysis and classification. These models provide a holistic understanding of memes by integrating both visual and textual information.

1.3.2.7 Autoencoders

Autoencoders are neural networks that are trained to reconstruct the input images. The hidden layers of the autoencoder can be used to capture a compressed represen-

tation of the image data [73]. Autoencoders are deep-learning models used for image representation in memes. They are neural networks designed to learn compact and meaningful representations of input data. In the context of memes, autoencoders aim to capture important features and patterns in the images, allowing for effective analysis and classification. The basic structure of an autoencoder consists of two main parts: an encoder and a decoder. The encoder takes in the input image and compresses it into a lower-dimensional representation, known as the latent space. The decoder then reconstructs the image from this latent space representation. The key idea behind autoencoders is that the model learns to reconstruct the input image as accurately as possible. By doing so, the encoder learns to capture the most essential information about the image in the compressed representation, while the decoder learns to generate an accurate reconstruction. An example of a reference paper in this area is by [11] they focused on classifying images of construction materials. They used two pre-trained models, GoogleNet and ResNet101, which were trained on a different dataset called ImageNet. The results were promising when they applied transfer learning to these models. For GoogleNet, they achieved a classification accuracy of 95.50% using a fine-tuning scheme. With ResNet101, the best result was 95.00% when using a fixed feature extractor approach. But they didn't stop there. They went further and tried using different methods to improve the results. One such method was called Autoencoder, which performed better than PCA (Principal Component Analysis) in all cases. When they used Autoencoder on the fixed feature extractor of ResNet101, they got an impressive 97.83% accuracy. In conclusion, just by adding Autoencoder on top of the pre-trained features, they could improve the performance significantly without needing to fine-tune the complex pre-trained model.

These image data representation techniques have been widely used in deep learning and have shown significant advancements in various computer vision tasks. The provided reference papers delve into the details of each technique, including their architectures, training methodologies, and experimental results. They serve as valuable resources to understand and explore these techniques further.

1.4 DATA FUSION TECHNIQUES

An imaging modality is a type of information that a sensor can provide for a specific purpose. Depending on the sensor used, we can collect different types of data for recognizing human actions. The accuracy and precision of this data may vary across different modalities. This group includes different types of data such as RGB images, skeletons, depth maps, and infrared images. Many studies on action recognition have explored different ways of combining multiple sources of information to improve how well actions can be recognized. Instead of analyzing each source separately, researchers have looked into methods for merging them. Multimodal fusion approaches can be divided based on the way they learn. One category includes approaches where the representation of the action is manually designed by choosing specific features that stand out. Newer methods for combining multiple sources of information in action classification using deep learning techniques. The approaches can be divided into two main types: early fusion and late fusion, based on where the

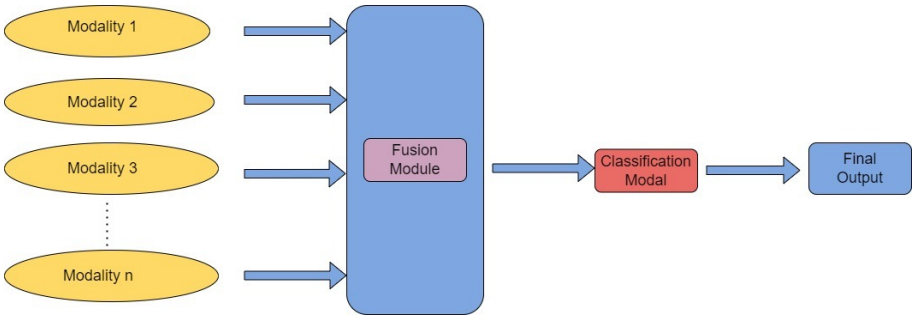

Figure 1.4 Early Fusion

fusion occurs in the processing chain. There are also hybrid fusion approaches that try to combine the characteristics of both methods. Multimodal fusion is a way of combining information from different sources, like images and text, to make predictions. It can be done in different ways, such as combining the information early on, combining it at the end, or using a mix of both. The goal is to use this combined information to predict things like whether someone is happy or sad, or how positive or negative their feelings are [17, 76].

1.4.1 EARLY FUSION

In early fusion, we can combine raw data or initial features. However, since raw data from different sources may have different resolutions or sampling rates, they need to be pre-processed before being combined for further processing. If we move forward in the network and start fusing at the level of initial features, we still need to ensure that the data is aligned in terms of space and time. There are different ways to extract these features, such as using convolutional units and training them from scratch or pre-training them on a similar dataset. Another option is to initialize the convolutional units with weights that have been learned from a similar dataset. Feature level fusion or early fusion is a useful method because it allows us to combine different types of information early on. This helps us do tasks better. Another benefit is that we only need to learn once using the combined information [68]. But it's difficult to show how the different features are synchronized in terms of time using this method. This happens because when we combine information from different sources that are connected, they may be collected at different times. Also, before combining the information, it needs to be in the same format [5].

1.4.2 INTERMEDIATE FUSION

Intermediate fusion is a way of combining different types of data within the recognition model itself. It merges the distinctive features from each type of data to create a new representation that is more powerful than individual representations. For instance, by combining features from RGB images and skeletal sequences, we can

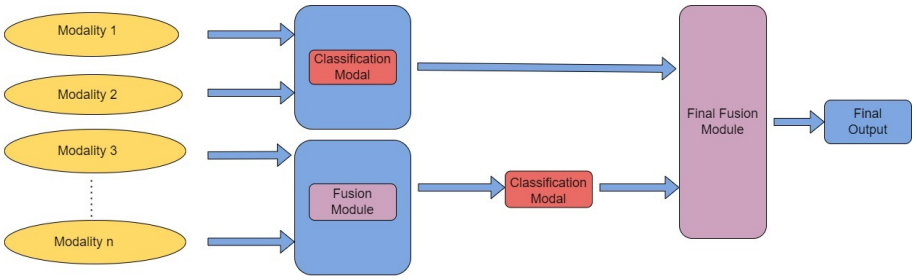

Figure 1.5 Intermediate Fusion

benefit from the strengths of both representations at the same time. This can lead to better recognition results compared to using just one type of representation alone. A hybrid fusion approach means combining the good parts of early and late fusion strategies. Lots of scientists use this method to solve different multimedia problems [9, 50, 78].

1.4.3 LATE FUSION

Late fusion is a merging technique that takes place outside the individual classification models. It takes the decisions made by each classifier and merges them to create new decisions that are more accurate and trustworthy. Unlike early fusion, late fusion does not involve deep neural networks for score fusion. Instead, existing methods retrieve scores from the individual classifiers' softmax layers and then use manually designed rules to combine these scores. The decision level fusion strategy or late fusion is better than feature fusion for a few reasons. With feature fusion, the different things we're combining, like sound and video, might be different. But with decision fusion, the things we're combining are usually the same. So it's easier to put them together. Decision fusion also lets us add or remove things we're combining more easily than feature fusion [6]. Late fusion strategy has another benefit. It lets us use the best methods to analyze different things, like using the hidden Markov model (HMM) for sound and support vector machine (SVM) for pictures. This gives us more flexibility compared to early fusion.

1.5 APPLICATIONS OF MEME CLASSIFICATION

The applications of meme classification refer to the various ways in which the categorization and analysis of memes can be utilized. Memes are widely shared and often humorous images, text that spread rapidly across the internet, typically through social media platforms. By applying classification techniques to memes, several practical applications can be achieved.

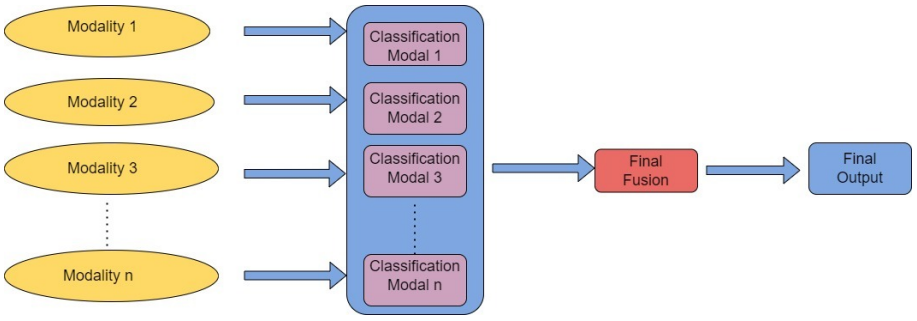

Figure 1.6 Late Fusion

1.5.1 SENTIMENT ANALYSIS

One of the most important applications of NLP (Natural Language Processing) is opinion mining, also referred to as sentiment analysis. In recent years, there has been a great deal of focus placed on sentiment analysis. The purpose of this article is to address one of the most fundamental challenges that comes up in the field of sentiment analysis, and that is the categorization of the different degrees of positive and negative feelings. A general method for classifying the degree of emotional polarity is suggested, along with descriptions of the process in further depth. Feelings can inspire a variety of mental states, including attitudes, thoughts, and judgements. Opinion mining and sentiment analysis are both terms that refer to the same thing: the gathering of information about how people feel about particular things. When it comes to obtaining information on people's feelings, the Internet is a rich source. People who utilize various forms of social media, such as forums, microblogs, or online social networking sites, have the ability to publish their own content. This is possible from the point of view of the user. Many social media sites publish their application programming interfaces (APIs), which encourages data collecting and analysis by researchers as well as developers [19].

1.5.2 HUMOR

Research on humor draws from a wide variety of academic fields. People in a wide variety of scientific domains, including psychology, linguistics, sociology, and literature, have been conducting studies on humor. Especially when it comes to the field of computer science (also known as artificial intelligence), the goal of humor study is to model humor in a form that can be easily processed by computers. When designing user interfaces, having computational models of humor gives interface designers the ability to make the computer generate and understand humor when it is engaging with users. When two people engage with one another, there are many different scenarios in which humor can play a vital part in maintaining the flow of the discussion [47]. In spite of the fact that Charles Gruner reworked this theory in the 21st century as the Superiority Theory of Humour [GRU97], it still has the appearance of

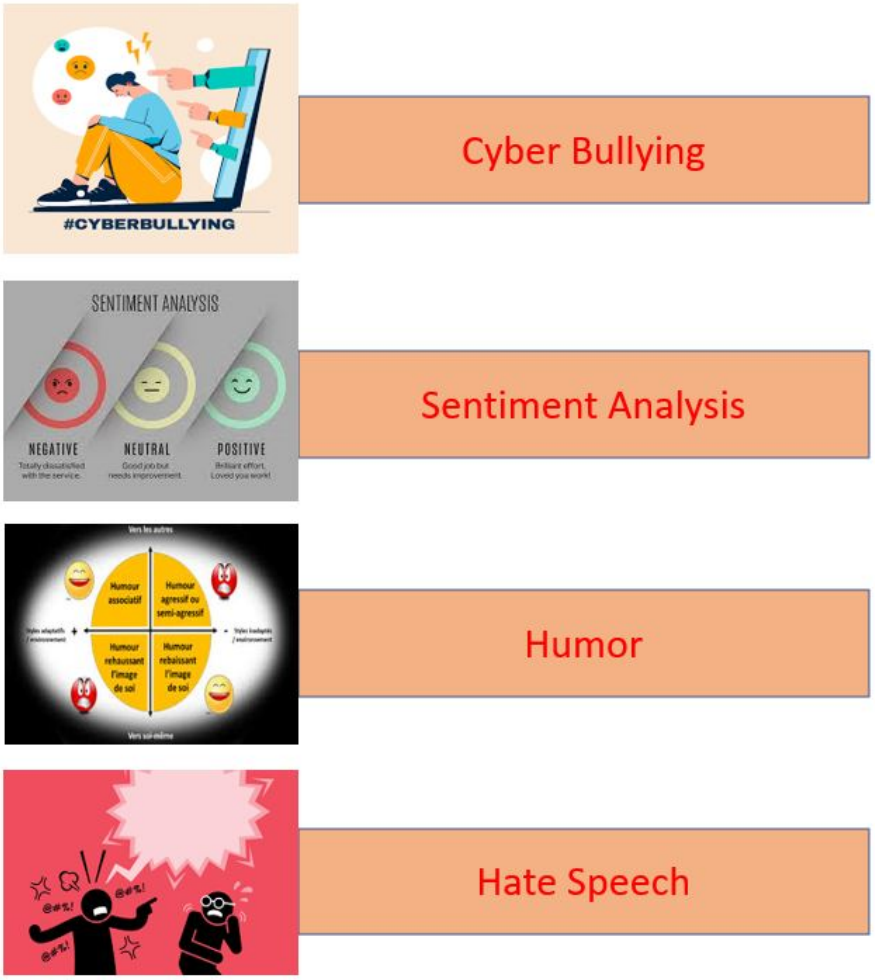

Figure 1.7 Applications of Meme Classification

being archaic. His hypothesis is built around a three-pronged thesis: In every funny scenario, there is at least one person who comes out on top and at least one person who comes out on bottom. In every humorous scenario, incongruity is present. Humor demands the presence of an element of surprise. This idea of superiority will be discussed in the first section of this thesis. Human nature provides the foundation for the presumption that there must always be a victor and a vanquished in comedic exchanges. Humans throughout history have used humor as a strategy to ”compete” with other people by making those people the target of their amusing comments. The ”winner” is the person who is able to poke fun at the ”loser” with greater ease.

1.5.3 CYBERBULLYING

Cyberbullying is an emerging societal issue in the digital era . The Cyberbullying Research Centre conducted a nationwide survey of 5700 adolescents in the US and found that 33.8% of the respondents had been cyberbullied and 11.5% had cyber bullied others. While cyberbullying occurs in different online channels and platforms, social networking sites are fertile grounds for online bullying. A recent large-scale survey conducted by Ditch the Label, an anti-bullying charity, showed the prevalence of cyberbullying on social networking sites 46% of the respondents reported being bullied more than once, and 20% reported bullying others [12]. These features alter the bullying dynamic and distinguish SNS bullying from traditional face-to-face (F2F) bullying and possibly other cyberbullying types on alternative digital communication media (DCM) such as e-mail, telephone, and text messaging service. Indeed, Lowry and colleagues emphasized, “most of these cyberbullying studies have glossed over the central issue: the role of ... social media artifacts themselves in promoting cyberbullying, The writers discuss the significance of schools, families, and communities cooperating with one another in an effort to put an end to cyberbullying. They recommend that prevention programmes and interventions be based on evidence and include participation from everyone. They also stress how important it is to educate people on how to utilize technology in a responsible and polite manner.

1.5.4 HATE SPEECH

An overview of the research conducted and methods utilized in relation to the automatic detection of hate speech in text can be found in the article by [21].The authors begin by providing an explanation of the meaning of hate speech as well as its characteristics. They define hate speech as language that encourages violence, prejudice, or animosity towards individuals or groups based on traits such as race, religion, gender, or ethnicity. Hate speech can also be referred to as bigoted speech. They emphasize the significance of combating hate speech due to the abundance of internet venues where it may be found and the potential damage it can do. After that, the article examines a variety of methods that are utilized for the automatic detection of hate speech. It includes conventional approaches to machine learning in addition to more contemporary developments in deep learning. Utilizing features derived from the text, such as n-grams or linguistic patterns, and combining them with machine

learning algorithms is how traditional methods are carried out. Deep learning, on the other hand, is a type of machine learning that uses neural networks to automatically discover patterns and representations from text input.

1.6 AFFECTIVE MULTIMODAL MEME ANALYSIS DATSETS

The development of memes categorization depends critically on the availability of high-quality datasets. Datasets give scientists the tools they need to create and test machine learning models and algorithms. Researchers can create more thorough and generalizable classification systems by carefully selecting datasets that include a variety of meme categories and cultural origins. An overview of major datasets utilized in meme classification research is provided in this section. Based on a variety of criteria, each dataset is examined. We describe the benefits and drawbacks of each dataset, outlining how well they fit the needs of various research goals.

Table 1.2
Different Datsets for Affective Meme Analysis

Dataset Name	No. of samples	No. of offensive memes	No. of non-offensive memes	References
Facebook Memes Challenge Dataset	10000	3,266	6734	[33]
MultiOFF dataset	743	305	438	[69]
Hateful memes dataset	10,000	3, 253	6747	[44]
HarMeme dataset	3,544	1,249	2295	[56]
Benchmark dataset	800	400	400	[23]
Task Dataset	Data	5, 504	5496	[20]

Facebook Meme Challenge Datset [33] :- This dataset, introduced by Kiela et al. in 2020. The dataset focuses specifically on memes containing hate speech, providing valuable annotations for identifying hateful instances. It includes both images and associated text captions, allowing for multimodal analysis. The dataset can contribute to developing models for hate speech detection in memes, addressing a significant concern in online content moderation. One potential disadvantage is that the dataset may have limitations in terms of size and diversity. It is essential to ensure that the dataset represents a wide range of hate speech instances to create robust and generalizable models. Additionally, the annotation process for identifying hateful instances can be subjective and may introduce biases, affecting the reliability of the dataset.
MultiOFF Dataset [69]:- Suryawanshi et al. introduced this dataset in 2020.This dataset focuses on political memes sourced from Reddit, offering insights into sentiment and political alignment. It can help in studying the role of memes in political discourse and understanding public opinion. The dataset's connection to a specific platform provides context for the memes and facilitates analysis of their impact on political discussions. A potential limitation is that the dataset is limited to political memes from Reddit, which may not represent the full spectrum of political discourse

on other platforms or in different contexts. The dataset's annotations for sentiment and political alignment can be subjective, and differing interpretations may affect the dataset's reliability and generalizability.

Hateful Meme Dataset [44]:- This dataset comprises multimodal memes extracted from social media platforms, providing a rich resource for studying both the sentiment and emotion expressed in memes. The inclusion of both text and visual features allows for a comprehensive analysis of the memes' communicative power. The dataset's multimodal nature reflects the complexity of memes and supports the development of models that consider both modalities. One potential disadvantage is that the dataset's multimodal nature introduces additional challenges in data representation and analysis. Combining text and image features can be computationally intensive, requiring sophisticated techniques for effective integration. Additionally, ensuring consistent and reliable annotations for both sentiment and emotion across modalities can be challenging and may impact the dataset's quality.

HarMeme dataset [56]:- The dataset focuses specifically on memes related to the COVID-19 pandemic, enabling researchers to study the unique characteristics and themes emerging during this global crisis. The dataset's identification of hateful or offensive memes related to COVID-19 can contribute to understanding the spread of misinformation, hate speech, or discriminatory content during public health emergencies. The dataset focuses specifically on memes related to the COVID-19 pandemic, enabling researchers to study the unique characteristics and themes emerging during this global crisis. The dataset's identification of hateful or offensive memes related to COVID-19 can contribute to understanding the spread of misinformation, hate speech, or discriminatory content during public health emergencies.

Benchmark Dataset on Migony Memes [23] :- This dataset focuses on misogyny in memes, shedding light on the discrimination or mistreatment of women in online spaces. It provides annotations for memes specifically related to misogyny, allowing for targeted analysis and the development of models to detect and combat sexist content. The dataset's thematic focus contributes to addressing gender-based harassment and promoting safer online environments. One potential limitation is that the dataset's narrow focus on misogyny may limit its generalizability to other types of hate speech or offensive content. The annotation process for identifying misogynistic memes can be challenging due to the subjectivity of judgments, potentially introducing inconsistencies or biases in the dataset.

Task Datset [20] :- This dataset specifically targets misogynistic memes, offering a large collection annotated for the presence of misogyny. It provides valuable resources for studying and combating gender-based discrimination in online spaces.

1.7 FUTURE SCOPE

Multimodal meme classification a branch of affective computimg for human emotion analysis has witnessed significant advancements through the application of deep learning techniques. However, there are several promising avenues for future research that can further enhance the accuracy and robustness of meme classification systems. One such direction is the exploration of attention mechanisms within

deep learning models. Attention mechanisms have shown great potential in capturing salient features and context in memes, enabling models to focus on relevant visual and textual elements. Future studies can investigate novel attention mechanisms tailored specifically for meme classification tasks, exploring both self-attention and cross-modal attention strategies. In addition to attention-based models, non-attention models also hold promise for meme classification. These models can provide alternative approaches that do not rely on explicit attention mechanisms, yet achieve competitive performance. Future research can explore and compare the effectiveness of various non-attention models, such as graph-based neural networks, capsule networks, or memory-augmented models, in meme classification tasks. These models can capture global dependencies and long-range contextual information, contributing to improved meme classification accuracy.

Furthermore, the representation of multimodal data, combining image and text, is a crucial aspect of meme classification therefore future studies can delve into advanced data representation techniques that effectively fuse information from both modalities.This includes investigating techniques such as graph-based fusion, cross-modal hashing, or multimodal embedding models. And even can include various other modalities such as audio, video etc,. By leveraging these approaches, researchers can extract more informative and discriminative features from memes, leading to enhanced classification performance. Another important aspect for future research is the availability and creation of diverse and large-scale meme datasets. While existing datasets, such as the [33], have contributed to the advancement of meme classification, there is a need for more comprehensive and domain-specific datasets. Also, future studies can focus on curating datasets that encompass a wide range of themes, sentiments, and cultural contexts. Moreover, the development of benchmark datasets with standardized evaluation protocols can facilitate fair comparisons between different meme classification approaches and drive further research progress. By addressing these areas, researchers can contribute to the development of more accurate and robust meme classification systems, enabling better understanding, moderation, and analysis of memes in various social media contexts. As timely recognition of offensive memes can help in reduction of spreading hatred among people and even reduces the activities like suicides caused due to them.

1.8 CONCLUSION

In conclusion, this chapter highlights the growing importance of identifying hate speech and analyzing the sentiment expressed in memes. With the widespread use of memes in the digital world, it has become crucial to develop techniques that can effectively classify and understand their content. Memes, with their combination of visual and textual elements, pose unique challenges for sentiment analysis. The multimodal nature of memes necessitates the consideration of both text and image components in classification models to accurately categorize them. Although there is a scarcity of studies and datasets specifically dedicated to hate speech detection and sentiment analysis in memes, the research in this field has shown promising advancements. Despite the existing gaps in the literature, recent approaches have

demonstrated improved results in addressing the challenges associated with meme analysis. The significance of identifying hate speech and examining the emotional aspects of memes is emphasized in this paper. Through a comprehensive analysis of prior research, the review evaluates the strategies proposed, including their strengths and limitations. It serves as a valuable resource for researchers in understanding the current state-of-the-art techniques and identifying future directions for study. In conclusion, this chapter contributes to the enhancement of techniques for identifying hate speech and analyzing meme sentiment. By addressing the challenges posed by memes' multimodal nature and synthesizing the existing research, it lays the foundation for the development of more effective methods in meme classification and contributes to creating a safer and more inclusive digital environment.

1. Apeksha Aggarwal, Vibhav Sharma, Anshul Trivedi, Mayank Yadav, Chirag Agrawal, Dilbag Singh, Vipul Mishra, and Hassène Gritli. Two-way feature extraction using sequential and multimodal approach for hateful meme classification. *Complexity*, 2021:1–7, 2021.

2. Nayan Varma Alluri and Neeli Dheeraj Krishna. Multi modal analysis of memes for sentiment extraction. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 213–217. IEEE, 2021.

3. Aaliyah Alshehri, Yakoub Bazi, Nassim Ammour, Haidar Almubarak, and Naif Alajlan. Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery. *IEEE Access*, 7:119873–119880, 2019.

4. Segun Taofeek Aroyehun and Alexander Gelbukh. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, 2018.

5. Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379, 2010.

6. Pradeep K Atrey, Mohan S Kankanhalli, and John B Oommen. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):2–es, 2007.

7. Jafar Badour and Joseph Alexander Brown. Hateful memes classification using machine learning. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2021.

8. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

9. Azzedine Bendjebbour, Yves Delignon, Laurent Fouque, Vincent Samson, and Wojciech Pieczynski. Multisensor image segmentation using dempster-shafer fusion in markov fields context. *IEEE Transactions on Geoscience and Remote Sensing*, 39(8):1789–1798, 2001.

10. Aruna Bhat, Vaibhav Varshney, Varun Bajlotra, and Vishesh Gupta. Detection of hatefulness in memes using unimodal and multimodal techniques. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 65–73. IEEE, 2022.

11. S Bunrit, N Kerdprasop, and K Kerdprasop. Improving the representation of cnn based features by autoencoder for a task of construction material image classification. *Journal of Advances in Information Technology*, 11(4), 2020.

12. Tommy KH Chan, Christy MK Cheung, and Zach WY Lee. Cyberbullying on social networking sites: A literature review and future research directions. *Information & Management*, 58(2):103411, 2021.

13. Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230, 2017.

14. Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

15. Abhishek Das, Japsimar Singh Wahi, and Siyao Li. Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2012.14891*, 2020.

16. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

17. Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

18. Christopher Ifeanyi Eke, Azah Norman, Liyana Shuib, Faith B Fatokun, and Isaiah Omame. The significance of global vectors representation in sarcasm analysis. In *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*, pages 1–7. IEEE, 2020.

19. Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):1–14, 2015.

20. Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, 2022.

21. Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

22. Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291–4308, 2020.

23. Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526, 2022.

24. Roushan Kumar Giri, Subhash Chandra Gupta, and Umesh Kumar Gupta. An approach to detect offence in memes using natural language processing (nlp) and deep learning. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2021.

25. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

26. Denis Gordeev and Vsevolod Potapov. Automatic meme generation with an autoregressive transformer. In *International Conference on Human-Computer Interaction*, pages 309–317. Springer, 2022.

27. Yimeng Gu, Ignacio Castro, and Gareth Tyson. Mmvae at semeval-2022 task 5: A multi-modal multi-task vae on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 700–710, 2022.

28. Xiaoyu Guo, Jing Ma, and Arkaitz Zubiaga. Nuaa-qmul at semeval-2020 task 8: Utilizing bert and densenet for internet meme emotion analysis. *arXiv preprint arXiv:2011.02788*, 2020.

29. Hritvik Gupta and Mayank Patel. Study of extractive text summarizer using the elmo embedding. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 829–834. IEEE, 2020.

30. Md Tarek Hasan, Md Al Emran Hossain, Md Saddam Hossain Mukta, Arifa Akter, Mohiuddin Ahmed, and Salekul Islam. A review on deep-learning-based cyberbullying detection. *Future Internet*, 15(5):179, 2023.

31. Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

32. Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. Iitk at semeval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes. *arXiv preprint arXiv:2007.10822*, 2020.

33. Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

34. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

35. Akshi Kumar and Nitin Sachdeva. Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. *Multimedia Systems*, pages 1–10, 2021.

36. Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, pages 1–36, 2023.

37. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

38. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

39. Gichang Lee, Jaeyun Jeong, Seungwan Seo, CzangYeob Kim, and Pilsung Kang. Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge-Based Systems*, 152:70–82, 2018.

40. Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147, 2021.

41. Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

42. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

43. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

44. Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. Findings of the woah 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, 2021.

45. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

46. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

47. Matthijs P Mulder and Antinus Nijholt. *Humour research: State of the art*. Centre for Telematics and Information Technology, University of Twente, 2002.

48. Tsendsuren Munkhdalai and Hong Yu. Neural tree indexers for text understanding. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 11. NIH Public Access, 2017.

49. Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2(1):1–21, 2015.

50. Jianjun Ni, Xiaoping Ma, Lizhong Xu, and Jianying Wang. An image recognition method based on multiple bp neural networks fusion. In *International Conference on Information Acquisition, 2004. Proceedings.*, pages 323–326. IEEE, 2004.

51. Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

52. Sofiane Ouaari, Tsegaye Misikir Tashu, and Tomás Horváth. Multimodal feature extraction for memes sentiment classification. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 285–290. IEEE, 2022.

53. John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, pages 571–576, 2019.

54. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

55. ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. arxiv preprint 2018. *arXiv preprint arXiv:1802.05365*, 1802.

56. Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*, 2021.

57. Raj Ratn Pranesh and Ambesh Shekhar. Memesem: a multi-modal framework for sentimental analysis of meme via transfer learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020.

58. Ailneni Rakshitha Rao and Arjun Rao. Asrtrans at semeval-2022 task 5: Transformer-based models for meme classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 597–604, 2022.

59. Anthony Rios and Ramakanth Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM conference on bioinformatics, computational biology and health informatics*, pages 258–267, 2015.

60. Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204951–204962, 2020.

61. Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*, 2019.

62. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

63. Mayukh Sharma, Ilanthenral Kandasamy, and Wb Vasantha. Memebusters at semeval-2020 task 8: feature fusion model for sentiment analysis on memes using transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1163–1171, 2020.

64. Ranti D Sharma, Samarth Tripathi, Sunil K Sahu, Sudhanshu Mittal, and Ashish Anand. Predicting online doctor ratings from user reviews using convolutional neural networks. *International Journal of Machine Learning and Computing*, 6(2):149, 2016.

65. Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065, 2019.

66. Manohar Gowdru Shridara, Daniel Hládek, Matúš Pleva, and Renát Haluška. Identification of trolling in memes using convolutional neural networks. In *2023 33rd International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 1–6. IEEE, 2023.

67. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

68. Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

69. Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41, 2020.

70. Radiathun Tasnia, Nabila Ayman, Afrin Sultana, Abu Nowshed Chy, and Masaki Aono. Exploiting stacked embeddings with lstm for multilingual humor and irony detection. *Social Network Analysis and Mining*, 13(1):43, 2023.

71. Teoh Hwai Teng and Kasturi Dewi Varathan. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, 2023.

72. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ■ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

73. Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

74. George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. Upb@ dankmemes: Italian memes analysis-employing visual models and graph convolutional networks for meme identification and hate speech detection. *EVALITA Evaluation of NLP and Speech Tools for Italian*, page 288, 2020.

75. Suryatej Reddy Vyalla and Vishaal Udandarao. Memeify: A large-scale meme generation system. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 307–311. 2020.

76. Martin Wollmer, Moritz Kaiser, Florian Eyben, Bjorn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.

77. Kecong Xiao, Zishuai Zhang, and Jun Wu. Chinese text sentiment analysis based on improved convolutional neural networks. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 922–926. IEEE, 2016.

78. Huaxin Xu and Tat-Seng Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):44–67, 2006.

79. Kejie Xu, Yue Zhao, Lingming Zhang, Chenqiang Gao, and Hong Huang. Spectral–spatial residual graph attention network for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

80. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

81. Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

82. Yichun Yin, Yangqiu Song, and Ming Zhang. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2044–2054, 2017.

83. Yue Zhang, Qi Liu, and Linfeng Song. Sentence-state lstm for text representation. *arXiv preprint arXiv:1805.02474*, 2018.

84. Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2020.

85. Jiawen Zhu, Roy Ka-Wei Lee, and Wen Haw Chong. Multimodal zero-shot hateful meme detection. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 382–389, 2022.