MULTIMODAL SENTIMENT ANALYSIS OF SOCIAL MEDIA THROUGH DEEP LEARNING APPROACH

AN JIEYU

UNIVERSITI SAINS MALAYSIA

2024

MULTIMODAL SENTIMENT ANALYSIS OF SOCIAL MEDIA THROUGH DEEP LEARNING APPROACH

by

AN JIEYU

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

June 2024

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to Dr. Wan Mohd Nazmee Wan Zainon for his invaluable support, guidance, patience, motivation, enthusiasm, and immense knowledge throughout the course of this research. His exceptional supervision played a pivotal role in shaping the direction and outcomes of this thesis. I also extend my heartfelt thanks to Associate Professor Gan Keng Hoon, Associate Professor Ahmad Sufril Azlan Mohamed, and Professor Ali Selamat for their insightful feedback and expertise, which greatly contributed to the refinement of my research and the quality of this thesis.

Furthermore, I would like to extend my heartfelt appreciation and recognition to my family, whose unwavering love and support have been indispensable throughout my academic journey. I am particularly grateful to my wife and son who have been my emotional backbone, not only accompanying me in my studies in Penang, but also giving me constant encouragement and understanding throughout my life. Their belief in my abilities has been a source of inspiration, empowering me to overcome challenges and pursue my research endeavours.

In addition, I am indebted to the members of the School of Computer Sciences at USM and all the individuals associated with the lab. Their invaluable assistance, collaboration, and friendship have fostered an environment conducive to my academic growth and development over the years. The insightful discussions, exchange of ideas, and shared enthusiasm for research have been truly enriching. I am grateful for the opportunity to be a part of such a vibrant and intellectually stimulating community.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANP	Adjective Noun Pairs
BERT	Bidirectional Encoder Representation from Transformer
CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Network
CV	Computer Vision
GPT	Generative Pre-training
GPU	Graphics Processing Unit
ICCI	Integration of Color Cues for enhanced Multimodal Sentiment Analysis
ITMSC	Improving Targeted Multimodal Sentiment Classification
LSTM	Long Short-Term Memory
MSA	Multimodal Sentiment Analysis
MLP	Multilayer Perceptron
NB	Naïve Bayes
NLP	Natural Language Processing
PTM	Pre-trained Model
ResNet	Residual Networks
RF	Random Forest
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SVM	Support Vector Machine
TMSC	Targeted multimodal sentiment classification
VLMSA	Vision-language Multimodal Sentiment Analysis
VSO	Visual Sentiment Ontology

ANALISIS SENTIMEN BERBILANG MOD MEDIA SOSIAL MELALUI PENDEKATAN PEMBELAJARAN MENDALAM

ABSTRAK

Data berbilang mod, dicirikan oleh kerumitan dan kepelbagaian yang wujud, memberikan cabaran pengkomputeran dalam memahami kandungan media sosial. Pendekatan konvensional kepada analisis sentimen sering bergantung pada model pra-latihan unimodal untuk pengekstrakan ciri daripada setiap modality dan ini telah mengabaikan hubungan intrinsik maklumat semantik antara modaliti kerana ia biasanya dilatih pada data unimodal. Selain itu, kaedah analisis sentimen berbilang mod sedia ada tertumpu terutamanya pada perolehan perwakilan imej tetapi mengabaikan maklumat semantik yang kaya terkandung dalam imej. Tambahan pula, kaedah semasa sering mengabaikan kepentingan maklumat warna, yang memberikan pandangan berharga dan ianya mempengaruhi klasifikasi sentimen dengan ketara. Menangani jurang ini, tesis ini meneroka kaedah berasaskan pembelajaran mendalam untuk analisis sentimen berbilang mod dengan menekankan perkaitan semantik antara data berbilang mod, interaksi maklumat dan pemodelan sentimen warna dari perspektif lapisan perwakilan berbilang mod, lapisan interaksi berbilang mod dan maklumat integrasi warna lapisan. Untuk mengurangkan perkaitan semantik yang diabaikan antara modaliti, tesis ini memperkenalkan "Pembelajaran Perwakilan Bersama untuk Analisis Sentimen Berbilang Mod" dalam lapisan perwakilan. Kaedah ini, disahkan oleh eksperimen yang ketat, mempamerkan peningkatan yang ketara dalam ketepatan, mencapai 76.44% pada MVSA-Single dan 72.29% pada dataset MVSA-Multiple, mengatasi metodologi sedia ada. Dalam lapisan interaksi berbilang mod, "Klasifikasi Sentimen Berbilang Mod Sasaran dengan Perihalan Imej Semantik" dicadangkan

untuk memperdalam pemahaman kandungan imej dan meningkatkan analisis terperinci. Pendekatan ini, setelah disahkan prestasinya, mencapai peningkatan ketepatan, masing-masing mencapai 78.59% dan 70.28% pada set data Twitter-2015 dan Twitter-2017, yang digunakan untuk analisis sentimen pelbagai mod yang terperinci. Kaedah ini telah terbukti sangat berdaya saing berbanding pendekatan sedia ada. Selain itu, pengiktirafan potensi maklumat warna yang kurang digunakan dalam analisis sentimen, "Mengintegrasikan Isyarat Warna untuk Analisis Sentimen Multimodal yang Dipertingkatkan" dengan memperkenalkan lapisan penyepaduan maklumat warna, menunjukkan peningkatan prestasi yang luar biasa dengan ketepatan mencapai 79.33% dan 73.29% pada MVSA-Single dan MVSA-berbilang set data, masing-masing. Tesis ini menyumbang kepada pemahaman semasa analisis sentimen berbilang mod dan membuka jalan untuk penyiasatan lanjut dan pembangunan teknik yang lebih maju dalam bidang ini.

MULTIMODAL SENTIMENT ANALYSIS OF SOCIAL MEDIA THROUGH DEEP LEARNING APPROACH

ABSTRACT

Multimodal data, characterized by its inherent complexity and heterogeneity, presents computational challenges in comprehending social media content. Conventional approaches to sentiment analysis often rely on unimodal pre-trained models for feature extraction from each modality, neglecting the intrinsic connections of semantic information between modalities, as they are typically trained on unimodal data. Additionally, existing multimodal sentiment analysis methods primarily focus on acquiring image representations while disregarding the rich semantic information contained within the images. Furthermore, current methods often overlook the significance of color information, which provides valuable insights and significantly influences sentiment classification. Addressing these gaps, this thesis explores deep learning-based methods for multimodal sentiment analysis, emphasizing the semantic association between multimodal data, information interaction, and color sentiment modelling from the perspectives of the multimodal representation layer, the multimodal interaction layer, and the color information integration layer. To mitigate the overlooked semantic interrelations between modalities, the thesis introduces "Joint Representation Learning for Multimodal Sentiment Analysis" within the representation layer. This method, validated by rigorous experiments, showcases a marked improvement in accuracy, achieving 76.44% on the MVSA-Single and 72.29% on the MVSA-Multiple datasets, surpassing existing methodologies. In the multimodal interaction layer, "Targeted Multimodal Sentiment Classification with Semantic Image Descriptions" is proposed to deepen the comprehension of image content and

enhance fine-grained analysis. This approach, having validated its performance, achieved accuracy gains, reaching 78.59% and 70.28% on the Twitter-2015 and Twitter-2017 datasets, respectively, which are used for fine-grained multimodal sentiment analysis. This method has proven highly competitive with existing approaches. Furthermore, recognizing the underutilized potential of color information in sentiment analysis, "Integrating Color Cues for Enhanced Multimodal Sentiment Analysis" introduces a color information integration layer, showing a remarkable enhancement in performance with accuracy reaching 79.33% and 73.29% on the MVSA-Single and MVSA-Multiple datasets, respectively. This thesis contributes to the current understanding of multimodal sentiment analysis and paves the way for further investigation and development of more advanced techniques in this field.

CHAPTER 1

INTRODUCTION

1.1 Background

Social media has emerged as a significant information source, primarily due to the enormous quantity of daily online messages. Consequently, it has become an extensively adopted online activity. The increasing popularity of social media has profoundly influenced individuals' lives, leading to a surge in the utilization of online social networking platforms such as Twitter, Facebook, Weibo, and Instagram. These platforms serve as avenues for individuals to share their experiences and express their opinions on diverse events and subjects. Moreover, the user-generated content found in social media data presents a promising opportunity to extract valuable insights into human sentiment.

The origins of sentiment analysis can be traced back to the discipline of computational linguistics and natural language processing, which have been the subject of scholarly investigation since the 1950s. However, the specific term "sentiment analysis" emerged in the mid-2000s, around 2004-2005, when researchers began focusing on automated techniques for determining the sentiment of texts, such as product reviews, movie reviews, and social media posts. Although different terms like opinion mining, opinion extraction, emotion analysis, sentiment extraction, sentiment classification, sentiment mining, subjectivity analysis, affect analysis, and review analysis have been used, they all share the same research objectives: detecting and categorizing feelings and attitudes towards specific events or objects (Liu, 2012).

Sentiment analysis plays a critical role in bridging the gap between user-generated data and inferred sentiment, thereby facilitating improved decision-making

in various applications. It involves categorizing input data, which can be extracted from sources such as reviews, comments, or social media posts, into three distinct categories: positive polarity, negative polarity, and neutral polarity. Consequently, sentiment analysis can be viewed as a classification task, wherein the objective is to assign each input data point to one of these categories. The significance of sentiment analysis lies in its ability to capture public opinion on a wide range of subjects, including products, services, marketing strategies, political preferences, and social events. As a result, sentiment analysis has garnered substantial interest from both industry and academia (Yue et al., 2019). To illustrate the growing prominence of sentiment analysis, Figure 1.1 showcases the increasing popularity of this field based on data from Google Trends spanning the period from 2010 to 2023.

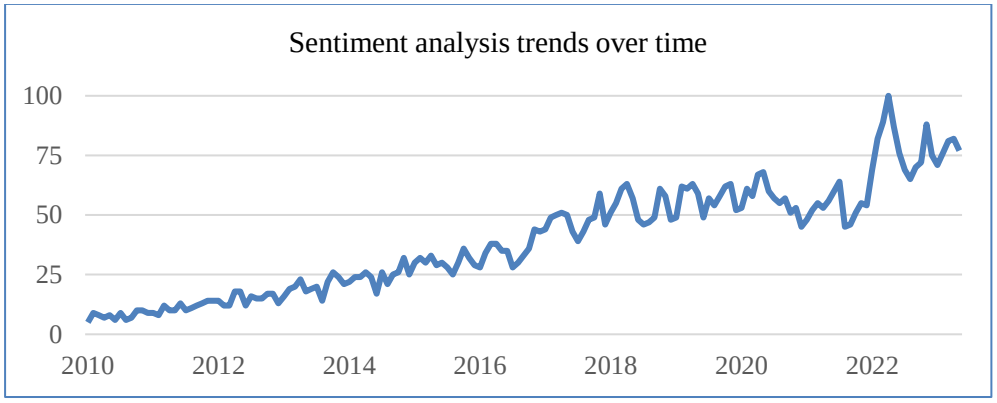

Figure 1.1 Sentiment analysis from Google Trends[1]

The ubiquity of smartphones and the widespread use of social media platforms have made multimodal content, such as images and videos, increasingly influential in capturing and conveying people's sentiments and opinions (Ji et al., 2016; You, 2016). Social media users often complement their textual content with accompanying images

[1] Google Trends. Sentiment analysis search trends. Retrieved December 5th, 2023, from https://trends.google.com/trends

to enhance their expression. Consequently, sentiment analysis for social multimedia has evolved beyond the confines of natural language processing and now encompasses various fields, including computer vision, pattern recognition, and artificial intelligence (Z. Li et al., 2019). Nevertheless, conventional unimodal sentiment analysis methods, which primarily focus on textual data, are limited in their ability to fully exploit the potential of multimodal information. These approaches fail to leverage the value inherent in diverse modal data sources for effectively capturing human emotional intentions. Hence, there arises a pressing need for multimodal sentiment analysis techniques that can harness the richness of multimodal content and extract comprehensive sentiment insights.

For instance, consider the analysis of a social media post that features an image of a serene sunset accompanied by a caption that reads, "Ending the day on a peaceful note." A unimodal sentiment analysis focusing solely on the text might classify this sentiment as positive. However, multimodal sentiment analysis delves deeper by examining both the textual and visual elements, recognizing the tranquillity and contentment conveyed by the image of the sunset in conjunction with the text. This integrated approach enables a more nuanced understanding of the post's sentiment, capturing the serene and reflective mood more accurately than text-based analysis alone. By leveraging both textual and visual data, multimodal sentiment analysis provides a richer, more accurate portrayal of sentiments, offering valuable insights for businesses, social media monitoring, and emotional research.

The primary objective of multimodal sentiment analysis is to comprehend and capture the polarity of sentiment expressed by users by leveraging expressive data from multiple modalities. Consequently, the fusion of information across modalities

not only enhances the accuracy and robustness of sentiment analysis models but also improves the interpretability and explainability of the results. In the past few years, notable advancements have been achieved in the domain of multimodal sentiment analysis, largely attributed to the rise and progress of deep learning technology. Deep learning, a subset of machine learning techniques, has revolutionized various domains within artificial intelligence and has exhibited promising outcomes in the processing of multimodal data (Yadav & Vishwakarma, 2020; P. Huang et al., 2021; Sunitha et al., 2022). The integration of deep learning technology has greatly enhanced multimodal sentiment analysis by enabling the learning of semantic mappings across distinct modalities. Utilizing deep learning models, researchers can effectively map features extracted from diverse modalities to sentiment labels. This approach capitalizes on the strengths of each modality, revealing intricate relationships and patterns within multimodal data.

Multimodal sentiment analysis is a burgeoning research domain that has garnered significant attention from scholars. This approach holds great promise in comprehending human sentiment by integrating diverse modalities, such as textual and visual elements. Deep learning techniques, coupled with advancements in natural language processing and computer vision, empower researchers to explore novel avenues and reveal latent patterns and relationships within multimodal data. With ongoing refinements in these techniques, the potential impact of multimodal sentiment analysis is expected to expand, offering enhanced insights into human sentiments in the digital era and finding applications across diverse domains.

1.2 Research Motivation

In recent years, the proliferation of smartphones has brought about a paradigm shift in communication and self-expression. The integration of high-definition cameras into these devices has made it effortless for individuals to share their thoughts along with captivating visuals, enhancing both the visual appeal and comprehensibility of their posts on popular social media platforms like Twitter and Facebook. Moreover, the practice of enhancing shared images with accompanying text has gained significant traction. These textual annotations serve as a semantic supplement, fostering a deeper understanding and interpretation of the images. Consequently, users can effectively convey their emotions and ideas to friends and followers irrespective of their geographical locations, benefiting from the global reach of social media platforms.

Sentiment analysis of social media has traditionally relied on text-based methods, wherein sentiment is determined by analysing individual words, phrases, and their relationships. While these methods have shown effectiveness, they overlook the wealth of insights that can be derived from multimodal information, covering both visual content and textual semantics (Poria et al., 2018). Focusing solely on textual data may lead to misinterpretations, particularly with short sentences lacking contextual information, ultimately limiting their ability to capture the full spectrum of sentiment expressed in diverse real-world scenarios.

To overcome the aforementioned limitation, there is an increasingly pressing need to investigate methodologies that go beyond text-based analysis and embrace a multimodal approach. By integrating textual data with visual information, valuable insights can be derived to facilitate the recognition of an individual's emotional state. Figure 1.2 displays a collection of tweets sourced from the social media platform

Twitter, exemplifying this concept. Notably, the tweets contain both text and corresponding image. In the left example, positive sentiment is expressed, while the middle example portrays neutral sentiment, and the right example conveys negative sentiment. It is evident that the fusion of textual and visual data produces sentiment information that goes beyond the ability to use either data type alone.

Happy snowman #illustration

A speck appears on the skyline of windy Beijing

Finally got a pic of my car from the accident …

Figure 1.2 Example of text and image combination from Twitter

The motivation for studying multimodal sentiment analysis arises from several key factors: 1) multimodal data surpasses unimodal data in terms of information richness due to the presence of multiple modalities that can mutually complement and enhance one another (Ju et al., 2021; X. Yan et al., 2022). Consequently, sentiment analysis conducted on multimodal data tends to yield more accurate results compared to its unimodal counterpart; 2) the widespread adoption of social media platforms has led to an exponential growth in the production and consumption of multimedia content (Gandhi et al., 2023). The ability to effectively analyse sentiment within these multimodal posts holds substantial value in gaining insights into public opinion, brand perception, and user experiences; 3) the practical applications of multimodal sentiment

analysis span across various domains. Monitoring social media platforms for public opinion on political issues, brand perception, or customer feedback is a crucial application (Birjali et al., 2021). By analysing multimodal content, decision-makers can gain a comprehensive understanding of how sentiment evolves over time and across different modalities; 4) multimodal sentiment analysis significantly contributes to research in social sciences and psychology (Pandey & Vishwakarma, 2023). It provides deeper insights into human behaviour, emotion detection, and communication patterns. By examining multimodal data, researchers can explore the interplay between different modalities and their impact on sentiment expression and interpretation.

Despite the richness of information inherent in multimodal data, a significant ongoing challenge persists in endowing computers with cognitive abilities comparable to those of humans (Mangaroska et al., 2021). This challenge encompasses the efficient extraction of meaningful information from diverse modalities, as well as the subsequent integration of emotional cues derived from these modalities. By overcoming obstacles such as accurately interpreting complex emotions from diverse data sources and integrating various types of information, researchers can significantly enhance our understanding of human emotions. This, in turn, lays the foundation for ground-breaking applications in numerous domains, including mental health assessment, customer service optimization, and enriched social media analytics. Advancing in this field involves not only technological innovation, employing advanced AI and machine learning algorithms, but also a deep interdisciplinary understanding of psychology and communication.

1.3 Problem Statement

The proliferation of multimedia content, such as text accompanied by images, on social media platforms has witnessed a substantial increase in recent times. Sentiment analysis, a crucial task in understanding user opinions or emotions, has traditionally concentrated on analysing either textual or visual data in isolation. However, this unimodal approach fails to capture the complete richness and complexity of user expressions, thereby resulting in suboptimal accuracy when performing multimodal sentiment analysis (Baltrušaitis et al., 2019). Consequently, the existing problem revolves around the need to develop an effective framework for multimodal sentiment analysis that leverages the complementary information present in both textual and visual modalities. This research endeavour aims to address the prevailing challenge by formulating an efficient framework for multimodal sentiment analysis that capitalizes on the inherent synergies between textual and visual modalities, thereby fostering a more comprehensive understanding of sentiment expression.

Despite the potential benefits of multimodal sentiment analysis, several challenges hinder its successful implementation:

Firstly, multimodal sentiment analysis lies in the challenge associated with effectively capturing and integrating features from various modalities to enhance the understanding of conveyed sentiment. To address this challenge, acquiring features using pre-trained models is a promising avenue. Pre-trained models, such as BERT in natural language processing and vision-based models like VGG or ResNet, have gained prominence for their exceptional ability to learn and represent intricate patterns and semantics within each modality (Huang et al., 2019). However, these pre-trained

models are often limited in their ability to effectively process multimodal data, as they are typically trained on unimodal data. Recent advancements in vision-language pre-trained models, which undergo pre-training on extensive image-text datasets, have mastered the ability to acquire universal cross-modal representations (Radford et al., 2021). These representations are advantageous for enhancing performance in a variety of downstream tasks that intersect vision and language domains. Such tasks include image text retrieval, visual question answering, and vision and language generation, among others (Du, Liu, Li, et al., 2022). Furthermore, beyond vision-language pre-trained models, the adoption of contrastive learning has garnered notable attention in the domains of computer vision and natural language processing, particularly for representation learning purposes. Despite the demonstrated effectiveness of contrastive learning, its incorporation into the domain of multimodal sentiment analysis remains underexplored within the academic community (Lin et al., 2022). Therefore, recognizing these hurdles, this thesis focuses on the effective utilization of vision-language pre-trained models while harnessing the capabilities of contrastive learning methods to enhance representation learning.

Secondly, in contemporary multimodal sentiment analysis techniques, the predominant emphasis lies on the fusion of textual and visual features as a means of capturing emotional expressions. Nonetheless, a frequently overlooked potential avenue involves the utilization of image descriptions, which are derived from images themselves (Cheema et al., 2021). These descriptions harbor a wealth of textual semantic information that can be harnessed for sentiment analysis. The necessity of feature extraction, which is alternatively referred to as feature engineering, is a pivotal stage in the sentiment analysis procedure, as it impacts the outcomes of sentiment classification in a direct manner (Avinash & Sivasankar, 2019). However, the visual

features themselves do not contain textual semantic information and can only capture the visual semantic information of an image. In contrast, the information conveyed through image descriptions encompasses a range of semantic elements, such as depicted objects, scenes, and backgrounds. These elements frequently possess shared semantic attributes with textual modalities, thereby facilitating improved alignment across different modalities (K. Li et al., 2022). The utilization of this shared semantic information assumes a critical role in the integration and synchronization of visual and textual data, thereby facilitating a deeper exploration of the sentiment conveyed within multimedia content(X. Xu et al., 2020). Consequently, there arises a compelling demand for the development of innovative methodologies that proficiently harness the semantic information derived from image descriptions, ultimately enhancing both the accuracy and comprehensiveness of multimodal sentiment analysis.

Thirdly, it is noteworthy that the existing body of literature on multimodal sentiment analysis has tended to overlook the vital role of color information in the conveyance of emotions within images. As illustrated in Figure 1.3, the left image, presented in grayscale, emphasizes a negative atmosphere. In contrast, the right image is noticeably more adept at evoking positive emotions, compared to the relatively less vibrant left image. When integrated with textual content, the right image demonstrates a superior ability to convey emotional messages compared to its counterpart on the left. This illustrative example underscores the crucial significance of holistically considering color features in the development of more precise and resilient models for multimodal sentiment analysis. Despite evidence from psychological research and art theory highlighting the importance of color features in understanding the emotional content of images (Bekhtereva & Müller, 2017; C.-E. Yu et al., 2020), current approaches have not adequately incorporate color information within the framework

of multimodal sentiment analysis. Hence, it is imperative to devise multimodal approaches for sentiment analysis that can efficiently incorporate both semantic and color data, thereby enhancing the accuracy and comprehensiveness of sentiment analysis.

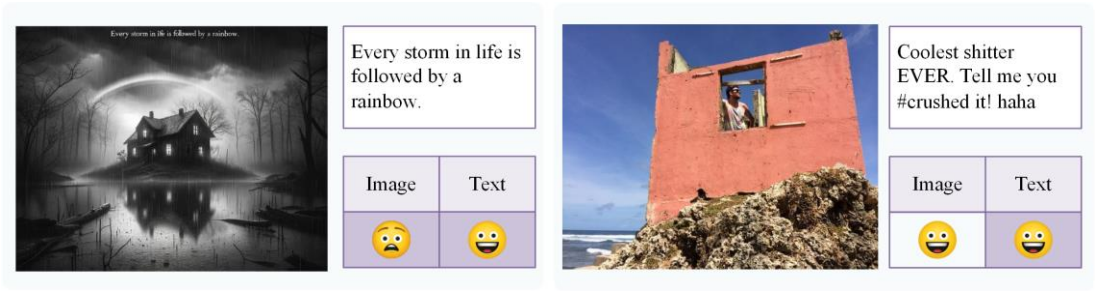

Figure 1.3 Multimodal sentiment analysis with color information

1.4 Research Questions

This thesis addresses the problem of multimodal sentiment analysis, with a specific emphasis on the integration of image and text data derived from social media platforms. The importance of this research becomes evident when considering the comprehensive exploration of the following three fundamental questions:

a) How can the integration of vision-language pre-trained models and contrastive learning techniques leverage the correlations and rich information across modalities for enhanced sentiment analysis?

b) How can semantic information extracted from image descriptions be effectively interacted and fused with other modal data for multimodal sentiment analysis?

c) How can color information be effectively integrated with semantic features to bridge the semantic gap and accurately capture the emotional impact of colors in multimodal sentiment analysis?

These fundamental questions are designed to confront the complex challenges inherent in the fusion of semantic information, the integration of vision-language pre-trained models, the application of contrastive learning techniques, and the efficacious incorporation of color information within the domain of multimodal sentiment analysis. By addressing these questions comprehensively, this study aims to significantly advance the development of sophisticated methodologies and approaches within the complex field of multimodal sentiment analysis.

1.5 Research Objectives

This thesis investigates the challenges of multimodal sentiment analysis with the goal of developing more effective analysis methods by combining visual and textual data. The research is structured around three key objectives, each aimed at addressing specific aspects of multimodal sentiment analysis:

a) To develop a joint representation learning method by integrating a vision-language pre-trained model with a multimodal contrastive learning method. This framework aims to effectively capture and represent sentiment semantic information in multimodal data, creating a strong and reliable base for thorough sentiment analysis.

b) To devise a fine-grained multimodal sentiment classification model that leverages the interaction between image descriptions, image semantics, and textual content. The fine-grained task involves the accuracy identification of

sentiment polarity pertaining to specific entities within a combined image and text pair. This objective seeks to uncover effective strategies for integrating descriptions of images with other modal data, aiming to improve the accuracy in detecting the sentiment polarity of entities within multimodal content.

c) To integrate color information within multimodal sentiment analysis, aiming to overcome the limitations of existing approaches that neglect color information affecting sentiment. This objective explores the impact of color information on sentiment conveyance, with the intention of constructing a comprehensive multimodal sentiment analysis model that efficiently integrates color information alongside textual and visual data.

1.6 Scope of the Research

The objective of this thesis is to advance the field of multimodal sentiment analysis for social media content through the integration of visual and textual modalities. Its primary focus is to develop techniques and models that effectively address the distinct challenges and explore untapped opportunities in this domain. By incorporating both visual and textual information, this research aims to enhance the understanding and interpretation of sentiment expressed in social media posts.

The scope of this thesis encompasses the development, evaluation, and analysis of proposed multimodal sentiment analysis models. In the development phase, models will be designed and implemented to leverage both visual and textual information, enabling them to capture and interpret sentiment expressed in social media content effectively. These proposed models will be subjected to comprehensive evaluation and analysis to ensure their reliability and effectiveness. Comprehensive experiments on various multimodal datasets, performance metrics measurement, and

comparison with state-of-the-art techniques will all be part of the evaluation process. Subsequently, the analysis phase of this thesis will involve a detailed investigation into the strengths and limitations of the developed frameworks.

By addressing the limitations of existing approaches and exploring the integration of image and text modalities, this thesis extends the current body of knowledge in multimodal sentiment analysis and its related fields. It aims to contribute valuable insights, practical implications, and future research directions. Specifically, this research emphasizes the potential benefits derived from combining image and text modalities in multimodal sentiment analysis. Through the identification of these benefits, the thesis aims to shed light on new avenues for research and practical applications in this area.

1.7 Contributions of the Research

This thesis delves into the utilization of multimodal sentiment analysis techniques that effectively leverage both textual and visual modalities. Figure 1.4 portrays the research perspectives, research challenges, and research methods, presenting a comprehensive overview of the study. The fundamental objective of this inquiry is to tackle the inherent challenges associated with multimodal sentiment analysis, thereby enhancing our comprehension of sentiment analysis as a whole. By attaining precise identification and interpretation of emotional states, this research endeavours to foster an improved understanding of sentiment.

The outcomes of this study possess broad implications across multiple domains, including social media analysis, market research, and opinion analysis. The knowledge gained from this research has the potential to enhance decision-making processes, inform product design strategies, and elevate overall user experiences.

Furthermore, the incorporation of multimodal sentiment analysis has the capacity to identify and track sentiment trends, enabling timely responses to crisis events. Consequently, this work not only contributes to the advancement of emotionally intelligent technologies but also propels the evolution of social communication mechanisms.

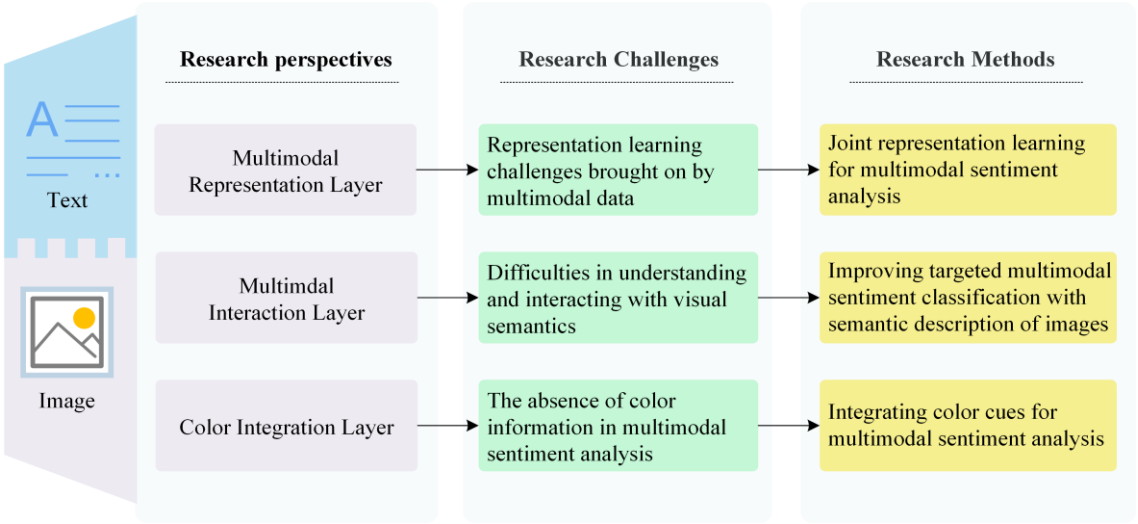

Figure 1.4 Comprehensive overview of the study

Alongside its societal impact, this thesis has made several significant contributions. These contributions involve exploring the semantic associations between multimodal data, information interaction, and color sentiment modelling within the context of multimodal sentiment analysis. These contributions are delineated from three distinct perspectives: the multimodal representation layer, the multimodal interaction layer, and the color information integration layer. Each of these perspectives contributes to a unified framework for multimodal sentiment analysis, and they are outlined below:

- In response to the problem of existing studies that typically rely on unimodal pre-trained models for feature extraction in the multimodal representation layer,

which leads to limitations in capturing representation information between multimodal data, this thesis introduces the concept of joint representation learning for multimodal sentiment analysis. This method involves the integration of a vision-language pre-trained model and the introduction of a multimodal contrastive learning method. This methodology enhances the accuracy and efficiency of multimodal sentiment analysis, offering a solution to a previously unaddressed gap in the field.

- In light of the prevailing trend in contemporary multimodal sentiment analysis methods, which frequently overlook the essential semantic information contained within images, this thesis aims to bridge this gap by integrating semantic descriptions of images into the fine-grained multimodal sentiment classification process. This approach is situated within the multimodal interaction layer, serving to not only confront the challenges associated with the interpretation and incorporation of visual semantics but also to model the intricate interactions between different modalities.

- Diverging from the conventional multimodal sentiment analysis methods that predominantly focus on extracting sentiment-related information from textual or visual content while often neglecting the potential contribution of visual cues, such as color, in understanding and categorizing emotions, this study introduces a color information integration layer on top of the existing multimodal sentiment analysis research framework. This layer is devised to model the intricate interplay between image color and the semantic information derived from diverse modalities, with the ultimate aim of achieving a more comprehensive comprehension and classification of sentiments.

1.8 Thesis Organization

This thesis is structured to effectively present the research conducted on multimodal sentiment analysis. The following sections provide a logical progression of the work accomplished:

Chapter 2: Literature Review offers a detailed theoretical overview and reviews the pertinent literature pertaining to sentiment analysis. Specifically, this chapter delves into textual sentiment analysis, visual sentiment analysis, and multimodal sentiment analysis. By examining the available quantitative evidence, it becomes evident that there is a growing interest among researchers in the field of artificial intelligence in multimodal sentiment analysis.

Chapter 3: Methodology provides a detailed account of the research methodology employed in this thesis. In alignment with the research objectives introduced in the preceding chapter, this section presents a comprehensive elucidation of the experimental setup, the rationale behind the choice of evaluation metrics, and the establishment of baseline models. These components form the essential building blocks of the methodological framework underpinning this thesis, facilitating a nuanced comprehension of the research methodology.

Chapter 4: Multimodal Sentiment Analysis discusses three multimodal sentiment analysis methods proposed in this study. The rationale and implementation of each method are described in detail. Subsequently, an experimental setup is performed for these methods, which includes the identification of evaluation metrics and benchmark models. The performance of these methods is evaluated through systematic comparisons and analyses conducted in the experiments. Conclusively, a further analysis of the proposed multimodal sentiment analysis approaches is

conducted, yielding significant empirical insights and theoretical contributions. This analysis elucidates the effectiveness and prospective impacts of the proposed methods, offering a comprehensive understanding of their potential to advance the field of sentiment analysis.

Chapter 5: Conclusion and Future Work summarizes the thesis's contributions to multimodal sentiment analysis, highlighting three proposed methods: joint representation learning, targeted multimodal sentiment classification with semantic image descriptions, and color information integration. It discusses the significant advancements these approaches offer in understanding sentiments across different modalities. Additionally, the chapter outlines limitations, such as dataset size and diversity, and suggests future research directions, including incorporating more modalities, exploring advanced deep learning architectures, improving cross-domain generalization, and expanding into multilingual sentiment analysis, to further advance the field.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Sentiment analysis, originally defined as the task of ascertaining the emotional state of text, has primarily focused on evaluating reviews and opinions found across diverse domains, such as products, movies, and various forums. However, with the exponential growth of social media platforms, sentiment analysis has become increasingly reliant on these platforms as a critical source of data. In particular, the rise of social networks, such as Flickr, Instagram, and Twitter, has facilitated the availability of multimodal data that combines both textual and visual elements. This evolution in user behavior, where individuals routinely share images along with textual content to convey their emotions or perspectives, has catalyzed the ascendancy of multimodal sentiment analysis as a prominent and burgeoning research domain.

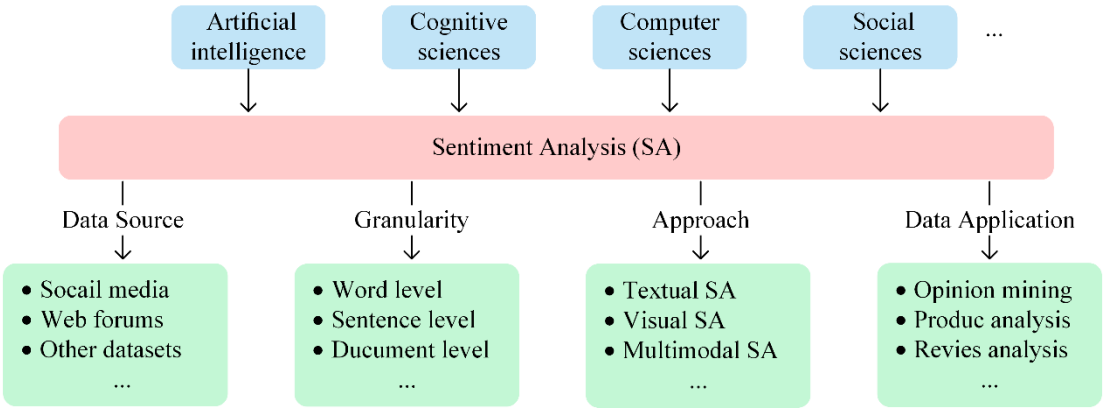

Figure 2.1 Sentiment analysis at a glance

An extensive illustration of the multidisciplinary nature of sentiment analysis, which lies at the intersection of computer science, cognitive sciences, artificial intelligence, and social sciences, is shown in Figure 2.1. This convergence allows sentiment analysis to draw upon the insights and methodologies derived from these

diverse fields, enriching its analytical capabilities. Historically, sentiment analysis has predominantly relied on three main approaches: textual sentiment analysis, visual sentiment analysis, and the fusion of text and visual modalities, forming the foundation of multimodal sentiment analysis. The focus of this thesis is specifically on multimodal sentiment analysis, which entails the integration of techniques that facilitate the analysis of both textual and visual content. Each of these approaches encompasses an extensive array of techniques and methodologies, carefully tailored to the intricate task of sentiment analysis.

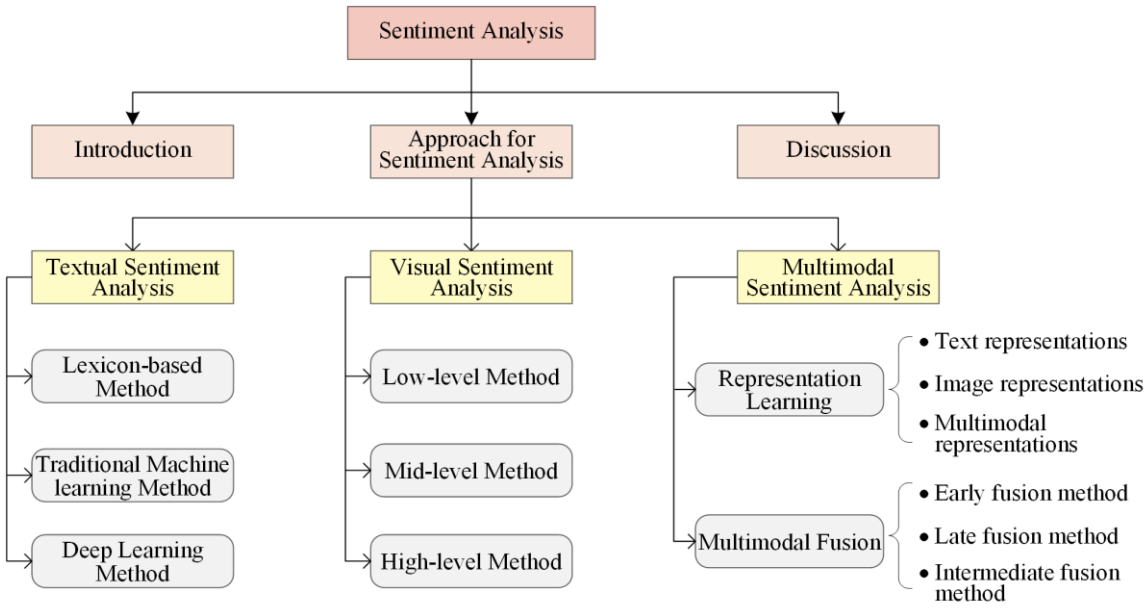

Figure 2.2 Organization of the sentiment analysis literature review

This chapter provides an extensive literature review that covers the different facets of sentiment analysis, progressing from textual sentiment analysis to visual sentiment analysis and ultimately to multimodal sentiment analysis, as shown in Figure 2.2. Section 2.2 presents an inclusive survey of textual sentiment analysis as applied to social media platforms, detailing the principal approaches and techniques employed within this domain. Subsequently, in Section 2.3, an overview of visual sentiment analysis is provided, detailing the primary strategies and challenges entailed

in the analysis of sentiment within visual content. In Section 2.4, an in-depth exploration of multimodal sentiment analysis is undertaken, wherein the fusion of textual and visual modalities is leveraged to infer sentiment. Finally, Section 2.5 critically evaluates the existing works pertaining to multimodal sentiment analysis, evaluating their strengths and limitations. Through an examination of diverse methodologies and frameworks, the objective of this literature review is to furnish a comprehensive comprehension of sentiment analysis, while simultaneously discerning prevailing gaps and challenges within the field.

2.2 Textual Sentiment Analysis

The rapid growth of contemporary social media and digital communication technologies has exposed us to a vast amount of textual data on a daily basis. This data comprises various forms of content, such as social media posts, news articles, online comments, and other text-based sources. Consequently, sentiment analysis has emerged as a prominent research area, as this textual data contains a substantial amount of sentiment information. Specifically, textual sentiment analysis, a subset of sentiment analysis, focuses on analyzing sentiment from these textual data sources. Its potential applications in diverse domains, including marketing, customer service, and political analysis, among others, have led to significant attention in the field of natural language processing.

The primary objective of textual sentiment analysis is to determine the attitude or sentiment of a writer or speaker towards a particular topic, product, service, or entity. Typically, this task involves assigning a polarity label to the text, indicating whether the sentiment expressed is positive, negative, or neutral. This polarity detection allows for a straightforward categorization of the sentiment, enabling quick assessments of

overall sentiment distribution in large datasets. Some approaches go beyond simple polarity detection and attempt to identify more nuanced emotions such as joy, anger, sadness, or fear. By capturing these subtler emotions, sentiment analysis can provide a more comprehensive understanding of the underlying sentiments expressed in the text.

Textual sentiment analysis can be conducted at different levels of granularity, including document level, sentence level, and entity level. The task of document level sentiment classification is a fundamental aspect of natural language processing, wherein the sentiment of an entire document is analyzed and determined. This task treats the document as the primary unit of information, aiming to classify it into either positive or negative polarity based on the overall sentiment expressed. By considering the document as a whole, this approach enables a comprehensive understanding of the sentiment conveyed and provides valuable insights into the overall opinion or attitude towards a particular topic or object. Sentence level sentiment classification, in contrast, is concerned with the task of classifying the sentiment expressed within individual sentences. This form of sentiment analysis aims to determine whether a given sentence conveys a positive, negative, or neutral sentiment. Through the isolation and examination of sentiments at the sentence level, researchers can attain a deeper understanding of the nuanced emotions and opinions articulated within a text, thereby enriching the overall comprehension of the conveyed sentiment. Entity level sentiment analysis aims to determine the sentiment expressed towards specific entities mentioned in a sentence, such as individuals, organizations, or brands. This analysis becomes particularly significant in the context of social media, where messages are typically constrained to a limited number of words. Consequently, the differentiation between document and sentence levels is unnecessary in this domain. Hence,

leveraging sentence level and entity level textual sentiment analysis proves to be highly valuable for social media sentiment analysis (Giachanou & Crestani, 2016).

Sentiment analysis, a significant task in natural language processing, has been the focus of numerous research efforts, leading to the development of diverse methodologies, each exhibiting unique strengths and constraints. The approaches to sentiment analysis can be broadly classified into three main groups: lexicon-based methods, traditional machine learning methods, and deep learning methods. The following subsections will provide an overview of the various approaches related to textual sentiment analysis, emphasizing their respective notable advantages. A comprehensive summary of the distinctive characteristics of each approach is available in Table 2.1, based on the works of Al-Tameemi et al. (2022), Birjali et al. (2021), Wankhade et al. (2022) and Das & Singh (2023).

Table 2.1 Advantages & disadvantages of textual sentiment analysis approaches

Method	Advantages	Disadvantages
Lexicon-based	• No labelling is required • Interpretability • Low computational cost	• Limited vocabulary coverage • Oversimplification • Difficulty of updating • Depends on the domain
Traditional machine learning	• Dictionary is unnecessary • Ability to learn from data • Generalize to new domains	• Need for labeled data • Feature engineering • Prone to overfitting • More time is required
Deep learning	• Automatic feature learning • Improved accuracy • Transfer learning • End-to-end training	• Large data requirements • Computational resources • Black-box nature • Hyper-parameter tuning

2.2.1 Lexicon-based Method

The Lexicon-based method, also known as the knowledge-based method, is one of the earliest methods used for textual sentiment analysis. This approach utilize precompiled sentiment lexicons, which consist of various words along with their associated polarities, to determine whether a given word falls into the positive or negative sentiment class. (Kiritchenko et al., 2014). Commonly, positive words are characterized by terms such as "good" or "beautiful," while negative words can be exemplified by expressions like "bad," "ugly," or "scary." The assigned score can be a polarity value, such as +1, -1, or 0 for positive, negative, and neutral terms, respectively, or a numerical value indicating emotional intensity or power. To calculate the final sentiment, the emotional scores of each word are summed or averaged. If a positive match is found, the total score of the input text is increased, and if a negative match is found, the total score is decreased.

The general flow of the Lexicon-based approach is illustrated in Figure 2.3. This approach is widely used in traditional textual sentiment analysis since it does not require training with labelled samples beforehand.

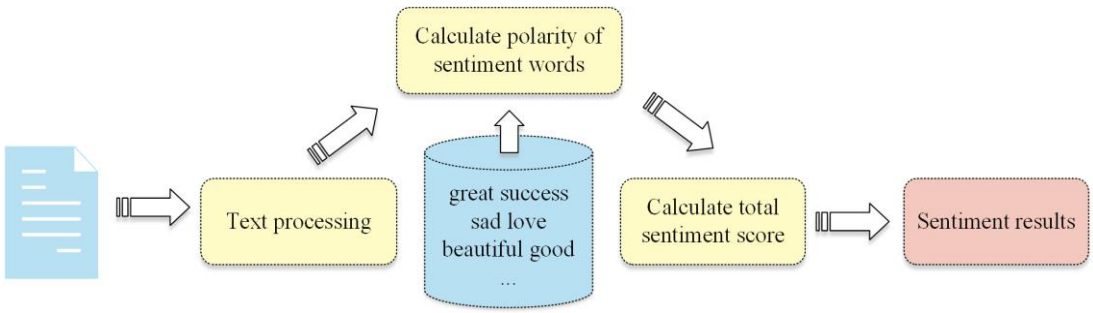

Figure 2.3 The general flow of the Lexicon-based sentiment analysis.

Dictionary-based and corpus-based are the two main methods used to develop sentiment lexicons. Starting with a limited set of sentiment words, the dictionary-