Research on Textual Emotion Recognition based on Deep Learning Methods

邓 佳文

A Thesis submitted to Tokushima University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March, 2021

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
Tokushima University, Japan

Contents

List of Tables

List of Figures

Acknowledgment

I would like to use this opportunity to express my gratitude to everyone who supported me throughout my Ph.D course. I thank all the teachers and friends for their aspiring guidance, friendly advice, encouragement, and assistance.

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Fuji Ren for the continuous support of my Ph.D study and related research. He provided me with an excellent research environment. With his patient guidance and encouragement, I have the chance to attend international conferences and publish papers in authoritative journals. These research experiences have broadened my horizons and increased my knowledge in the field of affective computing and NLP, which will benefit me throughout my life.

I would thank Prof. Kenji Terada and Prof. Masami Shishibori for their time and effort in reviewing my thesis. Their valuable suggestions helped improve this thesis. I would like to thank all teachers and classmates in Ren Lab for their help during this period. Thanks to Shun Nishide and Xin Kang for their practical suggestions and technical help. Thanks to Duo Feng, Ziyun Jiao, Tianhao She, Mengjia He and Yangyang Zhou for their generous sharing of experiences and ideas in research. Working with you is always a relaxing time for me. I would also like to thank the graduated seniors, Ning Liu and Chao Li. It is your help that made my research and life at Tokushima University start so smoothly. I especially want to appreciate Siyuan Xue and Qian Zhang for their continuous encouragement and company in these three years of study and life. Many thanks all for creating and keeping a harmonious team.

I give my gratitude to the financial support of China Scholarship Council (CSC), which enabled me to pursue my study in Tokushima. I also appreciate all my friends supporting me all the time. I appreciate my family for keeping me warm and inspired all the time. I know that you are always there, whenever for whatever.

Finally, thank all the people around me or related to me during my Ph.D course. With you all, living and studying at Tokushima University was an unforgettable experience!

Abstract

Textual emotion recognition (TER) is the process of automatically identifying emotional states in textual expressions. It is a more in-depth analysis than sentiment analysis. Owing to its significant academic and commercial potential, TER has become an essential topic in the field of NLP. Over the past few years, although considerable progress has been conducted in TER, there are still some difficulties and challenges because of the nature of human emotion complexity. This thesis explores emotional information by incorporating external knowledge, learning emotion correlation, and building effective TER architectures. The main contributions of this thesis are summarized as follows:

(1) To make up for the limitation of imbalanced training data, this thesis proposes a multi-stream neural network that incorporates background knowledge for text classification. To better fuse background knowledge into the basal network, different fusion strategies are employed among multi-streams. The experimental results demonstrate that, as the knowledge supplement, the background knowledge-based features can make up for the information neglected or absented in basal text classification network, especially for imbalance corpus.

(2) To realize contextual emotion learning, this thesis proposes a hierarchical network with label embedding. This network hierarchically encodes the given sentence based on its contextual information. Besides, an auxiliary label embedding matrix is trained for emotion correlation learning with an assembled training objective, contributing to final emotion correlation-based prediction. The experimental results show that the proposed method contributes to emotional feature learning and contextual emotion recognition.

(3) To realize multi-label emotion recognition and emotion correlation learning, this thesis proposed a Multiple-label Emotion Detection Architecture (MEDA). MEDA comprises two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). MEDA captures underlying emotion-specified features with MC-ESFE module in advance. With underlying features, emotion

correlation learning is implemented through an emotion sequence predicter in ECorL module. Furthermore, to incorporate emotion correlation information into model training, multi-label focal loss is proposed for multi-label learning. The proposed model achieved satisfactory performance and outperformed state-of-the-art models on both RenCECps and NLPCC2018 datasets, demonstrating the effectiveness of the proposed method for multi-label emotion detection.

Chapter 1

Introduction

1.1 Motivation

Emotion interaction is a common psychological phenomenon in human's daily life. Accurate emotion recognition is the premise of effective human communication, interaction, and decision making. With the development of big data technology and artificial intelligence, the research of emotion recognition systems has become a typical project in both academia and industry.

As the most basic and direct carrier, textual data in emotional communication is commonly utilized to infer emotional states such as joy, sadness, and anger. Moreover, in recent decades, with the rapid development of social media, mobile networks, and smart terminals, online social networks (such as Facebook, Twitter, Weibo, and Line) have become an unprecedented global phenomenon. Through these platforms, humans express their thoughts, connect, and communicate with each other. In this way, a large amount of data is generated, including posted ideas, feelings, and even photos and videos. These resources contain rich emotional information and provide a data foundation for emotion-related research. On the other hand, people's demands for mental health care and emotional management are continually increasing, helping to understand themselves better and find a more effective way to learn, work, and live. These facts have accelerated the need for fine-grained emotion recognition.

Emotion recognition task has gained considerable interest from the research community, aiming to point out the subtle differences in emotional content. Textual

emotion recognition (TER) automatically identifies human emotions in their textual expression, such as Happy, Sad, Angry, Fearful, Excited, Bored, or anything else. This task can be easily complete by humans based on their subjective feelings. Nevertheless, for automated TER systems, computational methodologies need to be continuously developed and optimized to achieve more accurate emotion prediction.

1.2 Significance of Research

The term of emotion recognition and sentiment analysis are often used interchangeably [1]. In fact, there are apparent differences between these two concepts [2]. Sentiment analysis mainly measures subjective attitudes to entities from the perspective of sentiment polarities, such as positive-negative dimensions or sentiment intensity. In contrast, emotion recognition involves identifying more detailed and explicit emotional states, referring to a wide range of mental states, such as happiness, anger, sadness, and fear.

As an essential element in human nature, emotions have been widely studied in psychology [3]. TER aims to classify a textual expression into one or several emotion classes depending on the underlying emotion theories employed [4], such as Ekman's six basic emotions, including fear, anger, joy, sadness, disgust, and surprise.

Current and potential applications of automatic TER systems have been entered various aspects in daily human life, including public opinion monitoring and mental health monitoring. In emotional management, according to the emotional rhythm displayed by Twitter users in different periods, the mood curve could be drawn to facilitate understanding of people's work status and mental state [5], [6], [7]. In marketing communications, it can help to improve business strategies based on consumer preferences [8]. In Social networks, emotion analysis can also be applied in sinister tone analysis, helping detect potential criminals or terrorists [9]. Emotion detection during crisis or disaster situations helps understand peoples' feelings towards a particular situation, useful to crisis management and critical decision-making [10]. During elections, public emotions can be tracked and predicted based on their speeches and comments online [11], [12]. In human-computer interaction systems, such as dialog systems, automatic question-answering systems, and companion robots, emotion recognition

technology is conducive to improving user experience. In an effective e-learning system, real-time emotion tracking improves students' motivation and effectiveness [13]. Text is the fundamental modality, and TER is the most convenient way to understand the emotional conversation, making human-computer interaction more accurate and intelligent. The nature of emotion indicates the potential applications of emotion recognition in various fields. More detailed information is shown in Table 1.1. The resulting market demand has greatly promoted TER in the field of NLP.

Table 1.1 Applications of emotion recognition

Domain	Applications	Details
Marketing	• Product marking	By conducting emotion analysis of product reviews online, companies can further understand consumers' attitudes and reviews on the product, thereby provide more relevant services to users and promoting marketing [14], [15], [16].
	• Prediction of purchase intentions	Purchase intention prediction by analyzing customers' emotional states [17].
	• Brand management	Brand merchants gain a competitive advantage by analyzing consumers' mental states and formulating effective marketing strategies [18].
	• Text-to-speech synthesis	Realize proper expressive rendering of text-to-speech synthesis in children's fairy tales [19].
Information prediction	• Financial prediction	Leverage emotional signals in financial materials to suggest trading decisions and forecast the economic climate [20], [21].
	• Election prediction	Emotion recognition of election tweets helps to understand how public emotion is formed and can further predict elections [22].
Personalized recommendations	• Music recommendation	Improve the availability of online music streaming services by developing emotion-based access methods and creating emotion-based playlists [23], [24], [25].
	• Children's information retrieval systems	Optimize children's information retrieval systems by analyzing children's information-seeking behavior and affective state, [26], [27].

Social networks	• Self-emotional management	Analyze people's personality characteristics and emotional states based on user's expression on social networks. This will help users better understand themselves and find more suitable ways to learn work and life [6], [7].
	• Public opinion monitoring	Through continuous monitoring of Internet public opinion trends, learn about users' different views on popular social issues. Maintain social security management by guiding the direction of online public opinion timely and correctly [28].
	• Disaster response	Information from bystanders and eyewitnesses in social media platforms can help law enforcement agencies obtain firsthand and credible information about an ongoing situation, thereby gaining situational awareness and other potential uses. [29].
Healthcare	• Emotional monitoring and guidance	According to user's expression, detect dangerous emotions, such as stress, pain, fear, and panic. When there is a dangerous emotional tendency, guide the user's emotions through intervention, especially for patients with autism or depression.
	• Suicide intervention	Conduct large-scale automated surveillance of social networking sites to identify people who may be at risk for suicide. Once discovered, suicide can be prevented through intervention and other measures [30].
Human-Computer Interaction	• Emotional Chatting Machine	Emotional dialogue system can generate appropriate responses with a particular emotion and increase user satisfaction [31], [32], [33].
	• Emotional care robot	Emotional care robots can mitigate stress, anxiety, and pain in hospital care [34].
	• Spoken tutoring systems	Improve online learning experience and increase the learning rate through real-time emotional feedback from students [35].

1.3 Main Research Contents

TER task brings some open challenges for NLP researchers, including the shortage of high-quality data, the complex nature of textual emotion expression, and of course, how

to design an effective TER system.

TER task is a classification problem aiming to predict all possible emotion labels for a given textual data. Although many public emotion-related databases have been proposed, the imbalanced data distribution between each category largely influences the classification performance. Besides, with the development of big data, the resources available online have shown explosive growth. How to effectively use these data, whether labeled or unlabeled, is a considerable challenge in the NLP field.

Human emotion expression and understanding are complex and subjective. The same expression may produce different emotional feelings in different scenes. Therefore, how to effectively recognize emotions according to the contextual information is worthy of attention. Furthermore, the nature of human emotional expression is complex. Many emotional categories have a particular connection, and there is no distinct boundary, such as love and happiness. In this way, the emotional classification task cannot be regarded as a simple single-label classification problem but a more complex multi-label classification task. Besides, emotions are related to and influenced by each other. How to improve emotion recognition performance by effectively emotion correlation learning is a core problem in this task.

This thesis revolves around the above challenges. The main research contents mainly include three parts.

(1) Text Classification with External Background Knowledge

TER task is a particular text classification task aiming to assign all possible emotion labels for a given textual data. As a fundamental and typical NLP task, the quality of datasets guarantees the performance of the model. However, most of the existing labeled databases are inevitably limited by data imbalance, resulting in the classifier tending its performance to those categories with more texts.

This thesis tries to incorporate external background knowledge in classification model to alleviate the problem of data imbalance. Therefore, a background knowledge-based multi-stream neural network is proposed, aiming to make up for the limitations of imbalance or insufficient information in training data. The multi-stream network mainly consists of the basal stream for retaining original sequence information and background

knowledge-based streams for information supplement. Background knowledge is composed of categorical keywords and co-occurred words, which are extracted from external corpora. Different fusion strategies are also proposed to fuse the features extracted from different streams effectively.

(2) Hierarchical Network with Label Embedding for Contextual Emotion Recognition

Recent researches about TER mainly conducted on sentence-level, which aimed to recognize subtle emotions based on word and concept-based features extracted from the given sentence. However, emotional expression is complicated, and the same sentence could present different emotions in different contexts. In the absence of contextual information, even humans cannot give confident emotional judgments. Therefore, it is necessary to utilize contextual information for sentence-level emotion recognition.

Considering the importance of contextual information and emotion correlation, a hierarchical model with label embedding is proposed in this thesis to realize accurate emotion recognition. The hierarchical model is utilized to learn the emotional representation of a given sentence based on its contextual information. To realize emotion correlation learning, the label embedding matrix is trained by joint learning, which is beneficial to emotion correlation-based emotion prediction.

(3) Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning

Most of the existing psychological emotion models divide emotions into several categories, which are oversimplified and ignore the diversity of emotions. Human emotion is complex in reality. Many emotional categories have a particular connection, and there is no distinct boundary. Thus it is difficult to match an accurate label for emotional expression. Besides, emotions are very subjective feelings. Treating the same emotional expression text, humans may feel different emotions according to their own experience. Emotion recognition task becomes challenging because of fuzzy emotional boundaries and human's subjective feelings.

Considering that multiple emotions often co-occur with non-negligible emotion correlations, this thesis tries to recognize all associated emotions with a Multi-label

Emotion Detection Architecture (MEDA). MEDA is mainly composed of two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). MEDA captures underlying emotion-specified features through MC-ESFE module, which is composed of multiple channel-wise ESFE networks. Each channel in MC-ESFE is devoted to the feature extraction of a specified emotion from sentence-level to context-level through a hierarchical structure. With underlying features, emotion correlation learning is implemented through an emotion sequence predictor in ECorL. Furthermore, a new loss function: multi-label focal loss, is proposed. With this loss function, the model can focus more on misclassified positive-negative emotion pairs and improve the overall performance by balancing the prediction of positive and negative emotions.

1.4 Organizational Structure

This thesis mainly investigates the background and research status of the TER task, and proposes some TER approaches for addressing some existing challenges. The organizational structure of this thesis is as follows:

Chapter 1: talks about the motivation and significance of TER task, and introduces the main research contents and organizational structure of this thesis.

Chapter 2: introduces the background and related works of TER task. This chapter first introduces some existing psychological emotion models and publicly available emotional resources. Then the research status and progress of TER technology in recent years is reviewed.

Chapter 3: introduces the multi-stream neural network with external background knowledge. Background knowledge mainly refers to keywords and co-occurred word pairs extracted from external corpora. This model tries to make up for the limitations of imbalance or insufficient information distribution in training data.

Chapter 4: explores the influence of contextual information on sentence-level emotion recognition. A hierarchical model with label embedding network is proposed to learn the contextual information and emotion correlation by joint learning.

Chapter 5: explores emotion correlations by proposing a Multi-label Emotion

Detection Architecture (MEDA). Furthermore, a new loss function named multi-label focal loss is defined to balance the prediction of positive and negative emotions.

Chapter 6: concludes the whole thesis and discusses future works.

Chapter 2

Background and Related works

2.1 Psychological Emotion Models

Emotion recognition is a classical problem in cognitive science and artificial intelligence. Before emotion recognition, it is necessary to establish the representation model of emotion states. In psychology, emotions are divided into basic emotions and complex emotions according to whether an emotion is difficult to summarize in one word. There are mainly two kinds of psychological emotion models describing how humans perceive and classify emotion: discrete categorical model and dimensional emotion model.

Discrete emotion models have been widely used in emotion recognition task because of simplicity and intuitiveness. Emotions are classified into several basic emotions, which are often relatively independent. Although many studies are devoted to classifying human emotions, there is no consensus on the definition of basic emotions. A typically utilized discrete emotion model is proposed by Paul Ekman [36]. He concludes some universal emotions and proposed six basic emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. Most emotion annotation tasks are based on Ekman's theory or its extension. Parrott [37] proposes a tree-structured emotion classification model for recognizing more than one hundred kinds of subtle emotions. This model has three layers in total, and the bottom layer consists of Ekman's basic emotions. Discrete emotion models intuitively represent emotions and make it easier to assign emotion labels in manual or automatic annotation tasks. In this way, emotion recognition task can be interpreted as a

classification task with categorical emotion labels. However, there are also some limitations in application. Each emotion category is always represented by a particular word, resulting in a limited range of emotional states that can be expressed. There is a high degree of correlation between some emotions, and the generation, development, and disappearance of emotions is a dynamic process, which is challenging to describe detailed by a discrete emotion model [38].

Dimensional emotion model measures emotion states with numerical dimensions. Each emotion is described as a multi-dimensional vector. In each dimension, the value is continuously changed to distinguish the nuances of emotion, and the extremes of two directions mean two polarities. PAD model is a typical used dimensional model [39], representing emotions with three dimensions: pleasure, arousal, and dominance. The pleasure dimension is also called the valence dimension, which is a measure of the degree of human pleasure from one extreme 'distress' to the other extreme 'ecstatic'. Arousal dimension is also called the activation dimension, measuring the level of physical activity and psychological alertness. Dominance dimension is also called attention or power dimension, referring to a feeling that affects the surrounding environment [40]. Russell's circumplex model [41] consists of bivariate classifications into valence and arousal. He thinks that the dominance dimension is more related to cognitive activities, and the two dimensions of VA could represent most of the different emotions. Depending on the strength of both components, certain regions in VA space are given explicit interpretations according to 28 emotional states. Plutchik proposes a wheel of emotions [42], [43] to describes how basic emotions are related. In the wheel, eight basic emotions are organized into four bipolar axes: joy-sadness, fear-anger, trust-disgust, and surprise-anticipation with different intensity levels. He thinks other complex emotions can be viewed as combinations of the relevant primary ones. This idea enables us to implement emotion detection more comprehensively. The Hourglass of Emotions [44] is a complex hybrid emotion model, in which discrete categories and four independent but concomitant affective dimensions are utilized to represents affective states.

Compared with the dimensional model, the discrete model is more widely used in emotion recognition research because of its intuitive and straightforward advantages.

2.2 Emotional Resources for Emotion Recognition

Deep learning-based TER system is data-driven and relies on a large amount of data. Standard, free, and generalized annotated databases are the guarantee of model performance. This section discusses the publicly available databases containing emotion knowledge.

Most existing prominent and available public corpora are annotated based on discrete emotion models. Their categorical emotion annotation schemes are often built based on Ekman and Plutchik's basic emotions or extension. SemEval-2007 Affective Text Task corpus contains 1250 news headlines, which are extracted from news websites and newspapers, and are annotated with Ekman's basic emotions [45]. ISEAR contains 7600 self-reported experiments of emotion-provoking text about their reactions to seven primary emotions [46]. NLPCC-2018 database contains 7928 code-switching texts with five emotion labels, and each text contains more than one language (Chinese and English) [47]. It was the benchmarking data for NLPCC Shared Task of Emotion Detection. Alm's fairy tale dataset consists of 1580 sentences from 185 children fairy tales and is annotated with eight emotion labels: Angry, Disgusted, Fearful, Happy, Sad, Positively Surprised, Negatively Surprised, and Neutral [48].Ren-CECps [49], [50] contains 1487 blogs with 34, 719 sentences in Chinese. It is hierarchically annotated in the document, paragraph, and sentence level. Each level is annotated with eight basic emotions and corresponding emotion intensity.

As the basic emotional knowledge, most emotion lexicons are categorically annotated. WordNet-Affect [51] and SentiSense Affective Lexiconv [52] are concept-based affective lexicons. They are suitable to represent affective concepts correlated with affective words. NRC Emotion Lexicon [53] (also called EmoLex) is a word-emotion association lexicon available in 40 languages. It is annotated with eight basic emotions and two sentiments. LIWC2015 [54] (Linguistic Enquiry and Word Count, LIWC) is hierarchically annotated with both sentiments and emotion labels. It is provided in English and has been translated into several languages, including Arabic, Korean, Turkish, and Chinese.

2.3 Conventional Approaches for TER

As obvious clues, lexicon resources contain emotional knowledge [55] and are widely used in the process of TER. Keyword spotting is the most naive approach because of its accessibility and economy. Emotion is recognized by mapping keywords into relatively unambiguous affect words in emotion lexicon. Lexical-affinity techniques [56] predict emotions according to measuring the relationships between co-occurred words in the same document, but it does not work well when it comes to negative words. One of the advantages of the keyword-based approach is the accessibility and economy of abundant emotional resources. However, it depends too much on the coverage and is significantly affected by the absence of emotional keywords [57]. The annotation quality also has a decisive influence on the prediction accuracy. In most emotion lexicons, the word is associated with only one emotion, which oversimplifies complex emotional nature and ignores the emotional correlation. The experimental results in [58] suggest that its performance cannot be guaranteed because the semantics of the keywords heavily depend on the contexts.

Machine learning-based statistic model has been widely used in this task, alleviating the aforementioned limitations to some extent. Emotion recognition can be treated as a regular text classification problem utilizing handcrafted syntactic and linguistic features. Bag of words (BoW) model along with feature-extraction techniques (such as TF-IDF weighting) are often applied to mapping the text into a feature vector [59]. Finally, machine learning algorithms are applied for final emotion prediction, such as LDA, SVM, KNN, Decision Tree, and Naive Bayes [60], [61]. Feature extraction is the critical step, and there are various basic features, such as TF-IDF, n-grams, and special symbols of emoticons and punctuations. Most features are extracted based on word occurrence frequency and BOW model. With these handcrafted features, texts are represented as feature vectors for final prediction.

Emotion lexicons are often served as prior knowledge to obtain emotional features[62], [63]. Searching for emotional words in input text is the most direct way. Frequencies of lexicon words are often utilized to construct the lexicon-based feature vectors. Emotional features are often combined with generic n-gram features to enhance text representation.

The works in [64] combine two kinds of features. One is corpus-based unigram features, and the other is lexicon-based features derived from Roget's Thesaurus (RT) [65] and WordNet-Affect. In [66], they consider both the appearance number and the intensity of emotional words. These emotional features are combined with Term weighting (TF) sentence representation for final classification. Emoticons are utilized hundreds of times more often on Twitter and contain direct emotional clues. With the lexicon of General Inquirer [67] and WordNet-Affect, S. Aman et al. [68] extracted emotional words and symbols (emoticons and punctuations) as features for the training of the classifier with popular used Naïve Bayes and SVM.

Utilizing implicit emotional words is not the only way of expressing emotion. A sentence without any emotional word may become emotion-bearing depending on the context or underlying semantic meaning. For instance, the sentence: 'What if nothing goes as planned?' implicitly expresses 'fear' without using any emotion bearing word. Therefore, contextual and semantic analysis is necessary for accurate emotion recognition.

Semantic-based features contribute to analyzing those expressions that do not convey explicit emotion but include emotional concepts. Such as the evaluation of the semantic similarity among generic terms and affective lexical concepts. In most cases, external knowledge, such as ontologies and lexicon, are often introduced to extract semantic features. They are often utilized as an indirect bridge for calculating the semantic similarity between text and an emotion state. In [69], LDA (latent semantic analysis) is used to calculate the semantic similarity between common words and emotional words, by which semantic features could be extracted for recognizing the emotions of social news. In [50], the polynomial kernel method is proposed to compute the similarities between sentences and emotional words. They utilize an emotion lexicon derived from Chinese emotion corpus: Ren-CECps, which is annotated with 8 basic emotions. In [70], semantic similarity between input words and lexicon words is computed and then combined with word embedding features. The works in [71] demonstrate the ability of word mover's distance (WMD) in measuring the semantic difference between sentences.

Commonsense knowledge is often introduced into TER model. The works in [72] propose a commonsense affect model to evaluate the affective qualities of textual underlying semantic content. They first mine affective commonsense from a large scale

of generic knowledge bases: Open Mind Common Sense (OMCS) [73]. This information is utilized to build a small society of commonsense-based linguistic models. This model is implemented in an emotionally responsive email browser called EmpathyBuddy and the feedback suggests its robust in emotion recognition.

Rule-based affect models are often utilized to determine the emotional affinity of individual sentences. Some rules are established in [74] to determine the emotions in sentences in blog posts. Their analysis relies on a manually prepared database of words, abbreviations, and emoticons with emotional annotation. In [75], according to whether input text containing the emotional keywords, there are two parallel classification methods. If containing emotional words, traditional keywords spotting method [76] is applied to predict the emotion. Otherwise, hierarchical semantic features are extracted based on emotion generation rules for final prediction.

Context emotion clues are essential attribution benefiting to emotion recognition, especially in long texts. For example, the emotion ‘angry’ is more likely to turn into ‘angry’ than ‘Joy’. These statistical rules revealed in emotion transfer indicate the continuity of emotions in adjacent contexts. In [77], such clues are considered in their weighted high-order hidden Markov models (HMMs). They find that the sentence emotions are clearly affected by the direct previous three consecutive sentences.

Traditional methods heavily rely on hand-crafted features along with shallow models, such as SVM and logistic regression. External emotional resources are often introduced as prior knowledge for emotional enhancement. However, these hand-crafted features are represented in a high-dimensional and sparse matrix, which are often incomplete and time-consuming, limiting the performance of TER model to some extent.

2.4 Deep Learning-based Approaches for TER

This section reviews some deep learning-based methods for TER tasks, including pre-trained word embedding models, basal neural networks, and some derived variations. Some techniques for performance enhancement are also given, including knowledge enhanced model and transfer learning.

TER task is a particular text classification problem. Most researches focus on creating

and training an effective network to learn multi-layer feature representation automatically. In deep learning-based neural networks, word embedding is often applied as the first step to obtain the distributed representation of input text. Each input word is represented in an n-dimensional vector by word embedding techniques. The distance between vectors corresponds to the semantic similarity of the word pairs. The word embeddings are then fed into neural networks, such as RNN and CNN, for further learning and emotion prediction.

2.4.1 Pre-trained Language Model

Word embedding is a technique based on distributional semantic modeling and aims to learn latent, low-dimensional representations from the language structure. Pre-trained word embedding alleviates the problems of sparse features and high-dimensional representation in traditional bag-of-words models. Some well-established embedding models are widely used in NLP tasks and have shown great success.

Early word embedding model is trained based on the syntactic context. It is believed that frequently co-occurred words are often similar in some semantic criteria. The typical models, Word2vec [78], [79] and GloVe [80], are trained on a large scale of unlabeled data to capture fine-grained syntactic and semantic regularities. Early word embedding performs better than randomly initialized word vectors and has shown great success in NLP tasks, such as word similarity tasks and named entity recognition benchmarks. However, it assumes that ‘A word is represented by a unique vector’ and ignores words differences in different contexts. Each word is embedded into a unique vector, whether monosemous or polysemy, thereby antonyms words with the same language structure often have similar vectors [81]. This meaning conflation deficiency largely hampers the effectiveness of word embedding.

Inspired by the successful transfer learning of CNNs from the image field to other computer vision fields, the emergence of pre-trained language model opened the pre-training era in the NLP field. The pre-trained language models generate contextualized word embedding with the general knowledge that can be easily transferred to almost all downstream tasks. In [82], LSTM-based network is trained on English-German

translation task for context learning. The output of hidden layer is obtained as context vectors (CoVe). CoVe can be taken as word embedding with contextual information and contributes to other NLP tasks, including sentiment analysis and question answering. The contextualized word embedding model, ELMo [83], captures the variation in the meaning of a word depending on its context. Word representations are the real-time output of pre-trained BiLSTM language model through the input sentence. Its word embedding is dynamic and context-sensitive. The same word with different contexts is represented in different word vectors, which significantly alleviates the ambiguity limitation. BERT [84], the Bidirectional Encoder Representations for Transformers, is a pre-trained model producing context representations that can be very convenient and effective. Through a deep network architecture, the language model learns to predict unseen words in the context by unsupervised learning. A large amount of unlabeled data is utilized during the training, which helps the model learn useful linguistic knowledge. Bert performs well in encoding contextual grammatical knowledge, and have achieved satisfactory results in many NLP task. The emergence and success of BERT inspires more and more pre-trained models, including GPT/GPT-2 [85], [86], Transformer-XL [87], XLNet [88], MASS [89], UNILM [90]. The pre-trained language models work well in practice on multiple tasks and have become a new paradigm in NLP field.

Emotional knowledge can be represented in different ways. Motivated by word embedding, word-level emotional representations have shown remarkable effectiveness in different emotion-related tasks. To learn generalized emotion representation, Emo2Vec is proposed in [91] to encode emotional semantics. This model is pre-trained in six different emotion-related tasks by multi-task learning, including emotion/sentiment analysis, sarcasm classification, and stress detection. Emo2Vec is often utilized by concatenating with other embeddings, such as GloVe, for more competitive performances. Emoji widely used in social networks. Emoji2vec [92], is released for all unicode emoji representations, which directly maps emojis to continuous representations. DeepMoji [93] is a pre-trained model with rich representations of emotional information. This model consists of two-layer BiLSTM network and an attention layer, and 1.2 billion tweets with emojis are utilized during training. Through transfer learning, DeepMoji obtains satisfying performance on various emotion-related tasks. In [94], the performance of

several well-known pre-trained embeddings are compared in SenEval 2019 corpus: DeepMoji, ELMo, GLoVe, Emo2vec, BERT, and Emoji2Vec. DeepMoji outperforms others by a large margin, mainly because of the similar emotional training datasets utilized in DeepMoji and the target task. It also suggests the importance of selecting an appropriate pre-trained model for target tasks, contributing to better feature extraction. For example, GLoVe and BERT are often applied in general embedding, while DeepMoji and Emo2Vec are more appropriate for emotion embedding.

Some works try to improve the performance of word embedding by incorporating emotional information. To distinguish words with similar syntactic context but opposite sentiment polarity, a sentiment-specific word embedding (SSWE) model is proposed in [95]. During the model training, sentiment information is integrated into the loss function. Its effectiveness has been verified in sentiment analysis of tweets in SemEval 2014 Task 9 subtask(b) [96]. A domain-sensitive and sentiment-aware embedding (DSE) model is proposed in [97], which jointly models the sentiment semantics and domain specificity of words. Their embedding model with emotional information leads to better performance for emotional recognition.

2.4.2 Knowledge Enhanced Representation

Prior knowledge is often incorporated into deep neural networks as auxiliary information for deeper language understanding, including emotional lexicon resources, commonsense, linguistic patterns, affective semantic rules, and any other emotion-related knowledge. Incorporating prior knowledge contributes to enhancing emotional feature representation and realizing more accurate emotion recognition.

Fusing deep learning-based features and lexicon-based term frequency features is the most direct way to realize emotion enhancement [98]. The combined features are fed into a deeper network for learning more high-level and abstract features or fed into classifiers directly for final prediction. In [99], to detect emotional states from health-related posts, they combined lexicon-based features and the outputs of CNN network, which are then fed into LSTM network for further learning. In [100], feature representation from an intermediate layer of pre-trained network is extracted and then concatenated with some

hand-crafted features such as TF-IDF weighted word vector and lexicon-based features. Their model captures the hidden semantics and provides a more insightful understanding of emotional texts. In [101], to alleviate the problem caused by misspelling and out-of-vocabulary words, the NELEC model is proposed to make the model more robust. In their model, lexical features are captured and combined with neural features to boost performance.

Some works concentrated on enhancing word-level representation with external knowledge. Implicit emotion in textual expression can be inferred more easily via its enriched meaning by associated commonsense knowledge. In [102], they try to explore implicit emotion from external knowledge, including external commonsense knowledge from ConceptNet and emotion intensity information from NRC_VAD lexicon. With a context-aware affective graph attention mechanism, they dynamically retrieve the context-aware concepts and obtain concept-enriched word representation. A knowledge-enriched two-layer attention network is proposed in [103]. Their primary word-level attention is applied to input word and related terms obtained by searching WordNet and Distributed Thesaurus [104], generating word embedding enhanced by the knowledge graph. The secondary attention mechanism works on sentence-level for further context learning. Their system performs remarkably well on the benchmark datasets of SemEval 2017 Task 5. NTUA-SLP embedding is proposed in [105] for affective information learning. Start from a set of seed words with affective ratings from -1 to 1, the embedding for new words is estimated by considering both semantic similarity and ten affect-related features (valence, dominance, arousal, pleasantness, anger, sadness, fear, disgust, concreteness, familiarity). In [106], emotional features of Tweets are generated based on seven lexicons, which is realized by the filter in the AffectiveTweets: TweetToLexiconFeatureVector [107].

The pre-trained word embedding model contributes to generating emotional word representation. In [93], pre-trained DeepMoji is utilized for learning richer emotional context representation at the sentence level. In [108], Sentiment and Semantic-Based Emotion Detector (SS-BED) is proposed for emotion detection. The input words are embedded by two word embedding matrices: sentiment representations obtained by SSWE (Sentiment Specific Word Embedding), and the semantic word representations

obtained by pre-trained Word2Vec, Glove, and FastText [109]. These two sequences of word embedding are separately fed into two LSTM layers to learn sentiment and semantic features. This approach significantly outperforms traditional machine learning baselines, including SVM, Decision Trees, and Naive Bayes.

Rule-based representation is another common formalism of knowledge representation. In [110], rule-embedded neural networks (ReNN) are proposed to encode domain knowledge and commonsense information. The rule-based knowledge reduces computing complexity and contributes to training a better model with a smaller dataset. To perform knowledge base completion, ITransF is proposed in [111], which discovers hidden concepts of relations and transfers statistical strength by sharing concepts.

Linguistic patterns play an important role in emotion recognition. For example, negative words can shift the emotional tendency of the entire sentence. [112] proposed the linguistically regularized LSTMs to enhance sentence-level emotion representation. The effect of linguistic role (such as sentiment lexicons, negation words, and intensity words) is considered by proposed sentiment, negation, and intensity regularizers. Their model addressed the sentient shifting effect of the linguistic role. In [113], domain knowledge is combined with CNN neural network in the task of sentiment classification. Their domain knowledge mainly has three parts, sentiment words with intensity, linguistic patterns modifying the sentiment of emotional words, and the sentiment ontology managing semantic relationships between sentiment terms and domain concepts. Their model considered the effect of different terms on a specific domain. This model is upgraded in [114], more techniques of data augmentation based on external knowledge are introduced to enhance the word embedding, including negation-based augmentation and transfer learning.

2.4.3 Transfer Learning for Emotion Recognition

In an emotion recognition system, the amount of resources is a guarantee of satisfying performance. Collecting and annotating large amounts of data is time-consuming and expensive. The shortage of quality emotional data is always the most urgent problem. How to train an effective TER model with smaller datasets has attracted increasing

attention.

By transfer learning, the knowledge learned from the source domain is transferred to the target domain, realizing performance improvement. In this thesis, the target task refers to TER, and the source task can be any other related NLP tasks, including sentiment analysis and machine translation. The valid information learned from related tasks can be transferred into the target TER model. Transfer learning can alleviate the problems caused by scarce training data and speed up training, which improved the performance of deep learning models. Gupta [115] investigates the application of semi-supervised and transfer learning methods in low-resource sentiment classification tasks and demonstrates that transfer learning could significantly improve the performance compared with the simple supervised method.

There are various kinds of architecture proposed based on transfer learning. According to whether labeled data are available in the target and source domain, transfer learning can be categorized into inductive transfer learning, transductive transfer learning, and unsupervised transfer learning [116]. Among them, training data of target domain are labeled in inductive transfer learning while unlabeled in another two. Transfer learning is often utilized to transfer emotional and semantical information from source domain for information supplement. In this thesis, target task refers to TER. To ensure the final recognition accuracy, the training data of the TER task is usually annotated with emotion labels. Therefore, recent works about TER with transfer learning are mainly based on inductive transfer learning. Based on some different situations, related works can be mainly categorized into two sub-cases. The first is sequential transfer learning, by which source and target tasks are learned successively. Another is multi-task learning, by which source and target tasks are learned simultaneously.

(1) Sequential transfer learning

Sequential transfer learning is arguably the most frequently used transfer learning scenario in the NLP field. The process generally consists of two stages. The first is pre-training on the source tasks, such as emotion-related tasks or other natural language understanding related tasks. The second is the transfer phrase, in which the knowledge

learned in the source domain is transferred to the target TER task.

During the pre-training phase, many efforts are devoted to train language models with universal knowledge of the natural language. The aforementioned pre-trained language models can be transferred to almost all NLP tasks and have advanced multiple state-of-the-art performances. Some works are devoted to pre-training their model on emotion-related source tasks with sufficient training data, such as sentiment classification and emotion intensity regression.

In the transfer phase, the pre-trained model is transferred to emotion recognition task. There are mainly two ways to realize the transformation. One is taking the pre-trained model as a feature extractor, and all parameters in this model are frozen. Aforementioned pre-trained language models are mainly treated as feature extractors, such as BERT and DeepMoji. Pre-trained text representations obtained by pre-trained model are fed into emotion classification model for further training and final prediction. Another is fine-tuning the pre-trained network on target task. This operation is often accompanied by minor modifications to the network architecture, such as replacing the prediction layer. The most common fine-tuning approach is freezing most of the network and fine-tuning only the top layers (similar to feature extractor). There are also some other fine-tuning strategies. In gradual unfreezing, the layers are gradually unfrozen, starting from the new output layer going down to the first layer [117]. In single bottom-up unfreeze (also known as chaw-thaw) [93], the new output layer is trained with all the layers frozen, and then the models are further trained layer-wise from the bottom layer to the top layer. In the end, the entire model is trained with all the layers unfrozen. In single top-down unfreeze, fine-tuning operation is similar to chaw-thaw while the direction is from the top layer to bottom layer. Layer-wise training is able to adjust the individual patterns across the network with a reduced risk of overfitting. Above fine-tuning strategies are compared in the task of cross-lingual emotion classification [10]. Their experimental results show that the performance of gradual unfreezing and single top-down unfreeze are slightly better on fine-tuning phrase.

Sentiment analysis and intensity regression are often treated as source task in transfer learning-based TER. Compared with existing emotional corpus, sentiment datasets with polarity annotation are abundant. Although their content and labeling system are different,

both of them contain rich emotional characteristics. Such resources can be utilized by knowledge transfer and enhance the performance of TEC model. For example, an extension of transfer learning called sent2affect is proposed in [118]. They train a sentiment classification model on 100,000 tweets with polarity annotation and then transfer this model to TER by exchanging output layer and fine-tuning. Similar work is conducted in [119], inductive knowledge is transferred from SA tasks to TER task, and their experiments span categorical and dimensional emotion models. With transfer learning, the deep attentive RNNs model proposed in [105] ranked 1st in semeval-2018 task 1 'Multi-Label Emotion Classification'. They firstly obtain affective word embedding based on a small number of emotional seed words. With emotional word embedding, they pre-trained their model, BiLSTM with a deep self-attention mechanism, on the dataset of Semeval 2017 Task 4A. The final layer is replaced with a task-specific layer model and then further fine-tuned with two fine-tuning schemes. In [120], an ensemble of transfer learning techniques is proposed to predict the emotions of removed emotion trigger words. They utilize three different pre-trained models to initialize some specific layers of their networks, including a language model, a word embedding model, and a sentiment model. Then, they ensemble and fine-tuning these models in their dataset and have achieved competitive experimental results. In [121], a dual attention-based transfer learning approach is proposed for multi-label emotion classification. They respectively captured typical sentiment features and emotion-specific features with a shared attention layer, which are respectively fed to the task-specific layer by a dual attention mechanism. Experimental results show that their dual attention transfer architecture can bring consistent performance gains compared to several existing transfer learning approaches.

In textual emotion recognition, the existing emotional resources are mainly in English, while there are low resources in most European languages. The quality of emotion recognition model with few resources is often limited. Therefore, to those low-resource languages, it is advantageous to leverage the emotional information from resource-rich languages. To solve the low-resources problem in Hindi emotion detection, the emotional knowledge from English is transferred to Hindi by a deep transfer learning framework [10]. In their model, cross-lingual word-embeddings are trained by mapping each

monolingual word-embedding into a shared space with the transformation strategy of alignment matrices [122]. Therefore, relevant information can be captured through the shared space. Based on the cross-lingual word embedding, the deep learning model is pre-trained on emotional dataset in English, and then fine-tuning is done for final emotion prediction in Hindi.

Cross-domain transfer learning aimed to transfer knowledge across different domains, utilizing a small amount of labeled data from the target domain and abundant labeled data from a different source domain. The data distribution and labeled emotions from different domains have a significant impact on the performance of transfer learning. In [123], they try to transfer emotional knowledge from the source domain through joint learning with a domain classifier, promoting the performance of emotion classification through the sharing of domain-specific representation. In [14], transfer learning is utilized to address the task of cross-domain and cross-category emotion tagging for comments on online news. They achieve domain adaption through reweighting instances from the source domain by modeling the distribution difference. They also model the relationship between different sets of emotion categories from each domain, enabling project data from one domain into the label space of another domain. To improve the robustness of transfer learning methods to unseen data, Adversarial Discriminative Domain Generalization (ADDoG) is proposed in [124], aiming to generalize the representation of cross-corpus. ADDoG follows a 'meet in the middle' approach, iteratively move their dataset representations closer to one another, and improved the cross-dataset generalization

(2) Multi-task learning

Multi-Task Learning (MTL, also known as joint learning) is another form of inductive transfer. Various studies have shown that MTL dramatically improves the performance of TER systems than single-task learning. In MTL, target and source tasks are related and trained simultaneously [125]. Through underlying shared representations, sub-tasks promote and supply each other to learn more relevant information. Compared to the single-task framework, multi-task learning on related tasks can significantly reduce the risk of overfitting, contributing to better generalization performance and the improvement

on all sub-tasks.

MTL framework targets to enhance the generalization performance by leveraging the inter-relatedness of multiple tasks [126], [127]. Taking emotion-related tasks as auxiliary tasks is ideal for the MTL-based emotion recognition system, which indirectly realized emotional information integration from different resources [128]. Most works are conducted based on the structure consisting of shared networks and some task-specific layers. In the Emo2Vec model proposed in [91], to encode emotional semantics into word-level representations, six different emotion-related tasks are trained simultaneously. Their generalized emotion representation outperforms multiple existing affect-related representations, such as DeepMoji, but with much smaller training data. In [129], a two-stage multi-task learning structure is proposed to complement the feature representation in the dimensional model with the knowledge transferred from the discrete model, thereby establishing a relationship between discrete and dimensional emotion. Rather than parameter sharing, the works in [130] realize label transformation from sentiment label to emotion label by joint learning and have improved the performance of emotion classification.

To address time-consuming limitation in emotion annotation, multi-task active learning for regression (ALR) is proposed in [131]. The most beneficial samples are selected in their model, benefiting the emotion estimation in three dimensions (valence, arousal, and dominance) simultaneously. In [100], a multi-task ensemble learning framework is proposed for several tasks related to category/dimensional emotion, sentiment, and intensity. They firstly pre-trained three individual networks by multi-task learning and obtained three task-aware deep representations. These representations are combined with other hand-crafted features and then fed into a multi-task ensemble model for further learning. This multi-task ensemble framework helps in achieving generalization and contributes to superior results.

The difference of label distributions between the training and test sets is considered in [132]. They find that most of the errors are raised by recognizing the ‘Others’ category. They think their performance could be better if firstly conduct the binary classification ‘Others’ versus ‘Not-Others’. This problem is considered in [133]. They achieve better performance by utilizing a multi-learning network and can better detect emotions from

the ‘Others’ class.

Sub-tasks in MTL are highly correlated, and label relationships among all tasks provide useful information. To address emotion ambiguity in the textual expression, a multi-task CNN model is proposed in [134], which learns emotion label distribution and emotion classification simultaneously. These two tasks boost each other and thereby generate a robust text representation. In [135], An Adversarial Attention Network (AAN) is proposed to conduct adversarial learning between each pair of emotional dimensions. They conduct multidimensional emotion regression tasks by multi-task learning, and they perform well on the EMOBANK corpus. In [136], a joint label space is induced to enable multi-task learning from both labeled and unlabeled data. They exploit the relationships between different labels from all tasks according to a label transfer network, demonstrating that potential synergies between label spaces can be leveraged for label transformation.

Chapter 3

Background Knowledge Enhanced Classification Network

3.1 Introduction

In contemporary society, textual data are continuously increasing and have become the commonly used information carriers [137]. As a kind of efficient information retrieval and data mining technology, text classification aims to associate the given document and one or more categories according to the features representation. It has been widely used in many fields, including sentiment analysis [138], [139], stock analysis [140], automatic news grouping, and so on.

Training corpora with enough data and accurate label annotation always contribute to classification tasks. However, most of the existing corpora are imbalanced in two aspects [141]. The first is the data imbalance between categories, which means that the amount of data in different categories is considerably different [142]. In classification tasks, data distribution is usually not considered, and traditional algorithms always tend their performance to the categories with more data. In the worst case, categories with fewer data may be regarded as outliers [143], [144]. The second refers to the feature imbalance within the category. For a broader category, it is challenging to include all sub-categories in collected training datasets, especially for those low-resource categories. Therefore, it is common for a broader category that training and testing data come from different sub-categories, and there are huge differences in the distribution of feature words. Such as the

sub-categories ‘Telecommunication and networking’ and ‘Programming languages’ of the broader category ‘Computer science’. Therefore, the classifier is easy to give biased predictions.

There is a great deal of textual data in this information age that existed online with abundant information. Although most of them are unstructured data without category labels, this information can indirectly expand the data coverage and increase the scale of training data. Motivated by this cognition, a background knowledge based multi-stream neural network is proposed to address the challenges of imbalanced data distribution. The background knowledge is extracted from an external corpus, which covers the data from almost all fields, and it serves as prior knowledge to complement the deficiency of the training data. In this chapter, background knowledge mainly consists of two parts: 1) categorical keywords, containing distinguishable category information; and 2) co-occurred words, which frequently co-occurred with keywords.

To better incorporate background knowledge into the feature extraction process, we propose a multi-stream neural network with different fusion strategies. This network is mainly composed of the basal stream and background knowledge based stream. The basal stream takes the original word sequence as input, retaining the semantic information of the original sequence. The background knowledge based streams take keywords and co-occurred words as inputs, realizing information supplement and reinforcement for the basal stream. Each stream is trained with GRU Network. Different fusion strategies are proposed to integrate the features extracted from different streams. Compared with basal model, the proposed method performs well in both Chinese and English corpus. The macro F1 score of Reuters 21578-R8 is up to 95.02%, which obtained 10.16% improvement, and the macro F1 score of Fudan University corpus is up to 85.03%, which obtained 8.75% improvement.

Our work makes the following contributions:

(1) Background extracted from external corpora is incorporated into the classification network, which indirectly enlarges the training corpus and makes up for the imbalance data distribution.

(2) The background knowledge-based stream is proposed to extract features based on the distribution information of keywords and co-occurred words. This information is

acquired from the external corpus, contributing to avoiding feature imbalance within a particular category.

(3) A multi-stream neural network with different fusion strategies is proposed for information integration, realizing information supplements and reinforcements.

The remainder of this chapter is organized as follows. Section 3.2 introduces the proposed background knowledge-based multi-stream neural network. Section 3.3 includes the experiments and discussions on both Chinese and English corpus. Finally, Section 3.4 summarizes our works and outlines the direction of future work.

3.2 Background Knowledge-based Multi-Stream Neural Network

This subsection detailed describe the overall methodology of background knowledge acquisition and multi-stream neural networks. The background knowledge is extracted from an external corpus and is composed of keywords and co-occurred words. To better incorporate background knowledge, a multi-stream neural network is proposed to extract in-depth features. Different fusion strategies: early-fusion and later-fusion are employed in the feature fusion layer to integrate features from each single-stream.

3.2.1 Acquisition of Background Knowledge

It is easy for humans to immediately determine the category of a document by looking at some specific words based on their abundant background knowledge stored in the brain. Therefore, the assumption is reasonable that incorporating background knowledge into the classification model will contribute to more accurate predictions.

In this chapter, background knowledge is incorporated into the classification network to supplement training data indirectly. The background knowledge is extracted from external corpora based on the following assumption: in the natural language, if two words w_1 and w_2 often appear together in the same unit window (such as paragraph and sentence), it is believed that they have some particular relationship, and the higher frequency of word co-occurrence, the closer relationship they have.

While word w_1 appears, the probability of word w_2 occurring is as follows:

$$R(w_2|w_1) = \frac{f(w_1,w_2)}{f(w_1)}, \tag{3.1}$$

Where $f(w_1, w_2)$ represents the counts of word $w_1$and w_2 appeared together, and $f(w_1)$ represents the counts of word w_1 appeared.

Therefore, a reasonable assumption is that if there are some keywords of a particular category, their co-occurred words with high frequency also contain categorical information. In this chapter, background knowledge is composed of a set of keywords and their co-occurred words. The keywords are extracted from the annotated training corpus. The co-occurred words refer to those words that appear together with keywords within a particular word distance. The external corpora are employed to search for the co-occurred words in a statistical unit (such as the sentence). Only those high frequency co-occurred words are remained in the co-occurred words set. The flow of background knowledge acquisition is shown in Figure 3.1.

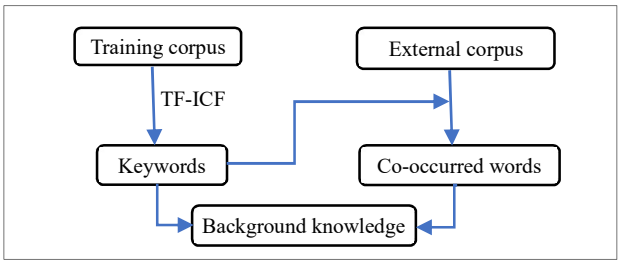

Figure 3.1 Extraction of background knowledge.

3.2.1.1 Keywords Acquisition

The training corpus has been classified, labeled with category, and pre-processed to filter the useless words, such as stop words. TF-ICF method [145] is applied to obtain the categorical keywords in training datasets.

$$tficf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log \frac{|c|}{|\{j: w_i \in c_j\}|} \tag{3.2}$$

In which $\frac{n_{i,j}}{\sum_k n_{k,j}}$ is the term frequency of word w_i occurred in category c_j, $|c|$ is the

number of categories, and $\left|\{j: w_i \in c_j\}\right|$ is the number of categories containing w_i.

For a particular category, a high $tficf$ value means a high term frequency in this category and a low category frequency in the whole corpus. Finally, those words with high $tficf$ weights composed the keywords set: $keywords = \{k_1, k_2, \dots\}$. These keywords are the basis of the acquisition of co-occurred word set.

3.2.1.2 Co-Occurred Words Acquisition

For a particular category, it can be further subdivided into many sub-categories, but it is difficult for the training data to cover all these sub-categories, especially for those low-resource categories. Therefore, the training and testing data may come from different subcategories and have different feature distribution, thereby affecting classification performance.

If two words often appear together, we can think they belong to the same category. The co-occurred words of keywords with high frequency carry discriminative category information as well. This way, the external corpora from various fields are utilized to obtain co-occurred words, aiming to cover more comprehensive data and realize information supplementation.

The co-occurred words and co-occurrence counts with keywords are obtained by scanning each sentence in the external corpora to obtain the word co-occurrence information. In this process, only the co-occurred words within a certain distance are taken into consideration. For each keyword k_i, a co-occurrence matrix is obtained, which column index is co-occurred word, and the value is co-occurred frequency $R(k_i|w_j)$.

$$R(k_i|w_j) = \frac{f(k_i, w_j)}{f(k_i)} \tag{3.3}$$

in which $f(k_i, w_j)$ is the co-occurrence count of k_i and w_j in the external corpus, and $f(k_i)$ is the occurrence count of k_i.

Finally, the co-occurred words of each keyword with high frequency are obtained, and after duplication remove, the set of co-occurred words are obtained: $Co\_words = \{cow_1, cow_2 \dots\}$.

3.2.2 Multi-Stream Neural Network

The multi-stream neural network performs well in features fusion, and is often applied in spatial and temporal networks in action recognition of videos [146]. To better incorporate background knowledge-based information, a multi-stream neural network is proposed, as shown in Figure 3.2, and different fusion strategies are used to extract comprehensive features.

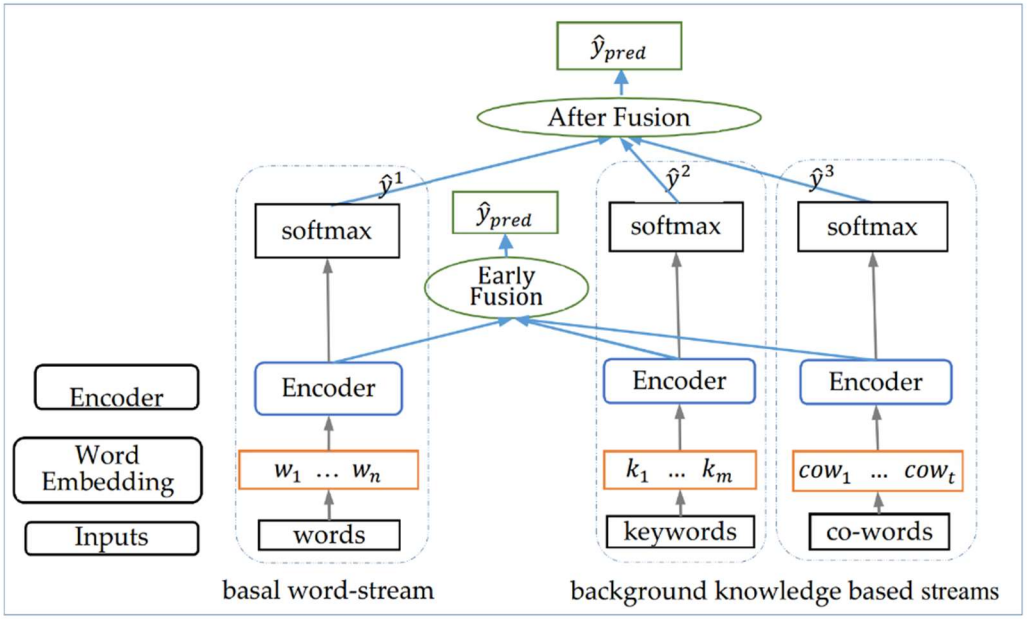

Figure 3.2 Multi-stream model based on background knowledge

The proposed multi-stream neural network consists of basal word-stream and background knowledge-based aid streams. It mainly has five parts: input layer, word embedding layer, encoder layer, model training, and fusion layer. Different feature sequences (detailed in Section 3.2.1) are fed into each stream separately. Each stream is trained on a mini-batch with Adam optimizer independently. To further combine the features extracted from different streams, different fusion strategies (detailed in Section 3.2.3) are employed for final prediction.

3.2.2.1 Input Layer

The input of basal word-stream is original texts, while another is background knowledge-based feature words. For each text s, the inputs of each stream in network

divided into three parts: words, keywords, and co-occurred words. Before the input layer, all words in texts s, keywords set $keywords$ and co-occurred words set Co_words are firstly transformed into real-valued word tokens by looking up in a pre-defined word tokens dictionary.

For each text $s = \{w_1, w_2, \dots w_n\}$, the different inputs in multi-stream are defined as follows:

(1) Word-stream: words = $\{w_1, w_2, \dots, w_n\}$, including all words in texts.

(2) Key-stream: for each word w_i in s, if w_i in keywords set $keywords$, append w_i to the keyword sequences. Finally, the input of keyword-stream obtained: $key = \{k_1, k_2, \dots, k_m\}$.

(3) Cow-stream: for each word w_i in s, if w_i in co-occurred words set Co_words, append w_i to the co-occurred word sequences. Finally, the input of co-occurred stream obtained: $cow = \{cow_1, cow_2 \dots cow_t\}$.

3.2.2.2 Word Embedding Layer

Word embedding is learned from massive unstructured textual data and is widely adopted in NLP tasks. By representing each word as a fixed-length vector, these embedding can group semantically similar words and implicitly encode rich linguistic regularities and patterns [80].

All words of corpus composed the word embedding matrix: $L \in R^{V*D}$, in which V is the vocabulary size and D is the dimension of word embedding. Each word in the input is represented as $w_i \in R^{1*D}$.

3.2.2.3 Encoding Layer

The encoder layer with RNN is connected to extract the high-order textual and semantic features. GRU (Gated recurrent units) cells are employed in RNN, and the parameters are not shared with each stream. GRU is a gating mechanism and performs well in in-depth feature extraction [147].

The final state of GRU is output as feature sequence: $H = \{h_1, h_2 \dots h_N\}$, in which N is the dimension of hidden layer in each stream.

The softmax function is stacked to the encoder to calculate the probability distribution,

and the output is $P = \{p_1, p_2 \dots p_n\}$, in which n is the number of categories, and p_i is the predicted probability of input belonging to the corresponding category i. The final predicted tag $\hat{y} = argmax(P)$.

3.2.2.4 Model Training

Each single-stream is trained respectively without parameter sharing. End-to-end backpropagation is employed in training, and the loss function is defined as follows:

$$cross\ \ entropy = -\frac{1}{n}\sum y ln\hat{y} + (1-y)\ln(1-\hat{y}), \tag{3.4}$$

The training of the model is to minimize the cross-entropy in each stream, and AdamOptimizer is used during training.

3.2.3 Fusion Strategy

Different fusion strategies are employed to integrate the information of corpus self and background knowledge. After the optimal parameters being trained in each stream, early-fusion and two after fusion strategies: average pooling and soft voting, are employed to obtain the comprehensive text representation.

(1) Early-Fusion

In early-fusion, the features extracted from each stream are concatenated together and then input to softmax for final prediction. The feature vector after early-fusion is:

$$H^{early} = [H^{word}, H^{key}, H^{cow}], \tag{3.5}$$

(2) After-Fusion

Probability distribution from each stream indicates the corresponding categorical predictions of different inputs. Therefore, appropriate weights can be assigned to each prediction to generate a more comprehensive probability distribution for final predictions. This way, two after fusion strategies: average pooling and soft voting, are proposed.

In average-pooling, a uniform weight is assigned to each stream, and the final probability distribution is:

$$P^{Avg} = w_1 * P^{word} + w_2 * P^{key} + w_3 * P^{cow} \tag{3.6}$$

in which P^{word}, P^{key}, and P^{cow} are predicted probability distribution from three streams. w_i is the weight parameter and $\sum w_i = 1$. The final predicted tag: $\hat{y}_{pred} = argmax(P^{Avg})$.

In the process of average pooling, a uniform weight w_i is distributed to each stream. It means that for a certain single-stream, the weight w_i is assigned on each category's final prediction. However, in the actual case, for a certain single-stream, the features extracted for a particular category may be more discriminatory than others, and thereby its prediction may be more accurate. This way, higher weight should be given to these categories while lower weight should be given to those inaccurate estimations. Therefore, another strategy of after-fusion: soft-voting, is proposed.

In soft-voting, a fully connected neural network is trained after softmax layer, as shown in Figure 3.2, to balance the weakness among each stream. The input is the concatenated probability distributions of all streams, and the final probability distribution is:

$$P^{soft} = \mathrm{w} * [P^{word}, P^{key}, P^{cow}] \tag{3.7}$$

In which $w \in R^{3n*n}$, and n is the number of categories. The final predicted tag: $\hat{y}_{pred} = argmax(P^{soft})$.

3.3 Experiments and Discussion

We conduct experiments in Chinese corpus: Fudan university corpus and an English corpus: Reuters-21578 R8. In order to evaluate the performance of the proposed multi-stream model based on background knowledge, multiple comparison experiments are conducted to:

(1) Investigate the performance of background knowledge based multi-stream neural network on text classification task;

(2) Investigate the contribution of background knowledge incorporation under different fusion strategies, especially the contribution to those categories with fewer data;

(3) Investigate the generalization of the proposed multi-stream neural network in different corpus and different language environments.

3.3.1 Dataset and Preprocessing

In our experiments, all datasets have been preprocessed. Referred Chinese corpus is preprocessed by word segmentation by Stanford-Segmenter, part-of-speech tagging by Stanford-Postagger, non-Chinese words removal, non-nouns removal, and stop words removal. The English corpus is preprocessed by stemming and stop words removal.

While reading an article, humans often can accurately judge the related area after reading some paragraphs instead of the whole. Especially for textual data, classification can be done after obtaining the category information. For the above common sense and to reduce the computing cost during the experiments, a fixed text length is set according to the length distribution of the corpus. If the length of the original text is higher than the fixed value, only the previous words are retained. Otherwise, 0-padding is employed.

(1) Training Corpus

The Chinese corpus: Fudan University text classification corpus and the English corpus: Reuters-21578 are used as training corpus to test the performance of the proposed method. The data distribution is shown in Figure 3.3 and Figure 3.4.

Fudan University text classification corpus (hereinafter referred to as Fudan corpus) is provided by the natural language processing group of international database in the computer information and technology department of Fudan University.

Reuters-21578, a collection of documents that appeared on Reuters newswire in 1987. This English corpus contains 90 classes of news documents. Reuters-21578 R8 (hereinafter referred to as Reuter R8) selects eight classes from Reuters-21578.

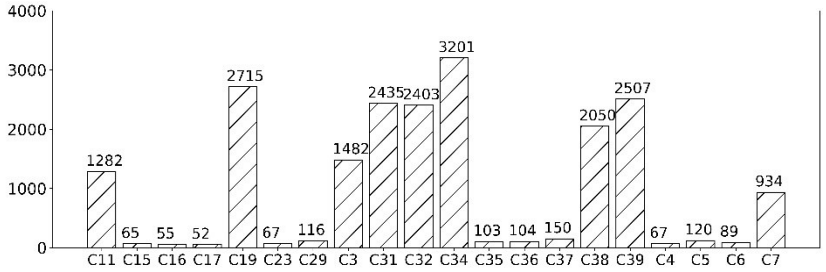

Figure 3.3 Data distribution of each class in Fudan corpus.

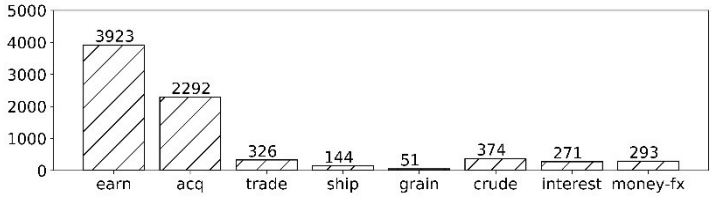

Figure 3.4 Data distribution of each class in Reuter-21578 R8.

(2) Background Corpus

There are two datasets severed as external corpus to extract background knowledge. One is Chinese corpus: People's daily news (http://paper.people.com.cn), which contains about 61 million sentences, and the average length is about eight characters. Another is English corpus: Reuters Corpus, which contains about 806 thousand texts, and the average length is 109 words.

3.3.2 Experimental Setup

To investigate the contribution of background knowledge incorporation under different fusion strategies, some comparison experiments are conducted in both single-stream network and multi-stream network. Especially, the single word-stream is employed as the baseline, which takes the original word sequence as input and takes GRU as encoder. Some abbreviations used in this section are shown in Table 3.1.

In our experiments, inputs of each stream are all uniformed to the same length by 0-padding. The max length setting is shown in Table 3.2. The dimension of word level embedding is set to 128. To extract features, RNN with GRU cell is employed as encoder in every single stream, and the hidden layer is set to 256. We optimize every single stream with Adam algorithm in mini-batch, and the batch size is 256. The dropout was 0.5, and

the learning rate is 0.002.

Table 3.1 Explanatory note of abbreviations in the experiments.

Abbreviations	Note
W	word-stream, input is $words$.
Key	key-stream, input is key.
Cow	cow-stream, input is cow.
KeyCow	keycow-stream, input is concatenate of key and cow.
W + Key	Fusion of word-stream and key-stream
W + Cow	Fusion of word-stream and cow-stream
W + KeyCow	Fusion of word-stream and keycow-stream
W +Key + Cow	Fusion of word-stream, key-stream and cow-stream

Table 3.2 Max length setting of inputs.

Parameters	Reuter R8	Fudan Corpus
Max length of word_stream (n)	128	300
Max length of key_stream (m)	45	130
Max length of cow_stream (t)	70	222

3.3.3 Results and Discussion

Results of Reuter-21578 R8 are shown in Table 3.3. From the results of single-streams based on background knowledge (key-stream, cow-stream, and keycow-stream), it can be seen that their macro precision, recall, and F1 score have been significantly improved while the accuracy is not much different compared with the basal word-stream. In the multi-stream network, the overall macro values increased significantly while the accuracy also improved, whether in early-fusion or after-fusion strategy. The highest three results under different evaluation indicators are bolded in the table. Under comprehensive consideration, the best one is obtained in three-stream network with average pooling (P = 95.28, R = 94.75, F1 = 95.02), which is superior than baseline (word-stream, P = 84.71, R = 85.37, F1 = 85.04). The improved classification results suggest that incorporating background knowledge has enriched the text representation to a great extent. Therefore, we can infer that the background knowledge can make up for the insufficient information

Table 3.3 Results of Reuters-21578 R8.

		Unit: %	Macro			Acc
			P	R	F1	
Single stream:		***W (Baseline)***	***84.71***	***85.37***	***85.04***	***96.12***
		Key	90.41	92.16	91.28	96.30
		Cow	91.35	89.84	90.59	95.39
		KeyCow	91.44	88.81	90.11	95.75
Early fusion:		W + Key	90.76	93.28	92.00	**97.44**
		W + Co	92.74	91.34	92.03	96.57
		W + KeyCow	**93.56**	91.03	92.28	96.67
		W +Key + Cow	93.09	92.39	92.74	96.85
After fusion:	Average pooling	W + Key	93.13	**93.85**	**93.49**	97.30
		W + Co	91.54	90.85	91.19	96.76
		W + KeyCow	90.24	89.63	89.94	96.67
		W + Key +Cow	**95.28**	**94.75**	**95.02**	**97.67**
	Soft voting	W + Key	91.75	92.52	92.14	97.08
		W + Co	91.00	89.89	90.44	96.07
		W + KeyCow	89.12	87.90	88.50	96.35
		W + Key +Cow	**95.15**	**94.30**	**94.72**	**97.67**

Table 3.4 Results of Fudan Corpus.

		Unit: %	Macro			Acc
			P	R	F1	
Single stream:		***W (Baseline)***	***76.70***	***76.41***	***76.55***	***95.43***
		Key	78.16	68.12	72.79	90.82
		Cow	83.27	78.15	80.63	95.22
		KeyCow	77.84	74.05	75.90	94.52
Early fusion:		W + Key	83.82	79.63	81.67	96.15
		W + Co	84.44	**81.73**	83.06	**96.79**
		W + KeyCow	83.02	80.80	81.90	96.53
		W +Key + Cow	**88.69**	**81.65**	**85.03**	**96.89**
After fusion:	Average pooling	W + Key	84.18	78.99	81.50	95.98
		W + Co	84.69	80.96	82.78	96.48
		W + KeyCow	83.90	78.66	81.19	96.13
		W + Key +Cow	85.26	**81.94**	**83.57**	**96.67**
	Soft voting	W + Key	83.60	76.59	79.94	95.64
		W + Co	**85.64**	79.40	82.40	96.30
		W + KeyCow	83.21	78.35	80.70	96.02
		W + Key +Cow	**88.66**	80.28	**84.26**	96.41

of training data and make up for the deep feature extraction of those data-sparse categories.

The effectiveness of background knowledge incorporation in multi-stream neural network is also verified in Fudan corpus. The results are shown in Table 3.4, and the best one is acquired in three-stream network with early fusion: P = 88.69, R = 81.65, F1 =

85.03 (while P = 76.70, R = 76.41, F1 = 76.55 in baseline), demonstrating that the background knowledge with multi-stream networks contributes a lot to this task.

The comparison results between the above optimal and basal models show that the macro indicators have significantly improved while the overall accuracy growth is slight. The reason may be that in these imbalanced corpora, there are some categories with less data, and their classification results have a slight effect on the overall accuracy results because of the data distribution. However, during the model training, accuracy is often used as the evaluation indicator, and the categories with fewer data are ignored. Therefore, macro evaluation can reflect the overall classification results more comprehensively for imbalance corpus.

In multi-stream network, the overall classification results are improved a lot after feature fusion. As the supplement, background knowledge-based features make up for the problem caused by data imbalance in the basal word-stream network. The imbalance refers to two aspects. The first is the feature imbalance in a particular category. For example, because of the limitation of data coverage, testing data and training data may come from different sub-categories of a broader category and contain different feature words. Background knowledge contains almost all sub-categories, serving as a bridge to connect training data and testing data. The second is the data imbalance among categories. There are some categories with fewer data and result in fewer features extracted. The incorporation of background knowledge can significantly alleviate the limitation caused by imbalanced data distribution.

To investigate the contributions of background knowledge on different categories, the optimal three groups of experimental results are compared with the basal word-stream model. The results of each category are shown in Figure 3.5 and Figure 3.6, respectively. The horizontal axis refers to the categories, the bar charts refer to the classification results, and the line graph refers to the number of training texts in the corresponding category.

According to the results of Reuters-21578 R8, as shown in Figure 3.5, the categories with fewer training data, like ‘ship’ with 36 training texts and ‘grain’ with 10 training texts, improved a large amount in macro precision, recall, and F1 score, while the

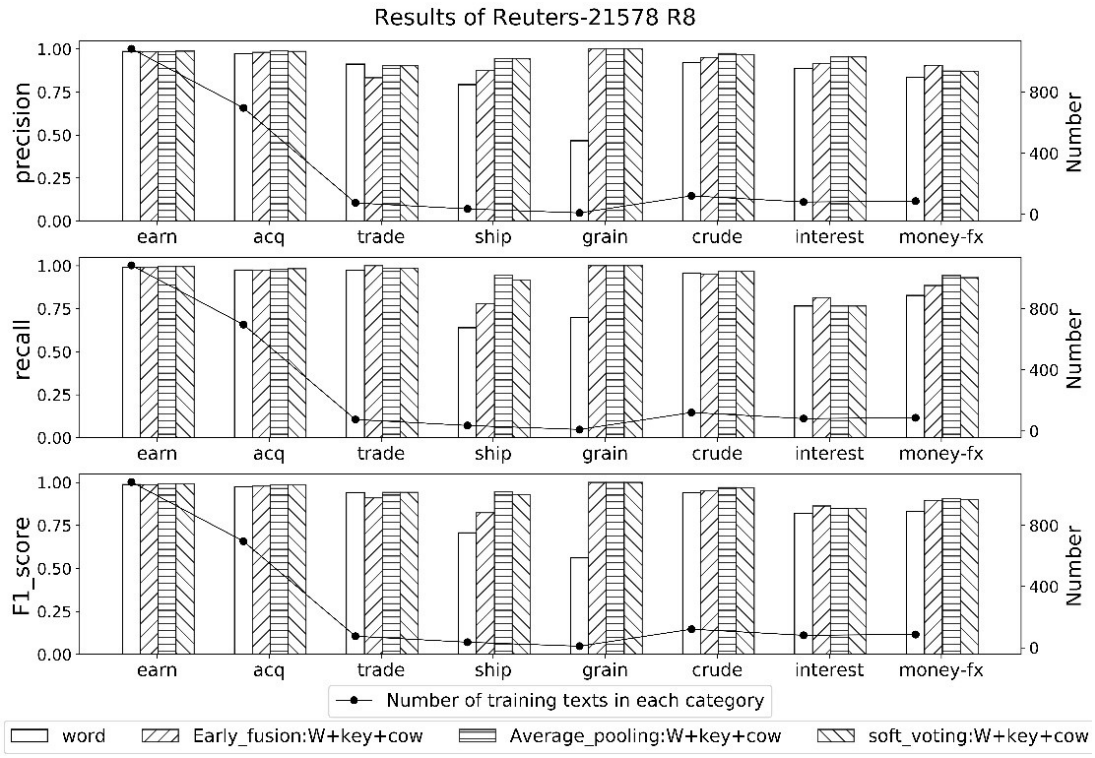

Figure 3.5 Results of Reuter-21578 R8.

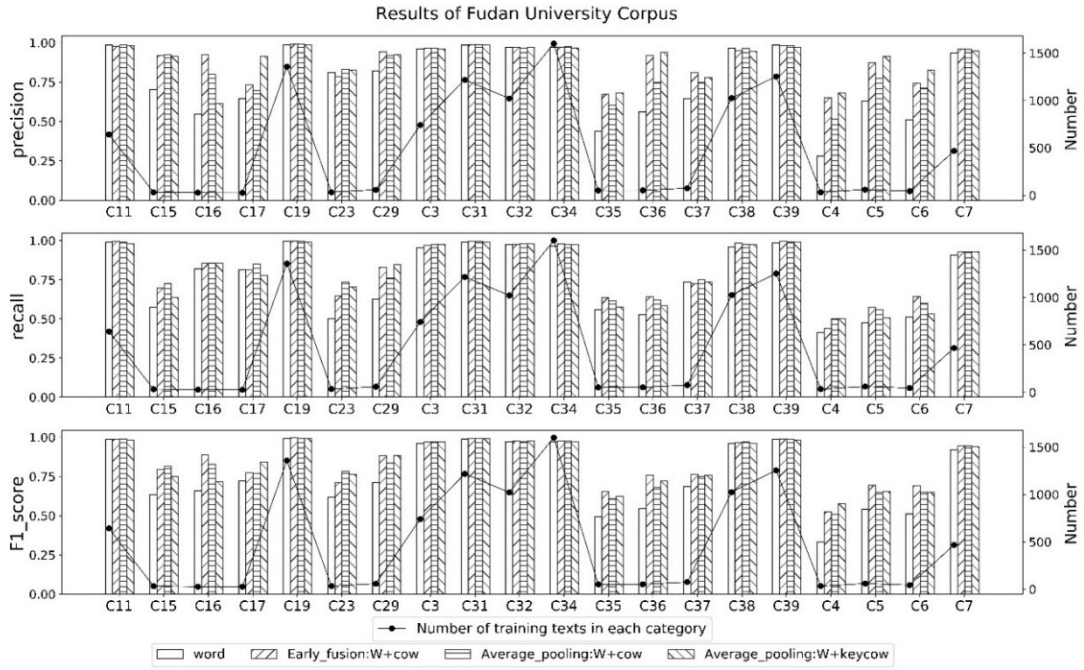

Figure 3.6 Results of Fudan University Corpus.

categories with relatively much training data, like ‘earn’ and ‘acq’, improved not very obviously. These results also suggest the previous viewpoint about imbalance corpus: the

incorporation of background knowledge can conspicuously make up for the insufficient information of some categories with fewer data, contributing to the overall performance.

The results of Fudan Corpus also verify the above viewpoints, as shown in Figure 3.6. In the categories with more than 600 training texts, such as ‘C11’ and ’C19’, the improvements are slight, while in other low-data categories, the improvement is relatively apparent.

3.4 Summary

This chapter focus on the research of incorporating background knowledge to make up for the limitation of imbalanced training data. To better fuse background knowledge-based features into basal model, a multi-stream neural network with different fusion strategies was proposed.

The experimental results obtained from different corpus showed that, compared with traditional RNN based text classification model, the proposed method performed better under different evaluation indicators. The macro F1 score improved up to 10.16% in Reuters-21578-R8, and 8.75% in Fudan corpus. According to the comparison results, the following conclusion can be drawn: as the supplement, the background knowledge can make up for the information neglected or absented in the basal text classification network, especially for imbalance corpus.

In the future, the proposed work can be extended by extracting more beneficial background knowledge from more comprehensive external corpora provided by this big data era. Furthermore, some state-of-the-art models can be utilized as encoders, and different feature fusion strategies can be further researched to achieve more comprehensive and in-depth feature information.

Chapter 4

Label Embedding for Contextual Emotion Recognition

4.1 Introduction

As an essential element in human nature, emotions have been widely studied in psychology [3]. Emotion recognition involves identifying detailed emotional states, which mainly refer to a wide range of mental states, such as happiness, anger, fear, etc. Textual emotion recognition (TER) is a fine-grained sentiment analysis, aiming to classify a textual expression into one or several emotion classes depending on the underlying emotion theories employed. In recent decades, TER tasks have gained considerable interest in the research community.

Recent researches about TER mainly conducted on sentence-level, which aimed to recognize subtle emotions based on word and concept-based features extracted from the given sentence. However, emotional expression is complicated, and the same sentence could present different emotions in different contexts. In the absence of contextual information, even humans cannot give confident emotional judgments. Therefore, it is necessary to utilize contextual information for sentence-level emotion recognition.

Given a sentence, its context generally refers to the sentences that appear around it. For example, given a sentence from a blog, its context refers to those sentences that appeared around the current sentence. Given an utterance from a dialogue, its context generally refers to the preceding occurred utterances. Such contextual information has been

explored in some preceding works, such as HANs [148], TreeLSTM [149]，and CLSTM [150]. Under the different circumstance, contextual sentences and current sentence have different contributions to final prediction, and attention mechanism based network is widely utilized to address this problem. Inspired by HANs (Hierarchical Attention Networks), we explore effective encoders for sentence-level encoding and contextual-level encoding to generate more accurate emotion representation expressed in the given sentence.

Emotional expression is very complicated. Some emotions often co-occurred with each other, such as the emotion pair of 'Joy' and 'Love', while some are opposed and rarely appear together, such as 'Joy' and 'Anxiety'. Emotion correlation has always been a significant factor in emotion recognition tasks. To accurately recognize emotions, it is necessary to fully consider the correlations between each emotion.

Above all, this chapter explores a hierarchical model to learn contextual representations, which encodes the emotional information of a given sentence based on its context. Besides, to realize emotion correlation learning, we trained a label embedding matrix by joint learning, which beneficial to emotion correlation-based emotion prediction. The contributions are summarized below:

(1) This chapter proposes a hierarchical model to learn contextual representations for sentence-level emotion recognition. Pre-trained language model BERT is taken as the sentence-level encoder, and attention-based bidirectional LSTM is taken as the context-level encoder, aiming to learn the emotional information of the given sentence based on its context.

(2) To give emotion correlation-based prediction, the label embedding matrix is learning by joint learning. Emotion correlation is obtained by calculating the similarity features between sentence representation and each label embedding, contributing to the final prediction.

(3) To guarantee the effectiveness of both emotion prediction and label embedding, the proposed network is trained by an assemble training objective. The experimental results indicate that proposed approach has a satisfying performance in TER task.

4.2 Methodology

4.2.1 Problem Definition

Assuming that we have N training samples X along with their contextual information C. Each sample $x \in X$ is often a sentence from the hierarchical text, such as dialogue or blog, and the contextual information $c = \{c_1, c_2, \dots c_n\} \in C$ often means the preceding n sentences appeared before x. Each sample x is annotated with K emotional labels: $\{e_1, \dots e_k, \dots e_K\}$, denoted as a one-hot vector $y = \{y_1, \dots y_k, \dots y_K\} \in \mathbb{R}^{1\times K}$, in which $y_k = 1$ is x contains emotion e_k otherwise $y_k = 0$.

For each sample $x \in X$, a multi-label emotion recognition model F is trained to transform x into predicted distributions $p = \{p_1, \dots p_k, \dots p_K\}$ based on its contextual information c, and then give a final prediction of all possible emotion labels. The function *F* is denoted as:

$$F(x, c) = \{p_1, \dots p_k, \dots p_K\} \tag{4.1}$$

4.2.2 Hierarchical Network with Label Embedding

To model a sentence x along with its contextual information $c = \{c_1, c_2, \dots c_n\}$, the simplest way is to utilize flatten context modeling, by which x and contextual sentence c are concatenated as $x' = \{c_1, c_2, \dots c_n, x\}$ and all tokens in x' are flattened into a word sequence. However, emotions flow naturally in each sentence, such flatten processing not only makes the sequence of words too long but also ignored the time step and destroyed the hierarchical structural information. The sequential nature of context is non-negligible, and such hierarchical information could contribute to the emotion prediction better.

Motivated by Hierarchical Attention Networks (HANs), we focus on hierarchical context modeling. Each sentence in $x' = \{c_1, c_2, \dots c_n, x, \}$ is first encoded into sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \dots h^s_{cn}, h^s_x\}$ by a sentence-level encoder En_s, and then contextual information is further encoded by a hierarchy context encoder.

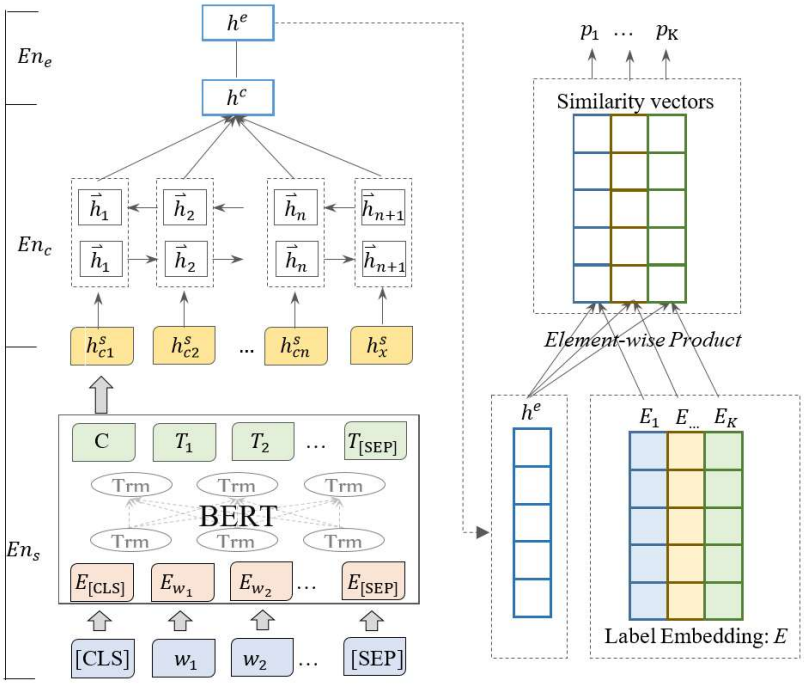

Figure 4.1 The framework of hierarchical network with label embedding.

4.2.2.1 Sentence-level modeling

At sentence-level, for each sentence $s = \{w_1, w_2, \ldots\}$ in $x' = \{c_1, c_2, \ldots c_n, x\}$, the function En_s encodes s into sentence-level representations h_s, denoted as:

$$h_s = En_s(s) \tag{4.2}$$

Inspired by the pre-trained language model and transfer learning techniques, pre-trained BERT model [84] is taken as sentence-level encoder En_s in this chapter. BERT stands for Bidirectional Encoder Representations from Transformers, and it is designed to pre-train deep bidirectional representations from unlabeled textual data by jointly conditioning on both left and right context in all layers. It remedies the limitation of insufficient training corpora and contributes to syntactic and semantic sentence representation.

In this way, for the sentences in $x' = \{c_1, c_2, \ldots c_n, x\}$, sentence-level representation $h^s = \{h^s_1, h^s_{c2}, \ldots h^s_{cn}, h^s_x\}$ is generated by pre-trained BERT model.

4.2.2.2 Contextual-level Modeling

In contextual-level, the function En_c encodes the sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \dots h^s_{cn}, h^s_x\}$ into a context-level representation h^c, which is denoted as:

$$h^c = En_c(h^s) \tag{4.3}$$

In the proposed model, the function En_c mainly consisting of two-layer networks: BiGRU (Bidirectional Gated Recurrent Neural Networks) and attention network.

BiGRU aims to deal with the sequential information of contexts. Take sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \dots h^s_{cn}, h^s_x\}$ as input, the output of the hidden state of BiGRU in each step is $h_i = [\vec{h}_i : \overleftarrow{h}_i]$, in which $\vec{h}_i$ and $\overleftarrow{h}_i$ are the output of hidden states from forward and backward directions respectively.

The attention network aims to make the network pay more attention to essential contexts. The attention mechanism considers the contributions of previous occurred contextual sentences $c_i \in c$ to the prediction of current sentence x. More attention weight will be assigned to related contexts. Attention weight a_i and weighted emotional feature vector h^c are defined as follows:

$$h^c = \sum_i a_i h_i \tag{4.4}$$

$$a_i = \frac{exp(e_i)}{\sum_{k=1}^{n} exp(e_k)} \tag{4.5}$$

$$e_i = w_2^T[\sigma(w_1^T \cdot h_i + b_1)] + b_2 \tag{4.6}$$

in which σ indicates the sigmoid activation function, w_1, b_1, w_2, b_2 indicate the model parameters.

In a typical contextual network, h^c is fed into the classifier for final prediction. The classifier typically consists of a linear transformation. It is followed by a sigmoid operation to normalize the outputs so that each element in the scale of [0,1]. A multi-label

neural network is typically trained by minimizing the Binary Cross Entropy (BCE) between the true labels distribution Y and predicted distribution P as the following:

$$BCE(P,Y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \cdot \log(p_{ik}) + (1 - y_{ik}) \cdot \log(1 - p_{ik}) \tag{4.7}$$

in which p_{ik} is the predicted probability of emotion e_k in ith sample, and y_{ik} is the true label, $y_{ik} \in [0,1]$.

Above mentioned typically network is intuitive and simple, and widely utilized in multi-label classification problems. However, emotion recognition is a more complex problem. This typical network with BCE loss function can be less effective and poor generalization due to its ignorance of label correlations. To capture label correlations, a joint learning label embedding network is proposed, which is detailed in Section 4.2.2.3.

4.2.2.3 Label embedding network

The label embedding is supposed to represent the semantics and relations between emotion labels. The embedding is denoted by

$$E = \{E_1, \dots E_k \dots E_K\} \in \mathbb{R}^{K \times d} \tag{4.8}$$

where K is the number of emotion labels, and d is the dimension of label embedding.

To make label embedding contribute to the emotion recognition network, the most intuitive way is to compare the emotion representation of contextual input with the label embedding of each emotion. Thus the prediction of all possible emotion labels could be given by calculating the similarity features.

Let the function En_e as emotion projector, maps contextual-level representation h^c into emotion representation h^e. In this way, the similarity vector sim_k between h^e and each label embedding $E_k \in E$ could be calculated:

$$h^e = \ En_e(h^c) = \ w_e^T \cdot h^c + b_e \tag{4.9}$$

$$sim_k = h^e \odot E_k, k \in [1, K] \tag{4.10}$$

in which $\odot$ is the element-wise product operation. In this way, the probability of containing emotion e_k is defined as:

$$p_k = \sigma(w_c^T \cdot sim_k + b_c) \tag{4.11}$$

in which σ indicates the sigmoid activation function, and w_c, b_c indicate the model parameters. The final prediction is given: $P = \{p_1, \dots p_k, \dots p_K\}$.

4.2.3 Training Objectives

For multi-label emotion recognition task, the training objective is often based on binary cross-entropy (BCE). However, BCE loss function takes each emotion as an independent individual and does not consider their relationships. Emotion correlation plays an important role in this task, which makes emotion recognition be a more complex problem than traditional text classification. To guide the model to learn the emotion correlation during the training process, we propose an assembled training objective to consider all aspects.

4.2.3.1 Training objective on output layer

To minimize the loss between the true label distribution and the output distribution, label-correlation aware multi-label loss function, is applied at the output layer, which is determined as follows:

$$loss_{ML} = \sum_{i=1}^{N} \frac{1}{|Y_i||\overline{Y_i}|} \sum_{(k,l)\in Y_i \times \overline{Y_i}} exp(-(p_k^i - p_l^i)) \tag{4.12}$$

where Y_i denotes the set of positive emotions for ith sample x_i, and $\overline{Y_i}$ denotes the negative emotions. p_k^i and p_l^i are the output possibility of positive emotion e_k and negative emotion e_l respectively. Therefore, training with the above loss function is

equivalent to maximizing the difference of $(p_k^i - p_l^i)$, which leads the system to output larger values for positive emotions and smaller values for others.

4.2.3.2 Training objective on label embedding

Given a contextual input x, its positive labels are Y_i and its negative labels are $\overline{Y_i}$, and $Y = Y_i \cup \overline{Y_i}$. Emotion representation h^e is learned as in Eq. (4.9). In the proposed network, nonlinear label embedding is utilized in the network to guiding the final prediction P by the similarity feature with h^e. In this way, we assume that h^e can in turn be used in the training of label embedding by being closer to the embedding of positive emotions while farther to other negative emotions.

To measure the distance of emotion representation h^e and label embedding, cosine embedding loss is utilized:

$$loss_{CosEmbed} = \sum_{i=1}^{N} \frac{1}{K} \cdot \sum_{k=1}^{K} CosLoss(h_i^e, E_k) \tag{4.13}$$

$$CosLoss(h_i^e, E_k) = \begin{cases} 1 - cos(h_i^e, E_k), & y_i \in Y_{pos} \\ max(0, cos(h_i^e, E_k) - margin), & y_i \in Y_{neg} \end{cases} \tag{4.14}$$

in which margin is a number from -1 to 1.

To guarantee label embedding can encode semantic features among labels, we introduce an additional network to recognize each emotion from corresponding label embedding. For each emotion e_i, its label embedding is E_i. The prediction $\hat{e}_i$ based on E_i is given as:

$$p_{e_k} = softmax(W_{ek} \cdot E_k + b_{ek}) \tag{4.15}$$

$$loss_{LabelEmbed} = \frac{1}{K} \cdot \sum_{k=1}^{K} - e_k \cdot log(p_{e_k}) \tag{4.16}$$

In summary, the assemble training objective of the proposed method is as follows:

$$Loss(x, y) = loss_{ML} + loss_{CosEmbed} + \ loss_{LabelEmbed} \quad (4.17)$$

4.3 Experimental Results and Discussions

4.3.1 Experimental Setup

The experiments are conducted on Chinese emotion corpus RenCECps to evaluate the proposed architecture (RenCECps: *http://a1-www.is.tokushima-u.ac.jp/member/ren/Ren-CECps1.0/DocumentforRen-CECps1.0.html*). RenCECps is an annotated emotional corpus with Chinese blog texts. The corpus is annotated in document, paragraph, and sentence level [49]. Each level is annotated with eight emotional categories ('Joy', 'Hate', 'Love', 'Sorrow', 'Anxiety', 'Surprise', 'Anger', and 'Expect').

Our experiments are conducted at sentence level, and the preceding two sentences of the current sentence are taken as the context information. After pre-processing, there is a total of 24310 contextual sentences in training data and 6746 in testing data. Label cardinality(LCard) is a standard measure of multi-labeled-ness and means the average number of emotions concluded per sentence of the corpus and [151]. In RenCECps, LCard is 1.4468.

4.3.2 Evaluation Metrics

The global performance is evaluated by micro and macro F1-score. F1 score is the harmonic mean of precision and recall. Micro F1-score gives each sample the same importance, while macro F1-score takes all classes as equally important. Some popular evaluation measures typically utilized in multi-label classification can be utilized to measure the efficiency of proposed methods. Hamming Loss (HL) is the fraction of labels that are incorrectly predicted. Coverage evaluates how far it is needed to go down the ranked emotion list to cover all the relevant emotions in the instance. One Error (OE) evaluates the fraction of sentences whose top-ranked emotion is not in the relevant emotion set. Ranking Loss (RL) evaluates the average fraction of label pairs that are

reversely ordered for instance.

4.3.3 Experimental Details

For a given sentence, its preceding two sentences are taken as contextual sentences. There are total 8 emotions labels annotated for each sentence, and the dimension of label embedding is set to 256. The dimension of hidden state of GRU cell is set to 768/2, and 768 is the dimension of sentence-level embedding.

During the model training, the learning rate is set to 2e-5, and the batch size is set to 128. Adam optimization method is applied to train the model by minimizing the proposed training objective.

4.3.4 Baselines

In this section, we report the experimental results of our proposed method and baseline models. Additionally, we analyze the influence of training objectives on output layer and label embedding.

We compare our proposed model with six baseline methods as follows.

(1) RERc [152]: a novel framework based on relevant emotion ranking to identify multiple emotions and produce the rankings of relevant emotions from text.

(2) HANs [148]: it has a hierarchical structure that mirrors the hierarchical structure of documents, and has two levels of attention mechanisms applied at the word-and sentence-level. In our experiments, sentence-level encoder of HANs is replaced by pre-trained BERT model.

(3) EDL [153]: Emotion Distribution Learning, it learns a mapping function from texts to their emotion distributions describing multiple emotions and their respective intensities based on label distribution learning.

(4) EmoDetect [154]: it outputs the emotion distribution based on a dimensionality reduction method using non-negative matrix factorization which combines several constraints such as emotions bindings, topic correlations, and emotion lexicons in a constraint optimization framework.

(5) ML-KNN [155]: Multi-Label k-Nearest Neighbor, which adapts traditional k-

nearest neighbor (KNN) algorithm to deal with multi-label data.

(6) Rank-SVM [156]: adapt maximum margin strategy to deal with multi-label data, focuses on distinguishing relevant from irrelevant while neglecting the rankings of relevant ones.

4.4 Experimental Results and Discussions

4.4.1 Experimental Results

The experimental results of our model compared with the baselines on RenCECps dataset are shown in Table 4.1. Results indicate that our proposed method outperforms other baselines to a great extent. For example, compared to the baseline RERc, our model achieves an improvement of 10.73% micro-F1 score. On multi-label evaluation measures, our model achieves a reduction of 46.15% ranking loss and 21.78% one error. Compared to other baselines, our model achieved satisfactory results as well, which demonstrated the effectiveness of the proposed method.

Table 4.1 Experimental results in RenCECps.

Metrics	Ours	RERc	HAN	EDL	Emo-Detect	ML-KNN	Rank-SVM
Micro F1 (↑)	**0.5665**	0.5116	0.5573	0.4620	0.4552	0.4720	0.4962
Macro F1(↑)	**0.4186**	0.4161	0.4003	0.3923	0.3622	0.3632	0.3965
Ranking loss (↓)	**0.1132**	0.2102	0.1136	0.2589	0.2781	0.2928	0.3024
One-Error (↓)	**0.3559**	0.4550	0.3623	0.5227	0.5352	0.5543	0.5606
Coverage (↓)	2.1272	**2.1268**	2.1272	2.1699	2.8956	2.4448	2.5962
Hamming loss (↓)	**0.1998**	0.2014	0.2075	0.2102	0.2202	0.2409	0.2585

4.4.2 Discussions of Label Embedding Layer

Our proposed model is an extension of the baseline of HANs. In our experiments, sentence-level encoder of HANs is replaced by pre-trained BERT model. Therefore, by comparing the results of these two models, it can be revealed whether the addition of label embedding layer is effective on the sub-task of emotion correlation learning.

As we can see from the results shown in Table 4.1, the proposed model significantly outperforms baseline HANs, which achieves the improvement of micro-F1 score from 0.5573 to 0.5665 and macro F1 score from 0.4003 to 0.4186. On multi-label evaluation measures, our model achieves a reduction of ranking loss from 0.1136 to 0.1131, one error from 0.3623 to 0.3559, and hamming loss from 0.2075 to 0.1998.

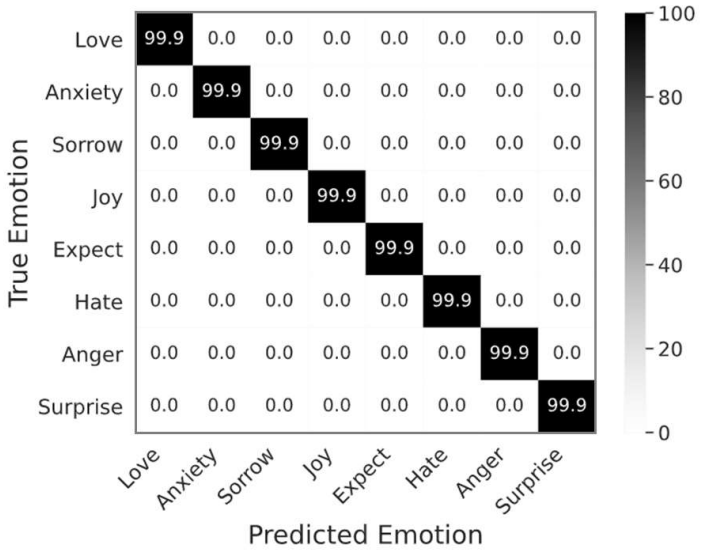

Figure 4.2 The prediction probability given by label embedding matrix

Both the proposed method and baseline HANs give predictions based on the contextual representation learned from a hierarchical network. HANs directly fed it into output layer for final prediction, which mainly consists of a fully connected layer and an activate function like sigmoid. This implementation is intuitive and straightforward, and it is also a typical processing method in most multi-label classification tasks. However, such implementation treats emotion recognition task as a general text classification task. It does not consider the correlation between emotion labels, such as the probability of co-occurrence of “Love” and “Happy” is higher than that of “Love” and “Sad”. In our proposed model, label embedding space is introduced for emotion correlations learning. The final prediction is based on the interaction of the emotion representation of input text and the label embedding matrix. To guarantee that the semantic features among labels can be learned in the label embedding matrix, an auxiliary training objective on label embedding is utilized to guide the training. The predicted probability given by label

embedding matrix, as Eq. (4.15), is visualized in Figure 4.2. The results in the figure clearly show that the label embedding matrix can accurately predict the corresponding emotion, which suggests that the emotional information of each label has been actually learned in the label embedding matrix.

4.4.3 Discussions of Training Objectives

As described in section 4.2.3, we proposed an assembled training objective to realize the joint learning of both emotion recognition task and label embedding task. To evaluate the effectiveness of training objectives and label embedding network, we train the proposed model with different training objectives. The results are shown in Table 4.2. The symbol 'M', 'C' and 'L' denotes the loss function of multi-label loss, as in Eq.(4.12), cosine embedding loss, as in Eq. (4.13) and label embedding loss, as in Eq. (4.16), which are utilized for training.

Table 4.2 Results of proposed models with different training objectives.

Metrics	M+C+L	M+C	M+L	M
Micro F1 (↑)	**0.5665**	0.5655	0.5570	0.5539
Macro F1(↑)	0.4186	**0.4246**	0.4156	0.4128
Ranking loss (↓)	**0.1132**	0.1209	0.1194	0.1272
One-Error (↓)	**0.3559**	0.3734	0.3719	0.3787
Coverage (↓)	**2.1272**	2.1778	2.1638	2.2234
Hamming loss (↓)	0.1998	0.1959	0.2040	**0.1957**

Note: 'M': multi-label loss, 'C': cosine embedding loss, 'L': label embedding loss.

As shown in Table 4.2, compared with the assembled training objective ('M+C+L'), the proposed model with only multi-label loss ('M') on output layer achieves a reduction of 2.22% micro F1 and 1.39% macro F1, and an improvement of 12.37% ranking loss, 6.41% one-error and 4.52% coverage. It suggests that the proposed ensemble training objective can contribute to the classification improvement.

The experimental results of the proposed model trained on 'M+C' and 'M' indicate the contribution of cosine embedding loss on the training of the label embedding matrix. Cosine embedding loss guides the training of label embedding by making the emotion

representation of input being closer to the embedding of positive emotion labels while farther to other negative emotion labels.

The comparison results of the proposed model trained on ‘M+L’ and ‘M’ indicate that the addition of label embedding loss is effective. Label embedding loss guarantees that the trained label embedding matrix can encode semantic features among emotion labels.

4.5 Summary

In this chapter, we proposed a hierarchical network with label embedding for contextual emotion recognition. Our method involves hierarchically encoding the given sentence based on its contextual information, and training a label embedding matrix with an assembled training objective to realize emotion correlation learning. The experimental results show the strong ability of proposed method to learn emotional features for contextual emotion recognition. In the future, it shall be interesting to incorporate background resources, such as emotion lexicon and knowledge graph, to make the system more satisfactory and robust.

Chapter 5

Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning

5.1 Introduction

With the rapid development of social media platforms, understanding the latent emotions expressed in user-generated content has gained much attention because of its vast potential applications [157], [158]. Although many studies have been conducted, most of them are done in the single-emotion environment [159]. They are based on the assumption that certain textual data is associated with only one emotion. However, in real-world conditions, people often hold multiple complex emotions simultaneously, and a textual expression is often associated with multiple emotions simultaneously. Therefore, multi-label emotion detection has gained burgeoning attention because of its vast potential applications.

Multi-label emotion detection aims to recognize all possible emotions in a piece of textual expression [160]. In conventional emotion detection networks, textual information is often encoded together into a representation vector and then directly fed into the classifier[161], [162]. However, in a textual expression with multiple emotions, there may be some emotions with relatively weaker intensity. If information of each emotion is mixed and encoded together into a shared vector, the weaker emotions with subtle features

could be covered by stronger emotions and be challenging to recognize. To accurately recognize the emotions expressed, the quality of underlying emotional feature representation has an essential influence on the final prediction.

In most previous researches, multi-label emotion detection task is often narrowed down into multiple binary classifications [163], in which each emotion is detected respectively without considering their correlations. However, emotion correlation information provides non-ignorable features and is useful for improving the performance of emotion detection. The definition of emotion correlation can be illustrated based on Plutchik's work. In an emotional expression, emotion correlation mainly refers to positive or negative emotional correlation. Positively correlated emotions are similar to each other and often appear together but with different intensities. Such as the emotion pair ‘Joy’ and ‘Love’ tend to appear simultaneously. Negatively associated emotions are often opposite to each other and rarely appear together, such as ‘Love’ and ‘Sorrow’. Emotion correlation can be utilized to facilitate more in-depth emotion analysis in multi-label emotion recognition task.

In this chapter, a Multi-label Emotion Detection Architecture (MEDA) is proposed to address the above challenges. MEDA is mainly composed of two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). MC-ESFE consists of multiple channels by which the features of each emotion are separately encoded. Each channel is devoted to the underlying feature representation of a specified emotion from both sentence-level and context-level. Furthermore, an external emotion lexicon is introduced as prior knowledge to integrate more detailed emotional information. ECorL module is devoted to learning emotion correlation based on extracted emotion-specified features from MC-ESFE. In ECorL, multi-label emotion detection task is transformed as an emotion sequence prediction task. Bidirectional GRU network is taken as the emotion sequence predictor, and the emotions are sequentially predicted in a fixed path. In the hidden state of each step, the emotion correlations of current emotion are learned by information interaction with the context of other emotions flowed from both forward and backward directions. Considering that the proposed MEDA network extracts emotional information from sentence level, context level, and emotion correlation level, an ensemble model called MEDA-FS is proposed to integrate emotional

information from different levels. MEDA-FS can realize the maximization of information retention and avoid information loss during bottom-up learning. During the training, positive-negative emotion correlation is incorporated into the proposed multi-label focal loss function. By introducing a weighting factor, our loss will focus more on misclassified emotion pairs and balance the prediction between positive and negative emotions.

Compared with existing multi-label emotion detection methods, the proposed MEDA architecture extracts both emotion-specified features and emotion correlations. Experimental results show that the proposed MEDA achieves state-of-the-art performance on this task and demonstrates its effectiveness.

The major contributions of this chapter can be summarized as follows.

(1) MEDA architecture composed of MC-ESFE and ECorL modules is proposed for the textual multi-label emotion detection task. MC-ESFE can encode emotion-specified features in the corresponding channel respectively, strengthening the underlying feature representation of each emotion. ECorL is proposed to learn emotion correlations by transforming multi-label emotion detection task into emotion sequence prediction task.

(2) MEDA-FS is proposed to fuse the information at sentence-level, context-level, and emotion correlation level, realizing the maximization of information retention during bottom-up learning.

(3) Multi-label focal loss function considering emotion correlation information is proposed for multi-label learning. This loss function contributes to model training by focusing on misclassified emotion pairs and balancing the prediction of positive and negative emotions.

The rest of this chapter is organized as follows: Details of the proposed MEDA are given in Section 5.2. The experimental setting and details are presented in Section 5.3. The performance of MEDA is discussed in Section 5.4. Finally, summary is drawn in Section 5.5.

5.2 Proposed Method

To comprehensively obtain emotional information of texts, Multi-label Emotion

Detection Architecture (MEDA) is proposed in this chapter. It mainly composes two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). The framework of MEDA is shown as Figure 5.1.

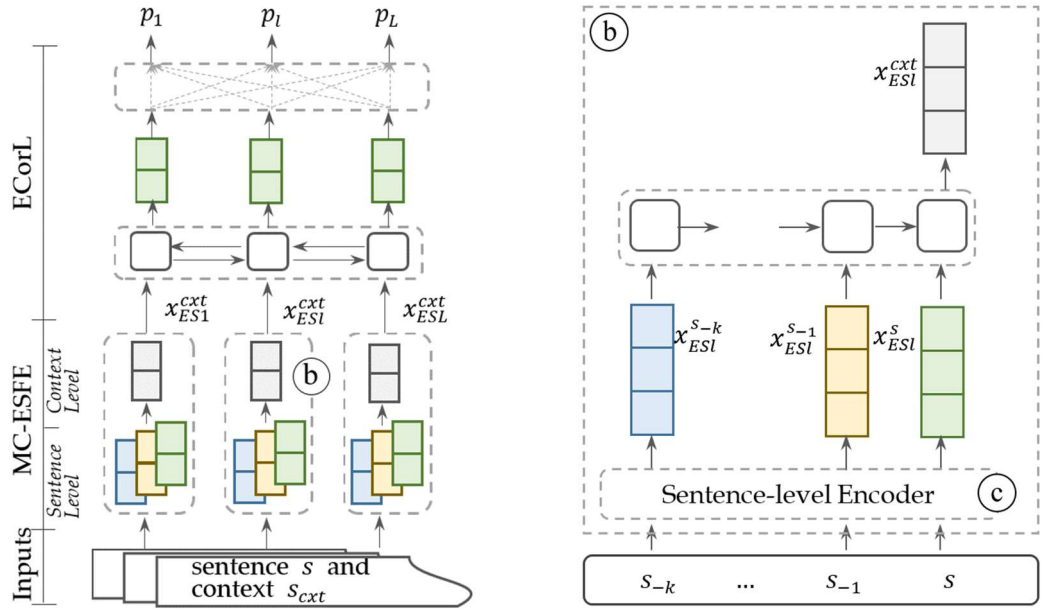

Figure 5.1 The illustration of MEDA architecture. Region b is the $l-t$h ESFE channel.

Multi-label emotion detection task aims to detect all possible emotions from the pre-defined emotional label set: $E = [e_1, e_2, \dots e_L]$. Considering the important influence of contextual information on this task, the previous k sentences occurred before current sentence s are taken as the context sentences: $s_{cxt} = [s_{-k}, \dots\ s_{-2}, s_{-1}]$. Given a sentence $s = [w_1, w_{2\dots} w_n]$ and its context s_{cxt}, our proposed multi-label emotion recognition model MEDA is trained to output the predicted probability distribution $P_{ML} = [p_1, p_2, \dots p_L]$ of each emotion, denoted as:

$$P_{ML} = f_{MEDA}(s, s_{cxt}) \tag{5.1}$$

MC-ESFE module is composed of L parallel channel-wise ESFE. In each channel, ESFE extracts emotion-specified features from sentence-level to context-level through a hierarchical structure. The output of each channel is combined into an emotion-specified feature matrix: $X_{ES}^{cxt} = [x_{ES}^{cxt}, x_{ES}^{cxt}, \dots, x_{ESL}^{cxt}]$. In ECorL module, emotion correlations are further learned from X_{ES}^{cxt} and multi-label emotions are predicted. Specifically, MEDA

architecture is very flexible, and the algorithm applied in each module can be replaced by other state-of-the-art algorithms.

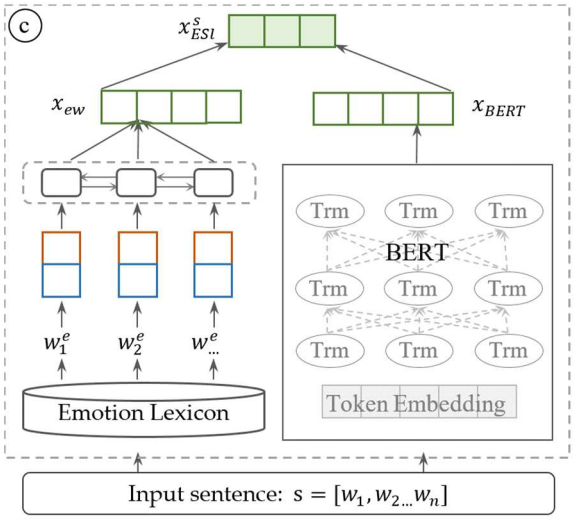

Figure 5.2 The illustration of sentence-level encoder of lth ESFE channel

5.2.1 MC_ESFE: Multi-Channel Emotion Specified Feature Extractor

In this chapter, a Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) is proposed for underlying fundamental feature extraction. MC-ESFE is composed of L channel-wise ESFE, and L is equal to the number of emotions. Each channel focuses on the feature extraction of a specified emotion, and each emotion's information is separately encoded in each channel. In this way, more details of each emotion could be summarized, and the features of weak emotions are prevented from being covered by strong emotions to some extent.

Figure 5.1 shows the hierarchical structure of lth ESFE-channel corresponding to emotion e_l, $l \in [1, L]$. Each channel contains a sentence-level encoder and a context-level encoder, which focus on feature extraction of emotion e_l on both sentence-level and context-level.

5.2.1.1 Sentence Level Encoder

In lth ESFE channel, given a sentence s, the sentence-level encoder f_{S-En}^{l} projects input sentence s to emotion-specified feature x_{ESl}^{s}:

$$x_{ESl}^{s} = f_{S-En}^{l}(s),\ l \in [1, L] \tag{5.2}$$

In sentence-level encoder f_{S-En}^{l}, as shown in Figure 5.2, two parallel architectures with different embedding methods are employed to generate: (1) emotional feature representation x_{ew}, (2) general sentence representation x_{BERT}. They are integrated into emotion-specified sentence-level representation x_{ESl}^{s} for further context-level learning.

General sentence representation. Inspired by the pre-trained language model learning approach and transfer learning techniques, pre-trained Chinese BERT model [84] is applied to yield general sentence representation in this chapter. BERT stands for Bidirectional Encoder Representations from Transformers. Chinese BERT is designed to pre-train deep bidirectional representations from unlabeled Chinese text by jointly conditioning on both left and right context in all layers. It remedies the limitation of insufficient training corpora and contributes to syntactic and semantic sentence representation. Given a sentence s, the general sentence representation x_{BERT} is generated from Chinese BERT.

Emotional feature representation. Arguably, it is accepted that general sentence representation generated by pre-trained language model does not contain specific emotional features, as no emotion-related knowledge has been included in the training process. To generate emotional sentence representation, emotional features are further extracted based on an external n-dimensional emotion lexicon.

With the input sentence $s = [w_1, w_2, \dots]$, emotional words $w^{\mathrm{e}} = [w_1^e, w_2^e, \dots]$ occurred in s are firstly extracted by matching the emotion lexicon. The embedding of emotional words consists of two parts. The first is general word embedding, which is realized by mapping the pre-trained Word2vec word embedding matrix. Each word is embedded as $v_{\mathrm{w2v}} \in R^{1*D}$, in which D is the embedding dimension. The second is emotional word embedding, which is realized based on n-dimensional emotion lexicon. Each word is embedded as $v_{\mathrm{emo}} \in R^{1*n}$, in which n is the number of emotions annotated in emotion lexicon and the value means the intensity of corresponding emotion. Finally, emotional word embedding is represented as $E = [v_1^e, v_2^e, \dots]$, in which $v_i^e \in R^{1*(D+n)}$.

Considered the polysemy of emotional words in different contexts, BiGRU (Bidirectional

Gated Recurrent Neural Networks) [147] and attention network [148] are utilized to make the network pay more attention to significant emotional words. Take emotional embedding E as input, the output of the hidden state of BiGRU in each step is $h_i = [\overrightarrow{h_i}: \overleftarrow{h_i}]$, in which $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are the output of hidden states from forward and backward directions, respectively. The attention mechanism considers the contributions of different emotional words to the prediction of specified-emotion e_l. More attention weight will be assigned to words related to emotion e_l in the current lth ESFE channel. Attention weight a_i and weighted emotional feature vector x_{ew} are defined as follows:

$$e_i = W_2^T[\sigma(W_1^T \cdot h_i + b_1)] + b_2 \tag{5.3}$$

$$a_i = \frac{exp(e_i)}{\sum_{k=1}^{n} exp(e_k)} \tag{5.4}$$

$$x_{ew} = [a_1 h_1 : \ a_2 h_2 : \dots a_i h_i : \dots] \tag{5.5}$$

in which σ indicates the sigmoid activation function, w_1, b_1, w_2, b_2 indicate the model parameters, and [:] indicates the concatenation operation.

Finally, emotional feature vector x_{ew} and general embedding x_{BERT} is integrated, and emotion-specified sentence-level representation x_{ESl}^s is generated as follows:

$$x_{ESl}^s = tanh(W_{\mathrm{e}} \cdot x_{\mathrm{ew}} + W_B \cdot x_{BERT} + b_s) \tag{5.6}$$

in which W_{e} , W_B and b_s indicate the model parameters.

5.2.1.2 Context Level Encoder

Context level encoding is channel-wise implemented as well. As shown in Figure 5.1, in lth ESFE channel corresponding to emotion e_l, given a sentence s and its context $s_{cxt} = [s_{-k}, \dots \ s_{-2}, s_{-1}]$, context-level encoder f_{C-En}^l gives contextual emotional feature x_{ESl}^{cxt}. GRU network is utilized to learn contextual information from previous k sentences, and the output of final step is captured as the context-level representation, which is denoted as:

$$x_{ESl}^{cxt} = f_{C-En}^{l}\left(x_{ESl}^{s-k}, \dots, x_{ESl}^{s-1}, x_{ESl}^{s}\right) = f_{\mathrm{GRU}}\left(x_{ESl}^{s-k}, \dots, x_{ESl}^{s-1}, x_{ESl}^{s}\right), l \in [1, L] \quad (5.7)$$

$$x_{ESl}^{s-i} = f_{S-En}^{l}(s_{-i}),\ i \in [1, k] \quad (5.8)$$

Contextual emotion-specified features x_{ESl}^{cxt} learned from each channel in MC-ESFE is output and combined as the emotional feature matrix: $X_{ES}^{cxt} = [x_{ES1}^{cxt}, x_{ES2}^{cxt}, \dots, x_{ESL}^{cxt}]$. X_{ES}^{cxt} is flowed into ECorL module for further emotion correlation learning.

5.2.2 ECorL: Emotion Correlation Learner

Emotion correlations are indispensable in multi-label emotion detection task. In this chapter, ECorL (Emotion Correlation Learner) is proposed to give emotion prediction based on emotion correlation learning.

MC-ESFE module project inputs into a sequence of continuous emotional representations $X_{ES}^{cxt} = [x_{ES1}^{cxt}, x_{ES2}^{cxt}, \dots]$. ECorL module takes X_{ES}^{cxt} as input. In ECorL module, multi-label emotion detection task is transformed as emotion sequence prediction task, and emotions are predicted in a fixed path. Refer to the previous work [170], the order of emotion sequence is set according to its occurred cumulative number in the corpus. BiGRU is taken as the emotional sequence predictor. The operation is formulated as follows:

$$H_e = f_{\mathrm{BiGRU}}(x_{ES1}^{cxt}, x_{ES2}^{cxt}, \dots, x_{ESL}^{cxt}) \quad (5.9)$$

$$P_{ML} = tanh(W_{ECor} \cdot H_e + b_{ECor}) \quad (5.10)$$

in which $H_e = [h_{e1}, \dots h_{el}, \dots h_{eL}]$, are the hidden states of each step, W_{ECor} and b_{ECor} are the learned weight and biases, and P_{ML} is the predicted probability of each emotion. In lth step of BiGRU, the learning of hidden state h_{el} can be viewed as the feature extraction of a specified emotion e_l. Invalid information of current input x_{ESl}^{cxt} can be filtered because of the gating mechanism. With the bidirectional network, emotional feature h_{el} is learned based on the information of other emotions flowed from both forward and backward hidden

state. In this way, emotional information interaction is realized. The hidden states of BiGRU are output and fed into emotion interaction layer. This layer is a fully-connected layer and aimed to realize further emotional information interaction. In this way, the final emotion prediction is obtained: $P_{ML} = [p_1, p_2, .. p_L]$.

5.2.3 Network Pre-training in MC-ESFE

Each channel in MC-ESFE is dedicated to obtaining corresponding emotional information, which belongs to the underlying feature extraction in the MEDA framework. The quality of feature representation has a direct impact on the performance of upper-level emotion predictions. To improve the underlying feature representation, network-based transfer learning is employed to pre-train the sentence-level encoder in each channel. During transfer learning, a prediction layer is added to emotional sentence representation x^s_{ESl} for single-emotion prediction:

$$p^s_{ESl} = \sigma(w_l \cdot x^s_{ESl} + b_l). \tag{5.11}$$

in which x^s_{ESl} is the sentence-level representation of input sentence s, and p^s_{ESl} indicates the predicted probability of emotion e_l expressed in sentence s. The first-step pre-training is implemented on positive-negative annotated emotional datasets. The second-step is fine-tuning. During fine-tuning, the multi-emotion annotation $\{s, [y_1, y_2, ... y_L]\}$ of each sentence s in dataset D is transformed into multiple single-emotion annotations: $\{s, y_1\}$, $\{s, y_2\}$…$\{s, y_L\}$. In this way, we reconstructed multiple binary-dataset: $\widehat{D} = \{\widehat{D}_1, \widehat{D}_2, ... \widehat{D}_L\}$. For each binary-dataset $\widehat{D}_l$, $l \in [1, \mathrm{L}]$, sentence $s \in \widehat{D}_l$ is annotated as $\{s, y_l\}$. $\widehat{D}_l$ is fed into lth ESFE channel to fine-tune the sentence-level parameters. During pre-training, binary focal loss [164] is utilized:

$$E_{FL} = \ -\alpha_t(1 - p_t)^r log(p_t) \tag{5.12}$$

$$p_t = \begin{cases} p^s_{ESl} & if\ \ y_l = 1 \\ 1 - p^s_{ESl} & otherwise \end{cases} \tag{5.13}$$

in which r is a modulating factor, and it aimed to reduce the relative loss of well-classified examples. $\alpha \in [0,1]$, is a weighting factor to address the problem of class imbalance. $\alpha_t = \alpha$ for positive label and $\alpha_t = 1 - \alpha$ for negative label.

5.2.4 MEDA-FS: Multi-level Information Fusion

The proposed MEDA architectural learns emotional information from sentence-level to context-level, from single-emotion level in MC-ESFE to multi-emotion level in ECorL module. Each layer in MEDA network learns different levels of information. To realize the maximization of information retention and avoid information loss during bottom-up learning, MEDA-FS is proposed to fuse the information from different levels. MEDA-FS consists of three sub-models: S-MC-ESFE, C-MC-ESFE, and MEDA, which give emotion predictions on sentence-level, context-level, and emotion correlation level, respectively.

S-MC-ESFE, gives sentence-level predictions $P_{ES}^{s} = [p_{ES1}^{s}, \dots, p_{ESL}^{s}]$. It is obtained during the pre-training step of sentence-level encoder in MC-ESFE, which is detailed in section 5.2.3. P_{ES}^{s} represents the prediction based on the underlying information, without considering the emotion correlations and contextual information.

C-MC-ESFE, gives context-level predictions P_{ES}^{cxt} based on sentence-level predictions of current sentence s, denoted as P_{ES}^{s}, and sentence-level predictions of its context s_{cxt}, denoted as $[P_{ES}^{s-k}, \dots, P_{ES}^{s-1}]$. GRU network is utilized to learn contextual information and its final output is taken as the prediction:

$$P_{ES}^{cxt} = f_{\mathrm{GRU}}([P_{ES}^{s-k}, \dots P_{ES}^{s-1}, P_{ES}^{s}]) \tag{5.14}$$

MEDA, gives prediction $P_{ML} = [p_1, p_2, .. p_L]$ by considering both contextual information and emotion correlation.

MEDA-FS, gives final predictions by comprehensively fuse the information from above three level, denoted as:

$$P = w_s \cdot P_{ES}^{s} + \ w_c \cdot P_{ES}^{cxt} + w_{ML} \cdot P_{ML} \tag{5.15}$$

in which w_s, w_c and w_{ML} are the weight parameters of each level's information.

5.2.5 Definition of Multi-label Focal Loss

Multi-label (ML) loss function [165] is the commonly used loss function in multi-label learning. Instead of concentrating on individual label discrimination like traditional cross-entropy loss function, ML-loss focused on considering the correlations between the different labels. Inspired by [164], ML-loss is rewritten as multi-label focal loss. Multi-label focal loss not only considers emotion correlation but also focus more on misclassified emotion pairs. Besides, it introduces a harmonic parameter to reduce the influence of the imbalance prediction of positive and negative emotions. The definition of multi-label focal loss is defined as follows:

$$E_{ML-FL} = \sum_{i=1}^{N} \frac{1}{|Y_i||\overline{Y_i}|} \sum_{(k,l)\in Y_i \times \overline{Y_i}} \alpha_{kl}^{i} \cdot exp(-(p_k^i - p_l^i) \tag{5.16}$$

$$\alpha_{kl}^{i} = \ w \cdot (1 - p_k^i)^r + (1 - w) \cdot (p_l^i)^r \tag{5.17}$$

in which Y_i denotes the set of positive emotions expressed in ith instance s_i, and $\overline{Y_i}$ denotes the negative emotion set. p_k^i and p_l^i are the predicted probability of positive emotion e_k and negative emotion e_l respectively. Therefore, the training with above loss function is equivalent to maximizing the difference of negatively related emotion pair of $(p_k^i - p_l^i)$. This leads the system to output a higher probability for positive emotion while a lower probability for negative emotion. In this way, the emotion correlation of negatively related emotion pairs can be taken into consideration.

α_{kl}^{i} is a weighting factor and mainly affected by two parameters: $w \in (0,1)$ is a harmonic factor aimed to balance the prediction between positive and negative emotions, and $r > 0$ is a modulating factor aimed to make the loss put more focus on hard and misclassified examples during training. Significantly, while $r = 0$, the proposed multi-label focal loss is equivalent to the multi-label loss function.

For well-classified positive-negative emotion pairs (e_k, e_l), predicted probability p_k^i tends to 1 while p_l^i tends to 0. In this case, the difference of $(p_k^i - p_l^i)$ tends to the maximum, which means the minimum of $exp(-(p_k^i - p_l^i)$, and the weighting factor α_{kl}^{i} tends to 0.

Thereby the loss of well-classified positive-negative emotion pairs is minimized. Conversely, for hard-classified emotion pairs, the difference of $(p_k^i - p_l^i)$ tends to the minimum, which could be caused by p_k^i tending to 0 or p_l^i tending to 1. In response to the above two cases, $w \cdot (1 - p_k^i)^r$ and $(1 - w) \cdot (p_l^i)^r$ are introduced to give more focus on misclassified p_k^i and p_l^i respectively.

5.3 Experiments Setup

5.3.1 Datasets

We employ two different datasets to evaluate the proposed architecture, which are listed below:

Ren-CECps Dataset is an annotated emotional corpus with Chinese blog texts. The corpus is annotated in the document, paragraph, and sentence level. Each level is annotated with eight emotional categories ('Joy', 'Hate', 'Love', 'Sorrow', 'Anxiety', 'Surprise', 'Anger', and 'Expect') and corresponding discrete emotional intensity value from 0.0 to 1.0. In our experiments, those emotions with an intensity greater than 0.0 are labeled as 1, otherwise 0. 'Neutral' is regarded as the $9th$ emotion label in case the sentence holds no emotion. After pre-processing, there is a total of 27091 sentences in training data and 7681 sentences in testing data. The average number of emotions expressed in a sentence is 1.4468.

NLPCC2018 Dataset consists of code-switching texts in Chinese, and concerns another language English on a small scale [47]. There are total 5 emotions annotated: 'Happiness', 'Sadness', 'Anger', 'Fear', and 'Surprise'. After pre-processing, there is 4611 texts in training data and 955 texts in testing data. The average number of emotions expressed in a sentence is 1.1466.

The cumulative number of each emotion e_i on Ren-CECps and NLPCC Datasets is calculated:

$$CN_i = \sum_{\mathrm{n}=1}^{N} (y_{n,i} = 1) \tag{5.18}$$

in which $y_{n,i}$ is the annotation of emotion e_i in n_{th} sample. The statistical results are

shown in Table 5.1.

Table 5.1 Cumulative number of each emotion in Ren-CECps and NLPCC2018 Datasets.

Ren-CECps				NLPCC2018	
Love	11909	Hate	3533	Happiness	2534
Anxiety	10099	Anger	2236	Sadness	1502
Sorrow	8184	Surprise	1121	Surprise	811
Joy	6223	Neutral	2488	Anger	765
Expect	4633	-	-	Fear	770

5.3.2 Experimental Details

In this section, we illustrate the experimental details during the model training.

In terms of the embedding of emotional words, it mainly consists of two parts. The first part is the general word embedding. It is initialized by 300-dimensional Word2vec word embedding, which is trained on Chinese microblog data [166]. The second part is emotional embedding by mapping an external n-dimensional emotion lexicon. Existing emotion lexicons are very rare due to the subjective and inconsistent annotation. In our experiments, a dimensional emotion lexicon is manually built based on word-level annotation on Ren-CECps. In our lexicon, each emotion word is annotated as an 8-dimensional vector v. Each dimension corresponding an emotion in [‘Love’, ‘Anxiety’, ‘Sorrow’, ‘Joy’, ‘Expect’, ‘Hate’, ‘Anger’, ‘Surprise’], and the value represents the emotion intensity. For example, the emotional word ’不幸’ (‘Unfortunately’ in English) is represented as [0., 0.23, 0.62, 0., 0., 0., 0., 0.], which means that this word expresses stronger emotion of ‘Sorrow’ and weaker emotion of ‘Anxiety’, and the intensities are 0.62, and 0.23 respectively. Specially, an extra token named ‘[EMO_PAD]’ is added to emotion lexicon, and its embedding vector is initialized by zeros. This token will be treated as emotional word if the current sentence does not contain any other emotional words.

For RenCECps, because of the non-coexistence of ‘Neutral’ label with other emotion labels, the final prediction is subject to the condition: only if the prediction of ‘Neutral’ label obtained the highest probability among all labels, the sentence is predicted as ‘Neutral’. Otherwise, it is predicted as emotions contained.

In terms of the number of context sentences k, we set $k = 3$, which means that the previous 3 sentences are taken as the contextual information. Particularly, replication padding with the last sentence is utilized while the number of contextual sentences is less than 3.

We set the dropout as 0.2 in EcorL module to avoid over-fitting. The hidden size of BiGRU in sentence-level encoder is 64 in each direction. For the binary focal loss utilized during the network pre-training, modulating factor r is set to 2, and the weighting factor α is set to 0.75. In multi-label focal loss, we set the modulating factor r to 2 and harmonic factor w to 0.4. Adam optimization method is applied to train the model by minimizing the proposed multi-label focal loss.

5.3.3 Metrics

In multi-label emotion detection task, the evaluation is more complicated than traditional single-label emotion recognition. In this chapter, some popular evaluation measures typically utilized in this task are utilized to measure the performance of proposed methods [165].

Micro F1-score and Macro F1-score are utilized as the main metrics to evaluate the global performance of each model. F1 score is the harmonic mean of precision and recall. Micro F1-score gives each sample the same importance, while Macro F1-score takes all classes as equally important. Hamming Loss (HL) is the fraction of labels that are incorrectly predicted. Average precision (AP) evaluates the average fraction of labels ranked above a particular label $y: y \in Y_i$ are actually in Y_i, in which Y_i is positive emotion set of sentence. Coverage evaluates how far it is needed to go down the ranked emotion list to cover all the relevant emotions in the instance. One Error (OE) evaluates the fraction of sentences whose top-ranked emotion is not in the relevant emotion set. Ranking Loss (RL) evaluates the average fraction of label pairs that are reversely ordered for instance.

5.3.4 Baseline Models

To demonstrate the performance of the proposed MEDA model, some baseline methods are compared in our experiments:

BR [167], Binary Relevance, based on the label independence assumption, transforms a multi-label classification problem into multiple binary classification problems.

CC [168], Classifier Chains, a multi-label model that arranges binary classifiers into a classifier chain to capture the label correlations.

LP, LabelPowerset, creates one multi-class classifier for every label combination attested in the training set.

BP-MLL [165], is derived from the backpropagation algorithm by employing a novel error function to capture the characteristics of multi-label learning.

DPCNN [169], a low-complexity word-level deep pyramid CNN network that can efficiently capture global representations of text.

HANs [148], hierarchical attention networks that mirror the hierarchical structure of documents. HANs can find the essential words and sentences in a document while taking the contextual information into consideration.

SGM [170], transfers multi-label classification task to a sequence generation problem and can capture the correlations between labels.

In previous studies, several emotion recognition methods have been implemented in RenCECps datasets and achieved the previous state-of-the-art performances. Therefore, we take them as baselines to verify the performance of our method in RenCECps, which includes:

DATN [121], divides the sentence representation into two different feature spaces, which aims to capture the general sentiment words and the other critical emotion-specific words via a dual attention mechanism.

SGM-IFC [171], utilizes the attention-based Seq2Seq model to solve the multi-label problem. An initialized fully

connection layer is employed to capture the correlation between any two different labels.

For the baselines of BR, CC and LP, we take pre-trained BERT model as sentence encoder and Gaussian Naive Bayes as the classifier, and all experiments are implemented based on Scikit-multilearn library. The results of baselines BP-MLL, SGM, DATN, and SGM-IFC on RenCECps dataset are adopted from the published papers [152], [121], [171]. For others, the comparison experiments are implemented based on the open-source codes shared on GitHub.

5.4 Experimental Results and Discussions

Experimental results of the proposed method and baseline models are reported in section

5.4.1. The discussions are organized into two sections. In section 5.4.2, we analyzed the contribution of multi-level information from each sub-model. In section 5.4.3, we evaluate the effectiveness of emotional features by ablation experiments. In section 5.4.4, we explore the effectiveness of proposed multi-label focal loss on this task.

Table 5.2 Comparison results on RenCECps Dataset

	Micro F1 % (↑)	Macro F1 % (↑)	AP % (↑)	HL (↓)	Coverage (↓)	OE (↓)	RL (↓)
BR	46.40	34.79	63.69	0.2464	2.8313	0.5221	0.1789
CC	46.97	33.62	63.16	0.2282	2.9721	0.5234	0.1965
LP	45.15	42.51	62.62	0.2069	2.9117	0.5275	0.1861
BP-MLL	48.89	38.13	55.45	0.2241	3.1272	0.4625	0.3234
DPCNN	49.99	35.47	65.43	0.1583	3.0555	0.4834	0.1993
HANs	54.54	41.36	70.65	0.1504	2.4631	0.4520	0.1362
SGM	55.60	-	-	0.1758	-	-	-
DATN	-	45.70	73.20	-	-	0.4150	-
SGM-IFC	58.60	-	-	0.1613	-	-	-
S-MC-ESFE	59.24	***47.73***	75.19	0.1367	2.3170	***0.3760***	0.1163
C-MC-ESFE	55.30	34.34	74.76	**0.1213**	2.2765	0.3915	0.1134
MEDA	***59.71***	47.25	***75.76***	0.1378	***2.2369***	0.3763	***0.1084***
MEDA-FS	**60.76**	**48.31**	**76.51**	***0.1249***	**2.2226**	**0.3618**	**0.1062**

Table 5.3 Comparison results on NLPCC2018 Dataset

	Micro F1 % (↑)	Macro F % (↑)	AP % (↑)	HL (↓)	Coverage (↓)	OE (↓)	RL (↓)
BR	48.92	41.07	67.74	0.2975	2.1645	0.4958	0.2771
CC	49.92	40.51	68.63	0.2790	2.1221	0.4883	0.2668
LP	47.67	36.81	67.04	0.2456	2.1592	0.5159	0.2758
BP-MLL	55.66	41.65	74.78	0.2584	1.8896	0.4002	0.2066
DPCNN	46.07	34.25	64.22	0.2420	2.3482	0.5414	0.3231
HANs	55.69	42.78	76.92	0.2805	1.7930	***0.3758***	0.1835
SGM	57.11	36.28	64.24	0.1843	2.7813	0.4395	0.4267
S-MC-ESFE	**63.32**	***49.23***	**77.19**	0.1849	***1.7340***	**0.3780**	***0.1694***
C-MC-ESFE	60.59	46.90	76.43	***0.1719***	1.7592	0.3895	0.1749
MEDA	61.21	47.70	75.90	***0.1696***	1.7665	0.4021	0.1775
MEDA-FS	***63.02***	**49.42**	***77.12***	0.1728	**1.7288**	***0.3812***	**0.1681**

5.4.1 Experimental Results

Experimental results of the proposed methods against baselines are shown in Table 5.2 and Table 5.3, the best two results on each metric are in bold and in bold italics, respectively.

As the results shown in Table 5.2, the proposed model significantly outperforms baseline models and achieves state-of-the-art performance on RenCECps. Compared with SGM-IFC [171], which has previously achieved the state-of-the-art performances, proposed MEDA-FS has improved micro-F1 score from 58.60% to 60.76% and reduced hamming loss from 0.1613 to 0.1249. Compared with DATN, the proposed MEDA-FS has improved macro-F1 score from 45.70% to 48.31%, improved average precision from 73.20% to 76.51%, and reduced one error from 0.4150 to 0.3618. Besides, our model outperforms other deep learning methods and commonly used machine learning methods to a great extent, such as BR algorithm and SGM model.

Table 5.3 shows the experimental results of proposed model and baselines on NLPCC2018 dataset. Our proposed model achieved excellent results on almost all metrics except hamming loss. The hamming loss of proposed MEDA-FS is 0.1728, while the best is 0.1617 (achieved by LP). HL is the fraction of wrong labels to the total number of labels and penalizes only the individual labels. There are mainly two reasons for the higher hamming loss. One reason is that weak emotions are difficult to predict accurately. MC-ESFE module can prevent the features of weak emotions from being covered by strong emotions to some extent, but not completely. Their emotional features are not noticeable and are difficult to recognize. The classifier tends to conservatively predict them as negative emotions to ensure the whole performance among all emotion labels. Another reason is that the data distribution is imbalanced. It is hard to guarantee the performance of low-source emotion categories. In future work, more attention will be paid to the detection of weak and low- source emotions. In addition to hamming loss, the global performance of proposed method can also be reflected by other multi-label metrics, such as micro-F1, macro-F1, and average precision, on which the proposed method has achieved satisfying performance.

5.4.2 Discussion of Sub-models

MEDA-FS is composed of 3 sub-models: MEDA, S-MC-ESFE and C-MC-ESFE. These sub-models are devoted to learning information from different levels and contributing to a more comprehensive ensemble model. To further explore the contribution of each sub-model, we further analyze their performance on RenCECps in this section. The comparison results are shown in Table 5.4 and Figure 5.3.

Table 5.4 Comparison results of sub-models on RenCECps.

	Micro			Macro		
	P	R	F1	P	R	F1
S-MC-ESFE	52.16	68.55	59.24	42.48	56.72	47.73
C-MC-ESFE	**59.34**	51.77	55.30	43.15	32.01	34.34
MEDA	51.81	**70.46**	59.71	41.44	**57.54**	47.25
MEDA-FS	55.77	66.72	**60.76**	**46.10**	52.21	**48.31**

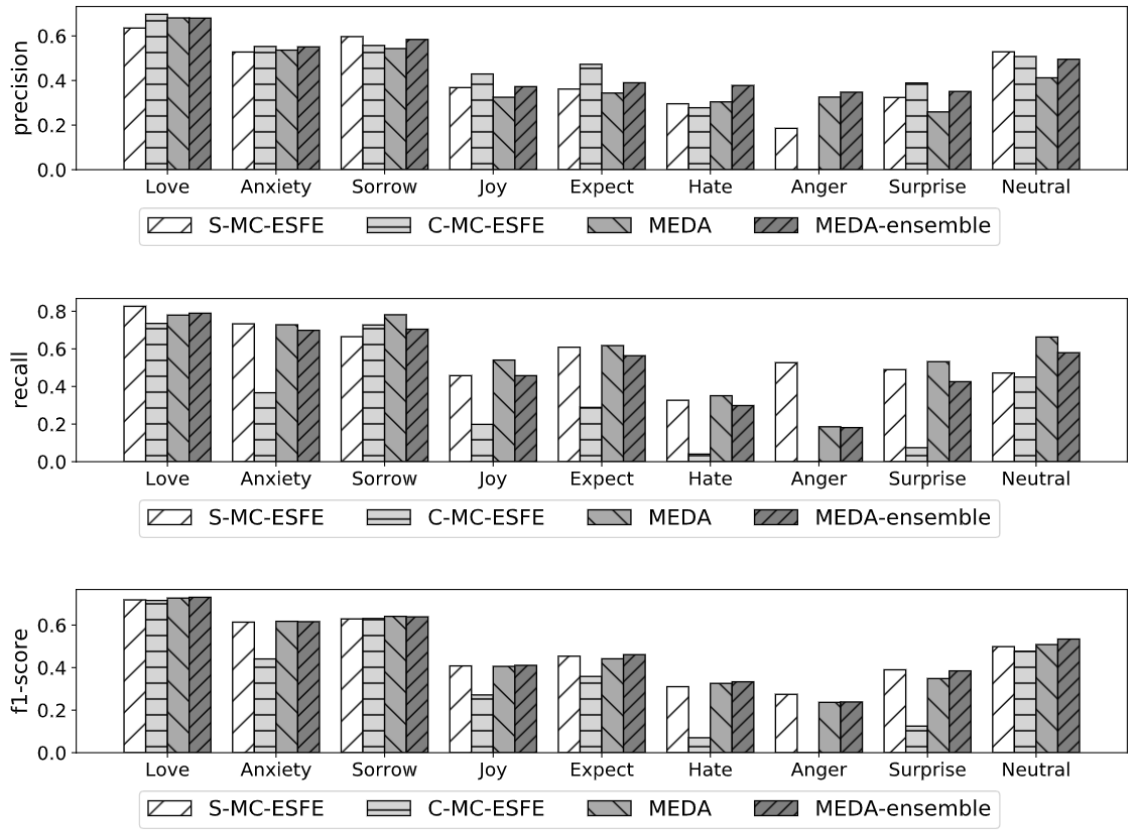

Figure 5.3 Comparison results of sub-models on RenCECps.

MEDA: As the global performance shown in Table 5.2, MEDA (micro-F1 = 59.71%, HL = 0.1378) outperforms the previous state-of-the-art model SGM-IFC (micro-F1 = 58.60%, HL = 0.1613), and outperforms another two sub-models on micro-F1, AP and ranking loss. MEDA network consists of two modules. The first is MC-ESFE, which is a hierarchical network and extracted emotion-specified features from both sentence-level and context-level in each channel. This feature matrix is extracted from the under-layer and each dimension focused on a certain emotion, which could conclude more detailed emotion-specified information. Another is ECorL module, which learns more global semantic information and emotion correlations based on above emotion-specified features. These two modules enable MEDA to give emotion predictions based on context and emotion correlation information.

To verify whether the emotion correlation information is learned in MEDA, we visualize the emotional correlation coefficients matrix. It is calculated with Pearson product-moment correlation coefficients, which indicates the level to which two emotions vary together:

$$R_{ij} = cov(E_i,\ E_j)/\sigma E_i \cdot \sigma E_j \tag{5.19}$$

Where $E_i = [E_{1i}, E_{2i,} \dots E_{Ni}]$ and E_{ni} is the emotional intensity of emotion e_i in the nth sample. $cov(E_i,\ E_j)$ is the covariance of e_i and e_j, and σ is the standard deviation. Figure 5.4 and Figure 5.5 show the comparison of the actual correlation coefficients matrix on Ren-CECps and the predicted correlation coefficients matrix in MEDA model. We can observe that the distribution of positively/negatively related emotion pairs predicted in MEDA is similar to the real distribution on Ren-CECps. Taking ‘Love’ as an example. Figure 5.4 shows that in actual distribution, the most positively related emotion with ‘Love’ is ‘Joy’ (+0.20) while the most negatively related emotion is ‘Anxiety’ (-0.38). This means that emotions ‘Love’ and ‘Joy’ often occur together while ‘Love’ and ‘Anxiety’ rarely appear together. The above emotion correlation information can also be learned by MEDA: correlation coefficient of ‘Love’ and ‘Joy’ is +0.51 while ‘Love’ and ‘Anxiety’ is -0.65. Besides, there are some emotion pairs with emotion correlation that have been learned, such as ‘Love’ and ‘Sorrow’ (-0.39), ‘Anxiety’ and ‘Joy’ (-0.43), ‘Hate’ and ‘Anger’ (+0.40), etc. The results demonstrate the ability of emotion correlation learning in proposed MEDA.

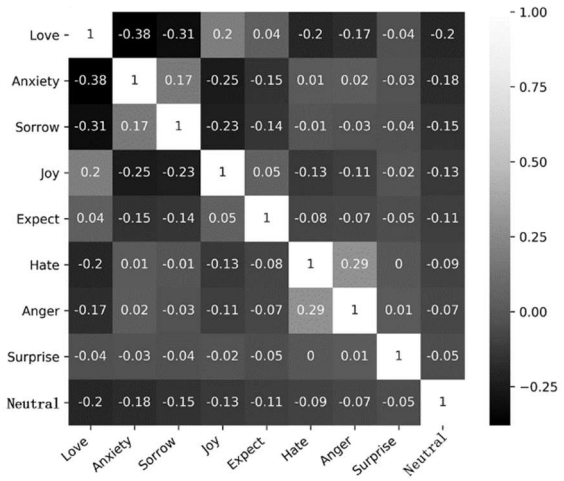

Figure 5.4 Emotional correlation coefficients matrix in RenCECp

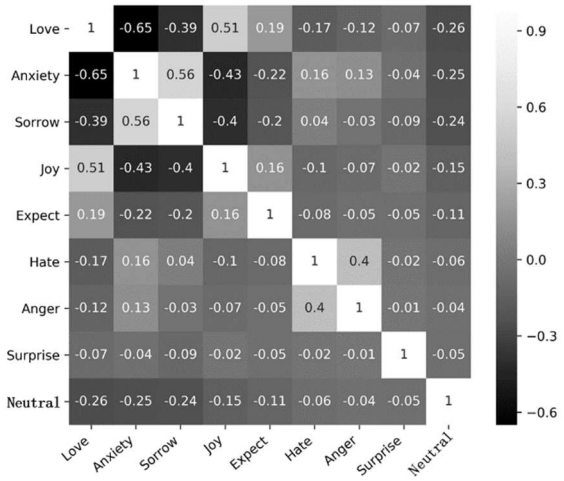

Figure 5.5 Emotional correlation coefficients matrix learned by MEDA

S-MC-ESFE: Results in Table 5.2 indicate that the prediction of S-MC-ESFE is better than baselines on most metrics. Compared with MEDA, S-MC-ESFE achieves a higher macro-F1 value (47.73% while 47.25%). Although this gap is small, it can reflect the average level of emotion detection of each emotion category in S-MC-ESFE. The higher macro-F1 of S-MC-ESFE suggests that for some sparse-resources emotion categories, it could give more accurate prediction than MEDA model. S-MC-ESFE is an intermediate model derived from MC-ESFE during sentence-level pre-training. In S-MC-ESFE, each channel is trained channel-wise and can be considered as multiple binary emotion classifier. During the training of each classifier, only the parameters of the corresponding channel are updated, which could make the model focus more on the feature extraction of a specified emotion. Take a sentence as an

example: ‘For a long time, I write about funny things in my blog, but this time, my heart is heavy.’ In the channel of ‘Joy’, feature extraction will pay more attention to the words ‘funny things’, while ‘Sorrow’ channel focused more on ‘heart is heavy’. Therefore, in each channel, the prediction of whether the sentence contains the corresponding emotion will be more accurate.

C-MC-ESFE: In C-MC-ESFE model, contextual information is further considered compared with S-MC-ESFE model. The results in Table 5.2 shows that both macro-F1 and micro-F1 are inferior to S-MC-ESFE. However, C-MC-ESFE achieves a better hamming loss (HL = 0.1213) than MEDA (HL = 0.1378). To further explore the role of C-MC-ESFE in MEDA-FS model, we further compared the micro/macro precision and recall of each sub-model. The results are shown in Table 5.4.

From Table 5.4, we can see that the lower F1-score of C-MC-ESFE mainly because of the lower recall during prediction. Its micro recall is 51.77% while S-MC-ESFE is 68.55%. Although its recall is lower, it can ensure that the prediction is more accurate: micro-precision of C-ESFE is 59.34% while S-MC-ESFE is 52.16%. This means that the prediction given by C-MC-ESFE model is more rigorous. Therefore, with higher precision, the C-MC-ESFE model improves the confidence for the final prediction of the ensemble model.

S-MC-ESFE, C-MC-ESFE, and MEDA mean different levels of information from sentence-level, context-level, and emotion correlation level. They are integrated into MEDA-FS and contribute to more accurate and stable predictions in emotion detection task.

5.4.3 Ablation Experiments

In the sentence-level embedding, we extract the emotional features based on the external emotional lexicon. To evaluate the effect of emotional features on experimental results, we train the model without this feature on RenCECps dataset. The experimental ablation results are shown in Table 5.5.

From Table 5.5, both MEDA model and MEDA-FS model with emotional features outperform the models without emotional features on almost all metrics. It is revealed that considering emotional features can make contributions to the classification improvement. In deep emotion recognition models, low-resource emotional datasets have been challenging,

and effectively incorporating existing emotional resources is the key to improving performance. In proposed MEDA, external emotional lexicon works as prior knowledge and is directly incorporated in sentence-level encoding. This method implements external knowledge supplementation in the simplest way and contributes to the effective extraction of emotional features.

Table 5.5 Ablation study on RenCECps dataset

	MEDA		MEDA-FS	
	With	Without	With	Without
Micro F1: %	**59.71**	56.22	**60.76**	57.33
Macro F1: %	**47.25**	42.91	**48.31**	44.16
AP: %	**75.76**	73.22	**76.51**	74.11
Hamming Loss	0.1378	**0.1342**	**0.1249**	0.1313
Coverage	**2.2369**	2.3703	**2.2226**	2.3322
One Error	**0.3763**	0.4101	**0.3618**	0.3959
Ranking Loss	**0.1084**	0.1239	**0.1062**	0.1189

'With' and 'without' denote with and without emotional features.

5.4.4 Discussion of Multi-Label Focal-Loss

In this section, we discuss the effectiveness of proposed multi-label focal loss (ML-FL) on emotion detection results. In the definition of multi-label focal loss, w is a harmonic factor aimed to balance the prediction of positive and negative labels. In this way, it has an effect on balancing the results of precision and recall, thus obtains an optimal F1 value. To verify the influence of w in emotion detection, we vary the value of w from 0. to 1. and compared it with two other commonly used loss functions: binary cross-entropy loss function(CE-loss) and multi-label loss function (ML-loss). The comparison experiments are implemented on RenCECps dataset, and the results are shown in Figure 5.6 and Table 5.6.

The results of CE-loss and ML-loss both show a higher recall (81.06% and 80.60% in micro-recall) while lower precision (41.95% and 44.39% in micro-precision). Precision is the average probability of relevant retrieval, while recall is the average probability of complete retrieval. They are two metrics restrain mutually [172]. In this emotion detection task, we hope to recognize as many emotions as possible, based on the premise of ensuring precision.

The comparison results of cross-entropy(CE), multi-label loss function(ML), and proposed multi-label focal loss with different weights are shown in Figure 5.6 The comparison results for cross-entropy(CE), multi-label loss function(ML), and proposed multi-label focal loss with different weights.a proper w can modulate the value between recall and precision, thus achieve both higher precision and F1-score to alleviate the above problems.

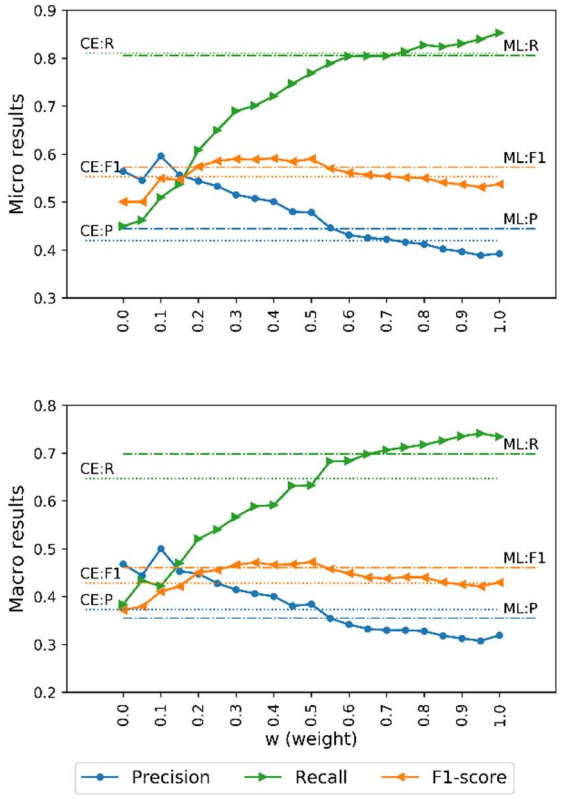

Figure 5.6 The comparison results for cross-entropy(CE), multi-label loss function(ML), and proposed multi-label focal loss with different weights.

Table 5.6 Results comparison of MEDA model with different loss functions

	Micro % (↑)			Macro: % (↑)			AP: %	HL	Coverage	OE	RL
	P	P	F1	P	P	F1	(↑)	(↓)	(↓)	(↓)	(↓)
CE-loss	41.95	**81.06**	55.29	37.26	64.71	42.79	70.64	0.1900	2.4496	0.4598	0.1349
ML-loss	44.39	80.60	57.25	35.49	**69.84**	46.07	72.27	0.1745	2.3652	0.4421	0.1251
ML-FL (w = 0.4)	**51.81**	70.46	**59.71**	**41.44**	57.54	**47.25**	**75.76**	**0.1378**	**2.2369**	**0.3763**	**0.1084**

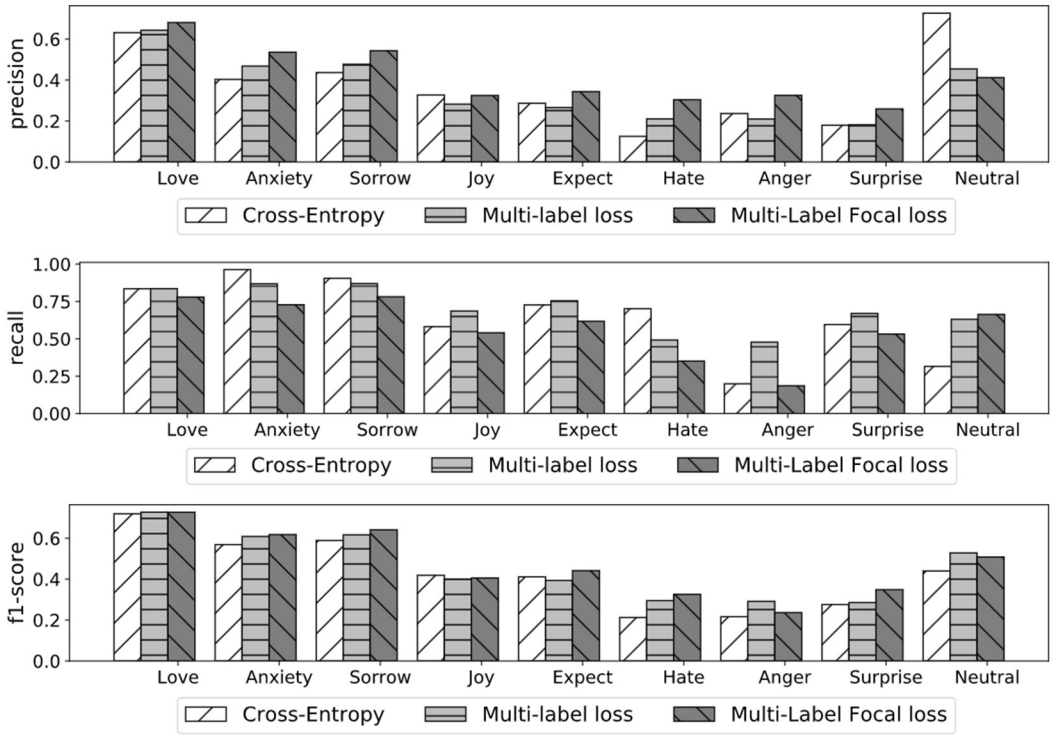

Figure 5.7 The comparison results of each emotion with different loss

We will analyze the role of the parameter w in the curve change. From the tendency of the curve in Figure 5.6 The comparison results for cross-entropy(CE), multi-label loss function(ML), and proposed multi-label focal loss with different weights. We can see that as the weight w increases, precision shows a downward trend, recall shows an upward trend while the overall trend of F1-score is to rise first and then fall. The loss function proposed in this chapter is committed to maximizing the prediction difference between positive-negative emotion pairs. The weighting factor α is dedicated to balancing the prediction of positive and negative labels, which aimed to recognize as many emotions as possible while ensuring the accuracy of prediction. The weight α is consists of two parts to control the prediction loss of positive and negative emotions, respectively:

$$\alpha_{pos}^{i} = w \cdot \left(1 - p_{k}^{i}\right)^{r}, \alpha_{neg}^{i} = (1 - w) \cdot (p_{l}^{i})^{r} \tag{5.20}$$

Considering the limit case, if harmonic factor w gradually increases to the maximum $w = 1$:

$$\alpha_{pos}^{i} \approx \left(1 - p_k^i\right)^r, \qquad \alpha_{neg}^{i} \approx 0 \tag{5.21}$$

In this case, as long as the model predicts all the emotion as $p^i = 1$, it is possible to minimize $\alpha_{pos}^{i} = 0$, thereby minimizing the loss. In this way, the prediction gap between positive and negative emotion pairs $exp(-(p_k^i - p_l^i)$ can only play a weak role. Therefore, the results show a higher recall while precision is difficult to be guaranteed: while $w = 1.0$, the micro-precision, recall, and F1-score are 39.22%, 85.29%, and 53.74%, respectively.

Conversely, as w gradually decreases, α_{neg}^{i} gradually increases. In this way, the prediction error of the negative label will bring greater losses. To reduce the loss, the model predicts the positive label more conservatively, and thus the recall decreased and precision could be guaranteed to some extent. Therefore, it can be assumed that by choosing an appropriate value of w, it is possible to reach a balance between precision and recall, and then achieve satisfactory results. As the results in Figure 5.6 The comparison results for cross-entropy(CE), multi-label loss function(ML), and proposed multi-label focal loss with different weights., take F1-score to measure the overall performance, while $w \in [0.3,0.5]$, proposed multi-label focal loss outperforms cross-entropy and multi-label loss function. To be specific, while $w = 0.4$, its micro-precision is 51.81%, micro- recall is 70.46%, micro-F1-score is 59.71%. Compared with ML-loss, although the recall drops, its micro-precision improved 7.42% and micro-F1 improved 2.46%. Table 5.6 shows the comparison results of different loss functions on multi-label metrics, which demonstrate that multi-label loss function outperforms others.

5.5 Summary

In this chapter, a Multiple-label Emotion Detection Architecture (MEDA) was proposed for the textual multi-label emotion detection task. MEDA was composed of two modules, its key idea was to capture emotion-specified features by MC-ESFE module in advance, and then learn emotion correlations based on above features in ECorL module. In MC-ESFE module, information of each emotion reflected in the text was separately encoded from sentence-level to context-level, which contributed a lot to underlying fundamental feature

extraction. In ECorL module, bidirectional-GRU network was utilized as emotion sequence predictor and emotion correlation learning was implemented among emotion-specified features. MEDA-FS integrated three sub-models derived from MEDA, and realized information fusion from sentence-level, context-level, and emotion correlation level. Furthermore, to incorporate emotion correlation information into model training, multi-label focal loss was proposed for multi-label learning. The proposed model achieved satisfactory performance and outperformed state-of-the-art models on both RenCECps and NLPCC2018 datasets, demonstrating the effectiveness of the proposed method for multi-label emotion detection.

Chapter 6

Conclusion and Future works

6.1 Conclusion

This thesis focuses on textual emotion recognition research and has proposed some related methods for perusing more accurate emotion prediction. Our work mainly revolves around some remaining challenges in this field: (1) the limitation of data imbalance, which is inevitable in the real-world database, (2) how to realize more accurately emotion prediction through contextual learning, (3) how to realize effective multi-label emotion recognition through emotion correlation learning. The concrete summaries and contributions are displayed as follows:

(1) An external background knowledge enhanced multi-stream neural network is proposed to address the limitations of data imbalance. The experimental results prove that the proposed network concisely and efficiently integrates external background knowledge, achieves information enhancement, and makes up for the neglected or missing information in the basic network.

(2) A hierarchical model with label embedding is proposed for contextual emotion recognition. The hierarchical model can effectively learn the emotional representation of a given sentence based on contextual information, while the label embedding matrix is conducive to realize emotion-correlation learning and emotion prediction. This method has the strong ability to learn emotional features for contextual emotion recognition, contributing to emotion correlation-based emotion prediction.

(3) A Multi-label Emotion Detection Architecture (MEDA) is proposed for multi-label learning and emotion correlation learning. MEDA can extract emotion-specified

features from sentence level to context level through a multi-channel hierarchical structure and realize emotion correlation learning and multi-label prediction through an emotion sequence predictor. Besides, the defined multi-label focal loss can make the model focus more on misclassified positive-negative emotion pairs, contributing to guarantee the overall performance.

6.2 Future Works

There is still much space for improvements in our works. To recognize all possible emotion labels in multi-label TER task, discernible feature representation of the weaker emotion category is a critical problem. This thesis has conducted some works to prevent weak emotion features from being covered by strong emotions to some extent, but not completely. Future work will try to explore more effective methods to recognize weak emotions more accurately. Abundant resources are the basis of neural network training, and external emotion resources can be severed as prior knowledge to enhance emotional feature representation. In future work, more emotional resources could be incorporated for better emotional understanding.

Furthermore, as the most extensive application scenario of the TER system, TER in dialogue has received continuous attention in NLP and affective computing. This thesis addressed some challenges, such as contextual-level emotion encoding, which also exists in the dialogue level TER system. Developing an intelligent dialogue-level emotion recognition system is worthy of further attention in future works, and some related sub-tasks, such as multi-party emotional interaction, personality modeling, and dynamic emotional tracking, can form new research directions.

Bibliography

[1] M.D. Munezero, C.S. Montero, E. Sutinen, and J. Pajunen. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, vol. 5, no. 2 pp. 101–111, Apr. 2014.

[2] A. Yadollahi, A.G. Shahraki, and O.R. Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–33, Jan. 2017.

[3] C. Strapparava, R. Mihalcea. Affect Detection in Texts 13. *The Oxford handbook of affective computing*, 2015.

[4] L.C. Zaragoza. Tackling the Challenge of Emotion Annotation in Text. *Ph.D. dissertation, Universitat d'Alacant-Universidad de Alicante*, Jun.2018.

[5] M.D. Choudhury, M.G. Scott, and C.E. Horvitz. Predicting depression via social media. *Seventh International AAAI Conference on Weblogs and Social Media*, pp. 128-137, Jun. 2013.

[6] S.A. Golder and M.W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, vol. 333, no. 6051, pp. 1878-1881, Sep. 2011.

[7] S. Kim, J. Lee, G. Lebanon, and H. Park. Estimating temporal dynamics of human emotions. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 168-174, Jan. 2015.

[8] E. Cambria. Affective computing and sentiment analysis. *IEEE intelligent systems*, vol. 31, no. 2, pp. 102–107, Mar. 2016.

[9] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, vol. 8, no. 1, pp.28, 2018.

[10] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya. Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, vol. 139, Jan. 2020.

[11] E.D. Aviles, C.O. Rodriguez, and W. Nejdl. Taking the pulse of political emotions in latin america based on social web streams. *2012 Eighth Latin American Web Congress, IEEE*, pp. 40–47, 2012.

[12] M. Anjaria and R.M.R. Guddeti. Influence factor based opinion mining of Twitter data using supervised learning. *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), IEEE*, pp. 1–8, 2014.

[13] C.S. Montero and J. Suhonen. Emotion analysis meets learning analytics: online learner profiling beyond numerical data. *in: Proceedings of the 14th Koli Calling International Conference on Computing Education Research, ACM*, pp. 165–169, 2014.

[14] Y Zhang, N Zhang, L Si, Y Lu, Q Wang, and X. Yuan. Cross-Domain and Cross-Category Emotion Tagging for Comments of Online News. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM*, pp. 627-636, 2014.

[15] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Strítesky, and A. Holzinger. Reprint of: Computational approaches for mining user's opinions on the Web 2.0. *Information Processing &*

Management, vol. 51, no.4, pp. 510-519, Jul. 2015.

[16] S.G. Kim and J. Kang. Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Information Processing & Management*, vol. 54, no. 6, pp. 938-957, 2018.

[17] S.H. Ang, S.Y.M. Low. Exploring the dimensions of ad creativity. *Psychology & Marketing*, vol. 17, no. 10, pp. 835-854, 2000.

[18] A.Z. Syed. Applying sentiment and emotion analysis on brand tweets for digital marketing. *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), IEEE*, pp. 1-6, Nov. 2015.

[19] C.O. Alm, D. Roth, R. Sproat. Emotions from text: machine learning for text-based emotion prediction. *Proceedings of the conference on human language technology and empirical methods in natural language processing*, ACL, pp. 579-586, 2005.

[20] E. Gilbert and K. Karahalios. Widespread worry and the stock market. *Fourth International AAAI Conference on Weblogs and Social Media*, pp. 58-65, 2010.

[21] P. Ormerod, R. Nyman, and D. Tuckett. Measuring financial sentiment to predict financial instability: A new approach based on text analysis, *arXiv preprint arXiv :1508.05357*, 2015.

[22] S.M. Mohammad, X. Zhu, and S. Kiritchenko. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, vol. 51, no.4, pp. 480-499, 2015.

[23] A. Aljanaki, F. Wiering, and R.C. Veltkamp. Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management*, vol. 52, no.1, pp. 115-128, 2016.

[24] K. Choi, G. Fazekas, M. Sandler, and K. Cho. Transfer learning for music classification and regression tasks. 2017, *arXiv preprint arXiv:1703.09179*.

[25] S. Brilis, E. Gkatzou, A. Koursoumis, and K. Talvis. Mood classification using lyrics and audio: a case-study in Greek music. *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, Berlin, Heidelberg, pp. 421–430, 2012.

[26] K.C. Wu. Affective surfing in the visualized interface of a digital library for children. *Information processing & management*, vol. 51, no. 4, pp. 373-390, 2015.

[27] T. Gossen and A. NüRnberger. Specifics of information retrieval for young users: A survey. *Information Processing & Management*, vol. 49, no. 4, pp. 739-756, 2013.

[28] Y.Z. Hao, Q.H. Zheng, Y.P. Chen, and C.X. Yan. Recognition of abnormal behavior based on data of public opinion on the web. *Journal of Computer Research and Development,* vol. 53, no.3, pp. 611-620, 2016.

[29] K. Zahra, M. Imran, and F.O. Ostermann. Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, vol. 57, no. 1, pp. 102107, 2020.

[30] D.D. Luxton, J.D. June, and J.T. Kinn. Technology-based suicide prevention: current applications and future directions. *Telemedicine and e-Health*, vol. 17, no. 1, pp. 50-54, 2011.

[31] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: emotional conversation generation with internal and external memory. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, arXiv preprint arXiv:1704.01074.

[32] X. Sun, J. Li, and J. Tao. Emotional Conversation Generation Orientated Syntactically Constrained Bidirectional-Asynchronous Framework. *IEEE Transactions on Affective Computing*, Jun. 2019.

[33] X. Zhou, X. Wan, and J. Xiao. Collective opinion target extraction in Chinese microblogs. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1840-1850, Oct. 2013.

[34] S. Jeong, D.E. Logan, M.S. Goodwin, S. Graca, and B. O'Connell. A social robot to mitigate stress, anxiety, and pain in hospital pediatric care. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ACM, pp. 103-104, 2015.

[35] H. Ai, D.J. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. *Ninth International Conference on Spoken Language Processing*, pp. 797–800, Sep. 2006.

[36] P. Ekman. An argument for basic emotions. *Cognition & emotion*, vol. 6. no. 3-4, pp. 169-200, 1992.

[37] W.G. Parrott. Emotions in social psychology: Essential readings. *Psychology Press*, 2001.

[38] X. Li, G.M. Lu, J.J. Yan, and Z.Y. Zhang. A Survey of Dimensional Emotion Prediction by Multimodal Cues. *Acta Automaica Sinica,* vol. 44, no. 12, pp. 2142-2159, Dec. 2018.

[39] A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, vol. 14, no. 4, pp. 261-292, 1996.

[40] S. Arifin and P.Y.K. Cheung. Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325-1341, 2008.

[41] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.

[42] R. Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*. Academic press, pp. 3-33, 1980.

[43] R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, vol. 89, no. 4, pp. 344-350, 2001.

[44] E. Cambria, A. Livingstone, and A. Hussain. The hourglass of emotions. *Cognitive behavioural systems*, Springer, Berlin, Heidelberg, pp.144-157, 2012.

[45] C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 70-74, 2007.

[46] E.S. Dan-Glauser and K.R. Scherer. The difficulties in emotion regulation scale (DERS). *Swiss Journal of Psychology*, 2012.

[47] Z. Wang, S. Li, F. Wu, Q. Sun, and G. Zhou. Overview of NLPCC 2018 Shared Task 1: Emotion Detection in Code-Switching Text. *CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, Cham, 2018.

[48] C.O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing,* ACL, pp. 579-586, 2005.

[49] C. Quan and F. Ren. A blog emotion corpus for emotional expression analysis in Chinese. *Computer Speech & Language*, vol. 24, no. 4, pp. 726-749, 2010.

[50] C. Quan and F. Ren. Sentence emotion analysis and recognition based on emotion words using Ren-CECps. *International Journal of Advanced Intelligence*, vol. 2, no.1, pp. 105-117, 2010.

[51] C. Strapparava and A. Valitutti. Wordnet affect: an affective extension of wordnet. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, vol. 4, pp. 1083-1086, May. 2004.

[52] J.C.D. Albornoz, L. Plaza, P. Gervás. SentiSense: An easily scalable concept-based affective lexicon for Sentiment Analysis. *LREC*, pp. 3562-3567, 2012.

[53] S.M. Mohammad and P.D. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pp. 26-34, Jun. 2010.

[54] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of LIWC2015, 2015.

[55] J. Li and F. Ren. Creating a Chinese emotion lexicon based on corpus Ren-CECps. *2011 IEEE International Conference on Cloud Computing and Intelligence Systems,* Sep. 2011.

[56] I.V. Willegen, L. Rothkrantz, and P. Wiggers. Lexical affinity measure between words. *Proceedings of International Conference on Text, Speech and Dialogue*, vol. 5729, pp. 234–241, 2009.

[57] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni. Emotion recognition from text based on automatically generated rules. *2014 IEEE International Conference on Data Mining Workshop*, pp. 383–392, Dec. 2014.

[58] F.R. Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. *In Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics*, pp. 422–425, 2007.

[59] L. Buitinck, J.V. Amerongen, E. Tan , and M.D. Rijke. Multi-emotion detection in user-generated reviews. *In European Conference on Information Retrieval*, pp. 43-48, Mar. 2015,

[60] S. Chaffar, D. Inkpen. Using a heterogeneous dataset for emotion analysis in text. *Advances in Artificial Intelligence*, pp. 62–67, 2011.

[61] T. Danisman, A. Alpkocak. Feeler: emotion classification of text using vector space model. *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 1, p.53-59, 2008.

[62] S.M. Mohammad, and P.D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, vol. 29, no. 3, pp. 436-465, Sep. 2012.

[63] M. Purver and S. Battersby. Experimenting with distant supervision for emotion lassification. *Proceedings of the 13th Conference of the European Chapter f the Association for Computational Linguistics*, pp. 482-491, 2012.

[64] S. Aman and S. Szpakowicz. Using roget's thesaurus for fine-grained emotion recognition. *Proceedings of the Third International Joint Conference on Natural Language Processing*, Vol.1, 2008.

[65] P.M. Roget. Roget's Thesaurus of English Words and Phrases. *TY Crowell Company*, 1911.

[66] F.M. Plaza-del-Arco, M.T. Martin-Valdivia, L.A. Ure-na-Lopez, R. Mitkov. Improved emotion

recognition in Spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems*, Sep. 2019.

[67] P.J. Stone, D.C. Dunphy, and M.S. Smith. The General Inquirer: A Computer Approach to Content Analysis, The MIT Press, Cambridge, 1966.

[68] S. Aman and S. Szpakowicz. Identifying expressions of emotion in text. *International Conference on Text, Speech and Dialogue*. Springer, Berlin, Heidelberg, 2007, pp. 196-205.

[69] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. *Proceedings of the 2008 ACM symposium on Applied computing.* ACM, 2008, pp. 1556-1560.

[70] O. Araque, G. Zhu, C.A. Iglesias. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, vol. 165, pp. 346-359, 2019.

[71] F. Ren and N. Liu. Emotion computing using Word Mover's Distance features based on Ren_CECps. *PloS one*, vol. 13, no. 4, pp. e0194136, 2018.

[72] H. Liu, H. Lieberman and T. Selker. A model of textual affect sensing using real-world knowledge. *In Proceedings of the ACM Conference on Intelligent User Interfaces*, pp. 125–132, 2003.

[73] P. Singh. The public acquisition of commonsense knowledge. *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*. 2002.

[74] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Narrowing the Social Gap among People Involved in Global Dialog: Automatic Emotion Detection in Blog Posts. *ICWSM*, 2007.

[75] Y.S. Seol, D.J. Kim, and H.W. Kim. Emotion recognition from text using knowledge-based ANN. ITC-CSCC: International Technical Conference on Circuits Systems, *Computers and Communications*, 2008.

[76] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis, *IEEE Intelligent systems*, vol. 28, no. 2, pp. 15–21, 2013.

[77] C. Quan and F. Ren. Weighted high-order hidden Markov models for compound emotions recognition in text. *Information Sciences*, vol. 329, pp. 581-596, 2016.

[78] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J.Deam. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111-3119, 2013.

[79] T. Mikolov, K. Chen, G. Corrado and J. Dean. Efficient estimation of word representations in vector space. 2013, arXiv preprint arXiv:1301.3781.

[80] J. Pennington, R. Socher and C. Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* pp. 1532-1543, 2014.

[81] J. Camacho-Collados and M.T. Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, vol. 63, pp. 743-788, 2018.

[82] B. McCann, J. Bradbury, C. Xiong and R.Socher. Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems*, pp. 6294-6305, 2017.

[83] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. 2018, arXiv preprint arXiv:1802.05365.

[84] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceeding of NAACL-HLT 2019*, pp. 4171-4186, Jun. 2019.

[85] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever. Improving language understanding by generative pre-training. available at https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf, 2018.

[86] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, vol. 1, no. 8, 2019.

[87] Z. Dai, Z. Yang, Y. Yang, W.W. Cohen, J. Carbonell Q.V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. 2019, arXiv preprint arXiv:1901.02860.

[88] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q.V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019, arXiv preprint arXiv:1906.08237.

[89] K. Song, X. Tan, T. Qin, J. Lu, and T.Y. Liu. Mass: Masked sequence to sequence pre-training for language generation. 2019, arXiv preprint arXiv:1905.02450.

[90] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H.W. Hon. Unified Language Model Pre-training for Natural Language Understanding and Generation. 2019, arXiv preprint arXiv:1905.03197.

[91] P. Xu, A. Madotto, C.S. Wu, J.H. Park, and P. Fung. Emo2vec: Learning generalized emotion representation by multi-task training. 2018, arXiv preprint arXiv:1809.04505.

[92] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel. emoji2vec: Learning emoji representations from their description. 2016, arXiv preprint arXiv:1609.08359.

[93] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. 2017, arXiv preprint arXiv:1708.00524.

[94] G.I. Winata, A. Madotto, Z. Lin, J. Shin, Y. Xu, P. Xu, and P. Fung. CAiRE_HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification. 2019, arXiv preprint arXiv:1906.04041.

[95] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1555–1565, 2014.

[96] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou. Coooolll: A deep learning system for twitter sentiment classification. *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 208-212, 2014, doi: 10.3115/v1/S14-2033.

[97]B. Shi, Z. Fu, L. Bing, and W. Lam. Learning domain-sensitive and sentiment-aware word embeddings. 2018, arXiv preprint arXiv:1805.03801.

[98] H. Meisheri and L. Dey. TCS Research at SemEval-2018 Task 1: Learning Robust Representations using Multi-Attention Architecture. *Proceedings of The 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pp. 291-299, 2018.

[99] H. Khanpour and C. Caragea. Finegrained emotion detection in health-related online posts. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pp. 1160–1166, 2018.

[100] M.S. Akhtar, D. Ghosal, and A. Ekbal. A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction. 2018, arXiv preprint arXiv:1808.01216.

[101] P. Agrawal and A. Suri. NELEC at SemEval-2019 Task 3: Think Twice Before Going Deep. 2019, arXiv preprint arXiv:1904.03223.

[102] P. Zhong, D. Wang, and C. Miao. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. 2019, arXiv preprint arXiv:1909.10681.

[103] A. Kumar, D. Kawahara, and S. Kurohashi. Knowledge-enriched two-layered attention network for sentiment analysis. 2018, arXiv preprint arXiv:1805.07819.

[104] C. Biemann and M. Riedl. Text: now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, vol. 1, no. 1, pp. 55-95, 2013.

[105] C Baziotis, N Athanasiou, A Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *Proceedings of The 12th International Workshop on Semantic Evaluation,* pp. 245–255, Jun. 2018.

[106] P. Goel, D. Kulshreshtha, P. Jain, and K.K. Shukla. Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 58-65, 2017.

[107] S.M. Mohammad and F. Bravo-Marquez. WASSA-2017 shared task on emotion intensity. 2017, arXiv preprint arXiv:1708.03700.

[108] A Chatterjee, U Gupta, MK Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, vol. 93, pp. 309-317, Apr. 2019.

[109] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efcient text classification. 2016, arXiv preprint arXiv:1607.01759.

[110] H. Wang. ReNN: Rule-embedded Neural Networks. *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 824-829.

[111] Q. Xie, X. Ma, Z. Dai and E. Hovy. An interpretable knowledge transfer model for knowledge base completion. 2017, arXiv preprint arXiv:1704.05908.

[112] Q. Qian, M. Huang, J. Lei, and X. Zhu. Linguistically regularized lstms for sentiment classification. 2016, arXiv preprint arXiv:1611.03949.

[113] K. Vo, D. Pham, M. Nguyen, T. Mai, and T. Quan. Combination of domain knowledge and deep learning for sentiment analysis. *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*. Springer, Cham, 2017, pp. 162-173.

[114] K. Vo, T. Nguyen, D. Pham, M. Nguyen, M. Truong, T. Mai, and T. Quan. Combination of Domain Knowledge and Deep Learning for Sentiment Analysis of Short and Informal Messages on Social Media. 2019, arXiv preprint arXiv:1902.06050.

[115] R. Gupta, S. Sahu, C. Espy-Wilson. Semi-supervised and transfer learning approaches for low resource sentiment classification. *2018 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP). IEEE, 2018.

[116] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2009.

[117] G Wiedemann, E Ruppert, R Jindal, and C. Biemann. Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. 2018, arXiv preprint arXiv:1811.02906.

[118] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, vol. 115, no. 2018, pp. 24-35, 2018.

[119] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel , and H. Prendinger. Decision support with text-based emotion recognition: Deep learning for affective computing. 2018, arXiv preprint arXiv:1803.06397.

[120] A. Chronopoulou, A. Margatina, C. Baziotis, and A. Potamianos. NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification. 2018, arXiv preprint arXiv:1809.00717.

[121] J. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1097-1102, 2018.

[122] SL Smith, DHP Turban, S Hamblin, and N.Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. 2017, arXiv:1702.03859v1.

[123] M. Yang, W. Yin, Q. Qu, W. Tu, Y. Shen, and X. Chen. Neural Attentive Network for Cross-Domain Aspect-level Sentiment Classification. *IEEE Transactions on Affective Computing*, Feb. 2019.

[124] J. Gideon, M. McInnis, and E.M. Provost. Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG). *IEEE Transactions on Affective Computing*, May 2019.

[125] S. Ruder. An overview of multi-task learning in deep neural networks. 2017, arXiv preprint arXiv:1706.05098.

[126] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. in *Proc. 25th Int. Conf. on Machine learning, ACM*, 2008, pp. 160–167.

[127] B. McCann, N.S. Keskar, C. Xiong, and R. Socher. The natural language decathlon: Multitask learning as question answering. 2018. [Online]. Available: https://arxiv.org/pdf/1806.08730.pdf

[128] G. Balikas, S. Moura, and M.R. Amini. Multitask Learning for Fine-Grained Twitter Sentiment Analysis. in *Proc. 40th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval (SIGIR '17),* Aug. 2017, pp. 1005-1008.

[129] X Wang, M Peng, L Pan, M Hu, C Jin, F Ren. Two-level attention with two-stage multi-task learning for facial emotion recognition. 2018. [Online]. Available: *https://arxiv.org/abs/1811.12139*

[130] W. Gao, S. Li, S.Y.M. Lee, G. Zhou, and C.R. Huang. Joint learning on sentiment and emotion classification. in *Proc. 22nd ACM Int. Conf. on Inf. & Knowl. Manage. ACM*, 2013, pp. 1505-1508.

[131] D. Wu and J. Huang. Affect estimation in 3D space using multi-task active learning for regression. *IEEE Trans. Affect. Comput.*, early access, May 2019, doi: 10.1109/TAFFC.2019.2916040.

[132] C. Huang, A. Trabelsi, O.R. Zaïane. ANA at SemEval-2019 Task 3: Contextual Emotion detection in Conversations through hierarchical LSTMs and BERT. in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019)*, Jun. 2019, pp. 49–53.

[133] A. Basile, M.F. Salvador, N. Pawar, S. Stajner, M. C. Rios, and Y. Benajiba. SymantoResearch at SemEval-2019 Task 3: Combined Neural Models for Emotion Classification in Human-Chatbot Conversations. in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019)*, 2019, pp. 330–334.

[134] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang. Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network. in *Proc. Twenty-Seventh Int. Joint Conf. on Artif. Intell. (IJCAI 18),* 2018, pp. 4595-4601.

[135] S. Zhu, S. Li, and G. Zhou. Adversarial Attention Modeling for Multi-dimensional Emotion Regression. in *Proc. 57th Annu. Meeting of the Assoc. for Comput. Linguistics*, 2019, pp. 471-480.

[136] I. Augenstein, S. Ruder, and A. Søgaard "Multi-task learning of pairwise sequence classification tasks over disparate label spaces. in *Proc. 2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, Jun. 2018, pp. 1896–1906.

[137] L. Qiu, Q. Lei, Z. Zhang. Advanced Sentiment Classification of Tibetan Microblogs on Smart Campuses Based on Multi-Feature Fusion. *IEEE Acces*s, 6, 17896–17904, 2018.

[138] A. Valdivia, M.V. Luzón, F. Herrera. Sentiment analysis in tripadvisor. *IEEE intelligent systems*, *32*, pp. 72–77, 2017.

[139] M. Bouazizi, T. Ohtsuki. A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access,* 2017, *5*, pp. 20617–20639.

[140] X. Du, L. Deng, K. Qian. Current Market Top Business Scopes Trend—A Concurrent Text and Time Series Active Learning Study of NASDAQ and NYSE Stocks from 2012 to 2017. *Applied Sciences*, *8*, 2076–3417*,* 2018.

[141] F.J. Castellanos, J.J. Valero-Mas, J. Calvo-Zaragoza, J.R. Rico-Juan. Oversampling imbalanced data in the string space. *Pattern Recognition Letters*, *103*, 32–38*,* 2018.

[142] Y. Li, H. Guo, Q. Zhang, M. Gu, J. Yang. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowledge-Based Systems*, *160*, 1–15, 2018.

[143] S. Tan. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems With Applications*, *28*, 667–671, 2005.

[144] P. Zhou, X. Hu, P. Li, X. Wu. Online feature selection for high-dimensional class-imbalanced data. *Knowledge-Based Systems*, *136*, 187–199, 2017.

[145] F. Ren, M.G. Sohrab. Class-indexing-based term weighting for automatic text classification. *Information Sciences,* 2013, *236*, 109–125.

[146] X. Luo, H. Li, D. Cao, Y. Yu, X. Yang, T. Huang. Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks. *Automation in Construction,* 2018, *94*, 360–370.

[147] J. Chung, C. Gulcehre, K. Cho, Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *Available online: https://arxiv.org/abs/1412.3555* (accessed on 11 December 2014).

[148] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. 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*, pp. 1480-1489, 2016.

[149] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks, *arXiv preprint arXiv:1503.00075*, 2015.

[150] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*, 2016.

[151] J. Lee, H. Kim, N. Kim, and J.H. Lee. An approach for multi-label classification by directed acyclic graph with label correlation maximization. *Information Sciences*, 351, pp. 101-114, 2016.

[152] D. Zhou, Y. Yang, and Y. He. Relevant emotion ranking from text constrained with emotion relationships, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 561-571, Jun. 2018

[153] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng. Emotion distribution learning from texts. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 638-647, Nov. 2016.

[154] Y. Wang, and P. Aditya. Detecting emotions in social media: A constrained optimization approach. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 996–1002, Jun. 2015,

[155] M. Zhang, Z. Zhou. Ml-knn: a lazy learning approach to multi-label learning. *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[156] M.L. Zhang, and Z.H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp.1819–1837, 2013.

[157] W. Liang, H. Xie, Y. Rao, R.Y.K. Lau, and F.L. Wang. Universal affective model for Readers' emotion classification over short texts. *Expert Systems With Applications*, vol. 114, pp. 322-333, Dec. 2018.

[158] F. Ren and K. Matsumoto. Semi-Automatic Creation of Youth Slang Corpus and Its Application to Affective Computing. *IEEE Transactions on Affective Computing*, vol.7, no.2, pp. 176-189, 2016.

[159] A. Bandhakavi, N. Wiratunga, and D. Padmanabhan. Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, vol. 93, no. 1, pp. 133-142, Jul. 2017.

[160] S.M. Liu, and J.H. Chen. A multi-label classification based approach for sentiment classification. *Expert Systems With Applications*, vol. 42, no. 3, pp. 1083-1093, Feb. 2015.

[161] N. Colneriĉ, and J. Demsar. Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. *IEEE Transactions on Affective Computing*, pp. 1949-3045, Feb. 2018.

[162] D. Pan, and J. Nie. Mutux at SemEval-2018 Task 1: Exploring Impacts of Context Information On Emotion Detection. *In Proceedings of the 12th international workshop on semantic evaluation*, pp. 345-349, Jun. 2018.

[163] H. Huihui, and R. Xia. Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification. *Nat. Lang. Process. and Chin. Compu. (NLPCC 2018)*, pp. 250-259, Aug. 2018.

[164] T.Y Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. *In*

Proceedings of the IEEE international conference on computer vision, pp. 2980-2988 2017.

[165] M.L. Zhang, and Z.H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering,*, vol. 18, no. 10, pp. 1338-1351, 2006.

[166] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du. Analogical reasoning on chinese morphological and semantic relations. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 138–143, 2018.

[167] O. Luaces, J. Díez, J. Barranquero, and J.D. Coz. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, vol. 1, no. 4, pp. 303-313, 2012.

[168] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, vol. 85, no. 3, pp. 333-359, 2011.

[169] R. Johnson, and T. Zhang. Deep pyramid convolutional neural networks for text categorization. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 562-570, 2017.

[170] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang. SGM: sequence generation model for multi-label classification. *In Proceedings ofthe 27th International Conference on Computational Linguistics*, pp. 3915-3926, Aug. 2018.

[171] W. Liao, Y. Wang, Y. Yin, X. Zhang, and P. Ma. Improved sequence generation model for multi-label classification via CNN and initialized fully connection. *Neurocomputing*, vol. 382, no.21, pp. 188-195, Mar. 2020.

[172] D.M. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of machine learning technologies*, vol. 2, no.1, pp. 37-63, Dec. 2011.