Multimodal Marketing Intent Analysis for Effective Targeted Advertising

Lu Zhang, Jialie Shen, Jian Zhang, *Senior Member, IEEE,* Jingsong Xu, *Member, IEEE,* Zhibin Li, Yazhou Yao, and Litao Yu

***Abstract*—People's daily information sharing and acquisition through the Internet has become more and more popular. The comprehensive multimodal marketing advertorial generated by 'We Media' accounts besides the normal social news is gaining its importance on social media platforms. In order to achieve effective advertising, the marketing intent understanding is a key step towards generating targeted advertising strategies (push advertorials to specific people at a specific time). However, advertorials in real are usually designed to pretend as normal social news with a wide range of contents. This poses big challenges to the platforms on accurately recognizing and analyzing the marketing intents behind the advertorials. As a pioneering study, we address this new problem of multimodal-based marketing intent analysis and answer three core questions: (1) does a piece of social news contain marketing intent? (2) what is the topic of marketing intent? (3) what is the extent of marketing intent? Towards this end, we propose a novel Multimodal-based Marketing Intent Analysis scheme (MMIA) to estimate the marketing intent embedded in the multimodal contents. Specifically, a novel supervised neural autoregressive model (SmiDocNADE) is proposed to enhance the discriminative capacity of the learned hidden features so that a single system is capable of solving the three questions. In order to effectively model inter-correlations between images and text in advertorials, we fuse multimodal data and extract features by Graph Convolution Networks as an enhancement to SmiDocNADE. The extensive evaluations demonstrate the advantages of our proposed system in multimodal-based marketing intent analysis from multiple aspects.**

***Index Terms*—multimodal, marketing intent analysis, targeted advertising**

I. Introduction

THE growing pervasiveness of the Internet media platforms has dramatically changed the way how people disseminate and acquire information. A recent survey found that over half (51%) of online social news consumers across 26 countries leverage various media platforms to access news and different kinds of other information [1]. 86% of marketers believe the Internet media platforms are important components of their marketing initiatives [2], [3]. The corresponding platforms have been emerging as novel channels to support effective promotion with 'We Media'-generated or user-generated content. Advertorial is such a kind of content published under a commercial arrangement between a promoter or sponsor of products/services and the publisher [4]. Surprisingly, according to The Global State of Digital Report in 2019, only 27% of media platform user respondents will be prompted to purchase products after experiencing the advertorials [5]. This not only demonstrates weak consumers' interest in the products but also reveals the importance and necessity of targeted advertising to specific persons or social groups (communities) during specific time periods. One of the typical examples is that a sauce

Fig. 1. Advertorial examples from media platforms. The above one is a piece of normal social news that shares knowledge about running. The middle one shows a biscuit advertorial with marketing intent embedded mainly in images. The below one shows a clothing advertorial with marketing intent embedded mainly in text.

Corresponding author: Jian Zhang. Lu Zhang, Jian Zhang, Jingsong Xu, Zhibin Li, and Litao Yu are with the Faculty of Engineering and Information Technology, University of Technology Sydney, NSW, 2007 Australia (e-mail: {Jian.Zhang; Jingsong.Xu; Litao.Yu}@uts.edu.au; {Lu.Zhang-5; Zhibin.Li;}@student.uts.edu.au). Jialie Shen is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom (e-mail: jialie@gmail.com). Yazhou Yao is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 210014 China (e-mail: Yazhou.yao@njust.edu.cn).

advertorial showed before dinner time may attract significant attention and recognition than a fertilizer advertorial.

In order to develop effective targeted advertising strategies, we study a novel problem of *multimodal-based marketing intent recognition and analysis*. Generally, marketing intent refers to the willingness of delivering value or benefits to potential consumers that makes the products/services desirable or interesting [6]. For example, Nutrition claims such as 'low in fat' or 'source of calcium'. Marketers post content with marketing intents to promote products/services, build customer relationships, or improve brand awareness, visibility, and engagement in the aim of benefits [7], [6]. For the emerging advertorial marketing, marketing intent recognition and analysis remain challenging compared with traditional public media (e.g., TV, radio, newspaper and movie) due to three main reasons. Firstly, an advertorial is designed in the form of editorial content to resemble a piece of normal social news and looks like describing an object or a story [8], which brings difficulties for distinguishing. The marketing intents camouflage in regular social news with the same, relative, or different topics. For example, a writer may first report a cold air strike then promote a new arrival down jacket. They want to almost 'trick' the readers into thinking that the advertorial is a part of the existing content but engage them in interesting advertising contents [9]. Secondly, an advertorial usually contains multimodal data (e.g., images and text). For some advertorials, the text still holds the mainstream standing as a medium to deliver commercial information. For other advertorials, marketers' marketing intents are only spread by images. Under other situations, images and text both show the promotion information. Figure 1 demonstrates one piece of normal social news example and two advertorial examples from the platforms. Obviously, images in the middle example are product images presenting the a brand of biscuit as shown by the red rectangular. However, the topic of the text is British leisure style. People will not easily discover the marketing intent without images. The other advertorial example introduces the clothing-related promotional information by text, such as 'new great discount' and 'put on a beautiful dress'. However images are essentially a set of artistic photos. Thirdly, advertorials generated by users and 'We Media' accounts are diverse without a uniform format and writing style. Social news on media platforms includes not only local affairs but also how-tos and tips in diverse areas, including fitness, health, relationships, career, and so on.

In order to effectively recognize and analyze marketing intents in the social news, it is important and essential to necessitate answers to the following core research questions:

- **RQ1:** Does the social news contain marketing intent?
- **RQ2:** What is the topic of the marketing intent?
- **RQ3:** What is the extent of the marketing intent?

With the emergence of advertorials, these questions have always been important but surprisingly not addressed properly in previous research. We are the first to study the problem systematically. Main focus of existing research is on human intent detection [10], [11], [12], [13], [14], which involves a variety of applications such as video search, intelligent vehicle, email conversation, E-commerce search, and so on. Some works focus on commercial intent detection [15], [16], [17], [18], [19], [20]. The aim is to discover latent business opportunities by analysing buyers' commercial intents. However, they focus on buyer intent analysis on search engine, E-commerce web sites, or social medias through click-through, mouse tracking, scrolling behaviours, or search queries. Our research aims at developing an intelligent scheme for marketing intent analysis on media platforms, where marketers conceal their business purposes in normal social news. Compared with the existing works, the goal and tasks are distinguishing and fundamentally different. Another set of previous works [21], [22], [23] focus on the detection of marketing content on media platforms. However, their research is coarse-grained and is lack of the comprehensive extraction and the deep analysis of the problem. For example, the problem is coarsely defined as a simple text classification task without image feature involved [21], [22]. Furthermore, none of these studies focus on the analysis of the marketing topics and the extents. Toward this end, we propose a novel Multimodal-based Marketing Intent Analysis System (MMIA). To estimate the latent intent distribution embedded in the multimodal content, a Supervised Multimodal Document Informed Neural Autoregressive Distribution Estimator (SmiDocNADE) is developed to learn a joint representation from images, texts, and class label information. The label information is taken into account to enhance the discriminative power so that our system can solve the proposed three different questions. The feature extracted by a two-branch Graph Convolution Network (GCN) [23] is integrated as an enhancement to SmiDocNADE for effective inter-correlation mining between images and text. The extensive experimental results demonstrate the advantages of our proposed system from multiple aspects.

In summary, the key contributions of our work are as follows:

- To our best knowledge, this work is truly pioneering in terms of the first attempt to deeply explore the marketing intents of multimodal data with three questions. We believe this work will have a significant impact in the multimodal marketing intent detection related literature, especially given the fact that related applications in social media are flourishing.
- We propose a novel MMIA scheme. Specifically, a supervised multimodal SmiDocNADE model is developed to discover latent intents of multimodal data in the social news. Further, the features extracted from multimodal data are integrated as an enhancement to achieve comprehensive multimodal data fusion and correlation mining.
- We conduct extensive experiments to evaluate the performance of the proposed model. To facilitate large scale test, three test collections are developed and a set of experiments are carefully designed to evaluate and compare the performance of our system with main competitors over a wide range of settings. The core empirical results show that the proposed method performed the best compared with the others, highlighting the effectiveness of the proposed model.

The rest of the paper is organized as follows. Section 2 gives a literature review in the related areas. In Section 3, we introduce our proposed scheme, giving the detailed structure of its component modules and its learning algorithms. Section 4 introduces about the application and the dataset construction. Section 5 introduces the experimental configuration and analyzes experimental results. Finally, in Section 6, we conclude the paper with key results and findings discussion, and directions of future work.

II. Related work

In this section, we briefly review the recent works in three directions, which are closely related to our study: (1) Human Intent Analysis, (2) Topic Model, and (3) Graph Neural Networks.

A. Human Intent Analysis

Human intent analysis refers to recognize and distinct what people aim to do by inferring what they truly intend to do. It has been proposed and studied in a variety of applications. For example, visual information is used to detect human intent in human-object interaction [10]. The relationship between the photographer's intent and the viewer's attention is explored to improve image analysis and understanding [11]. A self-paced learning mechanism is proposed for email intent detection by leveraging user actions as a source of weak supervision [12]. The Markov decision process is adopted to formulate the user intent prediction for user intent prediction in customer service bots [13]. Some works focus on discovering latent commercial opportunities by analyzing buyers' consumption intents. For example, the identification of the corresponding products of the hot trend on social media is proposed [15]. The correlations of the potential sellers and buyers is analyzed by discovering the keywords and learning a classification model [16]. The relationships between query terms and advertisement click behavior are also investigated [20]. These works focus on human behavior analysis. Compared with these works, our task and the application scenario are totally different. We focus on marketing intent inferring based on multimodal data.

On the other hand, some works [21], [22], [23] focus on marketing information detection in media platforms. A stacking-based ensemble learning method (SBEL) is proposed to reflect semantic information of texts to identify marketing intent [21]. A graph-based approach (GBA) is proposed to extract the self-defined graph-related and community-related features to detect content marketing articles [22]. These works follow the traditional NLP approach and treat the problem of marketing content detection as simple text classification task. However, in reality, marketing intents are usually delivered by multimodal data. The aforementioned methods only focus on text but can not detect marketing information in images. They concentrate on marketing semantic detection but not on marketing intent inference. An influencer profiling approach is proposed recently to classify influencers with their marketing preference and further classify their posts using multimodal data into several categories [24]. A brand-based post popularity detection model is proposed [25] to predict popularity of multimodal post of specific brand. While our study is highly related to these two previous explorations, the key focus is different and mainly on multimodal marketing intent analytics with specific aims to support effective advertisement. The datasets they constructed also cannot be directly used in our work to evaluate the performance of our proposed method on solving these research questions. A Two-Branch Graph Neural Network method (TBCGNN) takes image and text information into account simultaneously [23]. However, the analysis of the topic and the extent of marketing intent is still ignored (RQ2 and RQ3). Table I gives the clear comparison results. The last row shows our proposed method MMIA. Compared with the other methods, the proposed MMIA focuses on revealing the essence of the research enquiry by answering the three questions. The marketing intent embedded in both texts and images are analyzed.

TABLE I: Comparison of related works.

Work	Modality		Task			Year
	Text	Image	RQ1	RQ2	RQ3	
SBEL [21]	✓	-	✓	-	-	2019
GBA [22]	✓	-	✓	-	-	2019
TBCGNN [23]	✓	✓	✓	-	-	2020
MMIA	✓	✓	✓	✓	✓	-

B. Sentiment Analysis

Sentiment analysis, also known as opinion mining is usually performed on text by natural language processing for affective states and subjective information extraction and identification. It has been successfully utilised on social media monitoring [26], [27], market research [28], customer service [29], user preference discovery [30] and so on. Recently sentiment analysis is utilised on visual information by deep neural networks. A weakly supervised coupled network is proposed for visual sentiment detection [31]. A joint image-text fusion method is developed for sentiment recognition on GIF video and its textual annotation [32]. A multimodal emotion analysis system is proposed to integrate multiple modalities to discover people's feelings and sentiments [33]. An attention-based unsupervised adversarial model is developed for movie review spam detection [34]. Works related to market research mainly focus on stock market prediction, price prediction, and financial market prediction and so on [35]. Some works focus on extracting customers' opinion (positive or negative) towards their purchased products [36]. These market-related works focus on market research such as market affective prediction which is similar but not the same with our proposed research problem. In our work, we focus on social news marketing intent analysis which is to identify whether a piece of news is actually an advertorial. We also focus on the marketing topic and extent of the marketing advertorial.

C. Neural Autoregressive Topic Models

Generative neural autoregressive topic models estimate the distribution of the document by maximizing the log-likelihood.

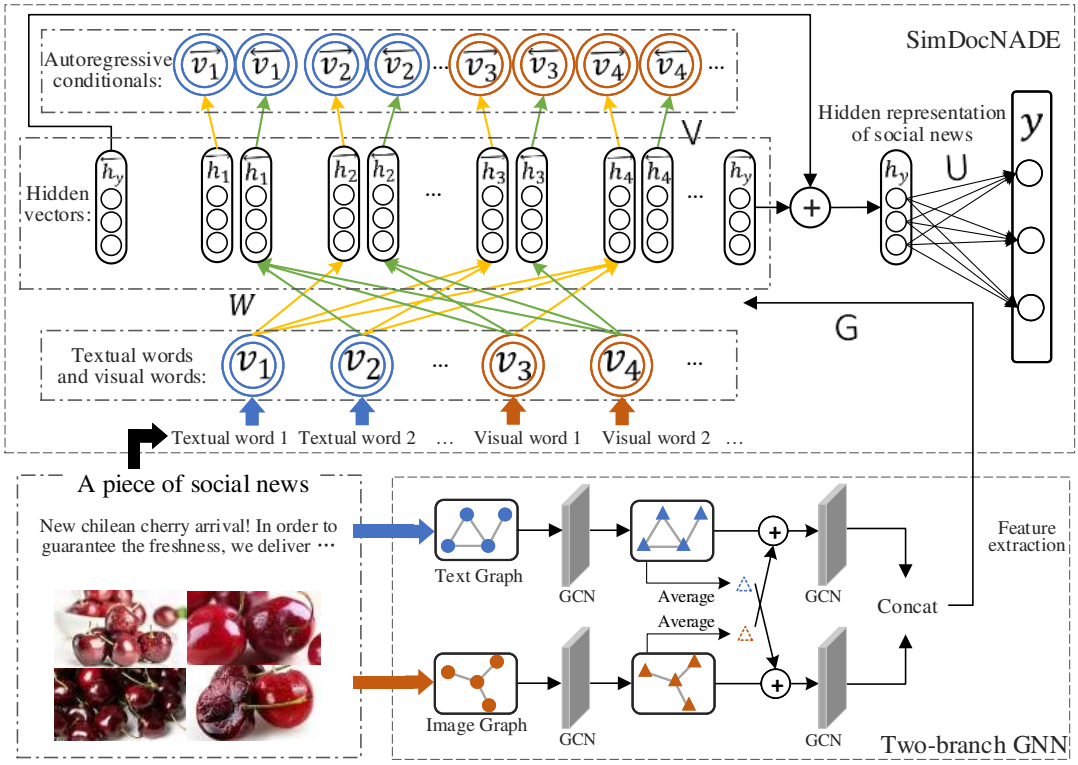


Fig. 2. Architecture of Multimodal-based Marketing Intent Analysis scheme. Text and images of the social news are first represented as several textual words and visual words. These words are fed into SimDocNADE as inputs. Then these text and images are represented as a text graph and an image graph. These graphs are used for feature extraction for Two-branch GNN. The feature extracted is integrated as an enhancement to SmiDocNADE.

DocNADE is a generative neural autoregressive to account for word counts [37], [38], [39] that relies on autoregressive neural network architecture. It defines the learning process as an estimation process of the probability of observing current words by the previously given words. iDocNADE [40] extends the DocNADE model to exploit the full context information. Previous works use iDocNADE to learn the representation of text documents by unlabeled document data. However, these methods are unsupervised models with single textual modalities which are not capable of dealing with multimodal data. With multimodal data becoming popular, some extensions of the basic unsupervised methods are proposed such as MDRF [41]. Based on DocNADE, a supervised extension SupDocNADE [38] and a deep extension of SupDocNADE [39] are proposed to model the multimodal data through deep neural networks. SupDocNADE focuses on learning a joint representation from three different modalities: image visual words, annotations, and class labels. However, the authors only consider the previous words but ignore the later contexts when inferring the hidden topic of a words. In comparison, our proposed SmiDocNADE is a supervised multimodal method by considering dual-direction contexts. This mechanism takes more contexts into account, which helps in determining the meaning and the intent of the words. Moreover, we fuse multimodal data and extract features as an enhancement of SmiDocNADE to effectively model inter-correlations between different modalities.

D. Graph Neural Networks

GNNs have shown their promising performance in a lot of applications [42], [43], [44], [45], because of the capability of directly operating on non-Euclidean data structures. GNNs use the graph structure and node features to learn a representation vector of a node or the entire graph. There are two kinds of GNNs: spectrum methods and non-spectrum methods. Graph Convolution Networks (GCNs) [46], [47] are one of the typical representation of spectrum methods. GCNs focus on local connections in a shared weight strategy by a multi-layer structure. Non-spectrum GNNs follow a neighborhood aggregation strategy, where the node representation is iteratively updated by aggregating representations of its neighbors [48]. The representative method is GraphSAGE [49], which generates embeddings by sampling and aggregating features from a node's local neighborhood. These methods can not be used directly on multimodal data fusion for inter-correlation mining. By a two-branch GCN with an image branch and a text branch, we can extract features from multiple modalities. DiffPool[50] and SAGPool[51] are two representative graph

pooling methods which can be combined with a lot of GNNs architectures.

III. Multimodal-based Marketing Intent Analysis System

Multimodal-based marketing intent analysis is very important in the emerging multimedia advertising. However, there is no standard way to conduct it. We propose a Multimodal-based Marketing Intent Analysis system (MMIA) to solve this problem. Inspired by the common assumption that textual and visual words are generated by a mixture of intents or topics, we propose a novel SmiDocNADE to estimate the latent intent distribution through the observable text and images of the social news. This supervised method is designed to increase the discriminative power of the model so that a single system is capable of solving the three proposed questions. Besides, a two-branch GCN [23] serves as an enhancement of SmiDocNADE to fuse arbitrary numbers of images and sentences and carefully mine inter-correlations between multiple modalities. Figure 2 illustrates the details of the system architecture, and we give a introduction in the Section III-B.

A. Definitions, Problem Formulation, and Notations

Marketing intent advertorial: Inspired by the definition on commercial intent [52] and the definition on advertorial by Australian Press Council [4], we define a piece of social news as a marketing intent advertorial (positive class in RQ1) if it contains the marketer's intent for commercial purpose within the normal news content. The intent can be explicitly or implicitly described.

Example: Advertorial *'Buy the new arrival Channel bag by clicking the following web link...'* is with an explicit marketing intent. Advertorial *'A Channel bag is all you need.'* is with an implicit marketing intent. In this advertorial, the marketer encourages the potential consumers to buy the bag by exaggerating its importance. Both of two advertorials express the marketing intents of marketers for the purpose of selling a bag, and they satisfy the criterion in the definition. Content *'Criminals who stole wealthy people were arrested for stealing ten Channel bags...'* is a piece of normal social news.

Topics of advertorial with marketing intent. Marketers' intents exhibit in advertorials may belong to different topics, and different topics of intents may be of interest to different potential consumers. No existing work has attempted to establish the topics for marketing intent advertorials. We define five topics of marketing intents: 'health', 'life', 'service', 'entertainment', and 'technology'. These topics are selected because they are important categories that are used to describe advertorials in the social news.

Extents of advertorial with marketing intent. The extent of marketing advertorials is the sentiment extent of marketing in marketers' expression. We quantify the extent of the marketing intents of each advertorial into three levels:

- **Weak** advertorials express very limited information about the promoting. The product names or logos appear in the social news which are highly related to the background content.
- **Medium** advertorials express the marketing intents to readers in a relatively soft way, which means readers may be more tolerant. The authors focus on describing the product properties and highlighting the benefits to customers, e.g., illustrating a badminton competition news with a picture of a brand of running shoes.
- **Strong** advertorials express their marketing intents to readers in a hard way, which means less tolerance from readers who are not interested in purchasing such products. The content usually contains 1) very clear purchase information such as the price of products, means to purchase, contact number, or address; 2) inflammatory emotion by using a specific expression such as 'last chance'; 3) activity direction such as 'go to buy', 'hurry up', 'click the web link', 'follow us', which means the marketer intents readers to take a kind of further action.

Problem formulation: main objective of the study is to model the marketing intent prediction with topic model generation and leverage the Neural Autoregressive networks to infer the marketers' intents. The core aim is to learn the function f from a piece of social news $\mathbf{v}$ to a vector indicating to 1) marketing intent advertorial and non-marketing intent advertorial for RQ1, 2) five topics of marketing intent advertorial for RQ2, and 3) three extents of marketing intent advertorial for RQ3.

Given:

1. $\mathbf{T}$: The text of a piece of social news.
2. $\mathbf{I}$: The images of a piece of social news.
3. $\mathbf{y}$: The label indicating vector.

Our goal is to find the function f which indicates a distribution:

$$f : [T, I] \rightarrow \mathbf{y} \tag{1}$$

The optimal parameters θ_o of the function f is found by minimizing the negative log-likelihood:

$$\theta_o = \arg\min_{\boldsymbol{\theta}} L(\mathbf{v}, y; \theta), \tag{2}$$

For RQ1, $\mathbf{y}$ is the one hot vector indicating whether a piece of social news contains marketing intent. For RQ2 and RQ3, the label vectors indicate the topic and extent class respectively.

Firstly, images and texts of a piece of social news are converted to a series of visual and textual words $\mathbf{v}$ as the observation values of the social news as shown in the Figure 2. The all notations and their definitions are summarised in Table II. $\mathbf{v} = [v_1^{(I)}, ..., v_{D_v}^{(I)}, v_{D_v+1}^{(S)}, ..., v_D^{(S)}]$ is of size D, where the first D_v observations are from images and the last $D - D_v$ observations are from text. We use a joint indexing of both visual and textual words. In order to simplify the formulation, we use v_i to indiscriminately represent a textual or a visual word, as shown in Figure 2. A vocabulary is a group of selected visual words and textual words. For text data, we select textual words with high frequency in the training dataset into the vocabulary. To select visual words, k-means is used to learn a cluster of SIFT [53], [54] descriptors densely. Each word $v_i \in \{1, ..., K\}$ is the index of the i^{th} word in the dictionary of vocabulary size K. These words are generated under a serious of latent intents, which means by the observation of words, the distribution of latent intents $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ can be inferred as shown in Figure 2. Marketing

intent is a kind of commercial intent, which is inferred by the observation of $\mathbf{v}$. The outputs $\overrightarrow{\mathbf{v}}_i$ and $\overleftarrow{\mathbf{v}}_i$ represent the autoregressive conditionals $p(v_i|\mathbf{v}_{<i})$ and $p(v_i|\mathbf{v}_{>i})$.

B. The MMIA System

To infer the hidden intent distribution of the social news for the proposed three questions, we propose a Supervised Multimodal Document Informed Neural Autoregressive Distribution Estimator model named SmiDocNADE. The SmiDocNADE incorporates label modality into the method, so that one model is capable of solving the three questions. To avoid repetition, the three questions are discussed without distinction.

Based on the probability chain rule, the joint probability of all observable words of a piece of social news can be decomposed as follows [37]:

$$p(\mathbf{v}) = \prod_{i=1}^{D} p(v_i|\mathbf{v}_{<i}), \tag{3}$$

where v_i is the i^{th} word, $\mathbf{v}_{<i}$ is the subvector of the first $i-1$ words. Conditional probability $p(v_i|\mathbf{v}_{<i})$ is the probability of the i^{th} word inferred by is the first $i-1$ words which is parameterized as follows:

$$\mathbf{h}_i(\mathbf{v}_{<i}) = \mathrm{g}(\mathbf{c} + \sum_{k<i} \mathbf{W}_{:,v_k}), \tag{4}$$

$$p(v_i = w|\mathbf{v}_{<i}) = \frac{\exp(b_w + \mathbf{V}_{w,:}\mathbf{h}_i(\mathbf{v}_{<i}))}{\sum_{w'} \exp(b_{w'} + \mathbf{V}_{w',:}\mathbf{h}_i(\mathbf{v}_{<i}))}, \tag{5}$$

where $\mathrm{g}(\cdot)$ is a non-linear activation function. $\mathbf{h}_i \in \mathbb{R}^H$ is the i_{th} latent intent distribution with total H intents estimated by $\mathbf{v}_{<i}$. $\mathbf{c} \in \mathbb{R}^H$, $W \in \mathbb{R}^{H\times K}$, $\mathbf{b} \in \mathbb{R}^K$, $V \in \mathbb{R}^{K\times H}$ are the parameters. W and V are the connection parameter matrices, $\mathbf{c}$ and $\mathbf{b}$ are bias parameter vectors.

To use the full context data around word v_i, the equation (4) is reformulated as a forward $\overrightarrow{\mathbf{h}}_i$ and a backward $\overleftarrow{\mathbf{h}}_i$[40]:

$$\overrightarrow{\mathbf{h}}_i(\mathbf{v}_{<i}) = \mathrm{g}(\overrightarrow{\mathbf{c}} + \sum_{k<i} \mathbf{W}_{:,v_k}), \tag{6}$$

$$\overleftarrow{\mathbf{h}}_i(\mathbf{v}_{>i}) = \mathrm{g}(\overleftarrow{\mathbf{c}} + \sum_{k>i} \mathbf{W}_{:,v_k}), \tag{7}$$

where $\overrightarrow{\mathbf{c}} \in \mathbb{R}^H$ and $\overleftarrow{\mathbf{c}} \in \mathbb{R}^H$ are the forward and backward bias parameter vectors. The equation (5) is reformulated as:

$$p(v_i = w|\mathbf{v}_{<i}) = \frac{\exp(\overrightarrow{b}_w + \mathbf{V}_{w,:}\overrightarrow{\mathbf{h}}_i(\mathbf{v}_{<i}))}{\sum_{w'} \exp(\overrightarrow{b}_{w'} + \mathbf{V}_{w',:}\overrightarrow{\mathbf{h}}_i(\mathbf{v}_{<i}))}, \tag{8}$$

$$p(v_i = w|\mathbf{v}_{>i}) = \frac{\exp(\overleftarrow{b}_w + \mathbf{V}_{w,:}\overleftarrow{\mathbf{h}}_i(\mathbf{v}_{>i}))}{\sum_{w'} \exp(\overleftarrow{b}_{w'} + \mathbf{V}_{w',:}\overleftarrow{\mathbf{h}}_i(\mathbf{v}_{>i}))}, \tag{9}$$

where $\overrightarrow{\mathbf{b}} \in \mathbb{R}^K$ and $\overleftarrow{\mathbf{b}} \in \mathbb{R}^K$ are the forward and backward bias parameter vectors. The log-likelihood function is:

$$\log p(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{D} \underbrace{\log p\left(v_i|\mathbf{v}_{<i}\right)}_{\text{forward}} + \underbrace{\log p\left(v_i|\mathbf{v}_{>i}\right)}_{\text{backward}}. \tag{10}$$

TABLE II: Notations and their definitions.

Notation	Definition
D_v, D	the number of visual words and total words
$\mathbf{v}$	a piece of social news
v_i	the index of the i^{th} word in vocabulary
$\mathbf{v}_{<i}$	subvector of the first $i-1$ words
K	the vocabulary size
H, L, $D_{\mathbf{e}}$	dimensions of hidden state, label, and feature vector
$\mathbf{y}$	the ground-truth label
$\mathbf{h}_i(\mathbf{v}_{<i})$	the i^{th} latent state
$\overrightarrow{\mathbf{h}}_i(\mathbf{v}_{<i})$	the forward i^{th} latent state
$\overleftarrow{\mathbf{h}}_i(\mathbf{v}_{>i})$	the backward i^{th} latent state
$\overrightarrow{\mathbf{v}}_i$, $\overleftarrow{\mathbf{v}}_i$	the autoregressive conditionals
W, V, G, U	connection parameter matrices
g, σ	activation functions
$W_{:,v_k}$	the v_k^{th} column of W
$V_{w,:}$	the w^{th} row of V
$\mathbf{c}$, $\mathbf{b}$, $\mathbf{f}$	bias parameter vectors
$\overrightarrow{\mathbf{c}}$, $\overrightarrow{\mathbf{b}}$, $\overleftarrow{\mathbf{c}}$, $\overleftarrow{\mathbf{b}}$	the forward and backward bias parameter vectors
C	the feature dimension of nodes in graph
$H_l^{(I)}$, $H_l^{(S)}$	outputs of the l^{th} layers of image, text branches
$\overline{h}_l^{(I)}$	image graph centre
$H_l^{(S)\prime}$	the updated output of the l^{th} layer in text branch
μ	an adjustive parameter
$\mathbf{e}$	the embedding multimodal feature vector
$\overrightarrow{\mathbf{h}}_y(\mathbf{v})$	the forward hidden representation of a piece of social news
$\overleftarrow{\mathbf{h}}_y(\mathbf{v})$	the backward hidden representation of a piece of social news
$\mathbf{h}_y(\mathbf{v})$	the backward hidden representation of $\mathbf{v}$

Considering that although the essence of RQ1, RQ2, and RQ3 is marketing intent analysis, in terms of specific content, they all have their own emphasis. To resolve the proposed three questions, we incorporate their different label information into SmiDocNADE to increase the discriminative power of the system on these specific tasks. Suppose given the class label $y \in \{1, ..., L\}$, the hidden representations of the social news in forward and backward networks are computed as:

$$\overrightarrow{\mathbf{h}}_y(\mathbf{v}) = \mathrm{g}(\overrightarrow{\mathbf{c}} + \sum_{k\leq D} \mathbf{W}_{:,v_k}), \tag{11}$$

$$\overleftarrow{\mathbf{h}}_y(\mathbf{v}) = \mathrm{g}(\overleftarrow{\mathbf{c}} + \sum_{k\geq 1} \mathbf{W}_{:,v_k}). \tag{12}$$

The final hidden representation is:

$$\mathbf{h}_y(\mathbf{v}) = \overrightarrow{\mathbf{h}}_y(\mathbf{v}) + \overleftarrow{\mathbf{h}}_y(\mathbf{v}), \tag{13}$$

which is used to perform classification. For modeling $p(\mathbf{y}|\mathbf{v})$ from $\mathbf{h}_{\mathbf{y}}(\mathbf{v})$, the architecture is:

$$p(y|\mathbf{v}) = \mathrm{softmax}(\mathbf{f} + \mathbf{U}\mathbf{h}_y(\mathbf{v}))_y, \tag{14}$$

where $\mathrm{softmax}(\mathbf{x})_i = \exp(x_i)/\sum_{j=1}^{C}\exp(x_j)$. $\mathbf{f} \in \mathbb{R}^L$ is the bias parameter vector and $\mathbf{U} \in \mathbb{R}^{L\times H}$ is the connection matrix. L is label dimension.

Algorithm 1 Computing $\log p(\mathbf{v}, y)$ using SmiDocNADE

Input A piece of training social news $\mathbf{v}$, label y, multimodal embedding vector $\mathbf{e}$

Output Log-likelihood function $\log p(\mathbf{v}, y)$

$\overrightarrow{\mathbf{a}} \leftarrow \overrightarrow{\mathbf{c}} + G\mathbf{e}$
$\overleftarrow{\mathbf{a}} \leftarrow \overleftarrow{\mathbf{c}} + \sum_{i \geq 1} W_{:,v_i} + G\mathbf{e}$
$p(\mathbf{v}) = 1$
for i from 1 to D **do**
  $\overrightarrow{\mathbf{h}}_i \leftarrow \mathrm{g}(\overrightarrow{\mathbf{a}})$, $\overleftarrow{\mathbf{h}}_i \leftarrow \mathrm{g}(\overleftarrow{\mathbf{a}})$
  compute $p(v_i = w|\mathbf{v}_{<i})$ by equation (7)
  compute $p(v_i = w|\mathbf{v}_{>i})$ by equation (8)
  $\overrightarrow{\mathbf{a}} \leftarrow \overrightarrow{\mathbf{a}} + W_{:,v_i}$, $\overleftarrow{\mathbf{a}} \leftarrow \overrightarrow{\mathbf{a}} - W_{:,v_i}$
  $p(\mathbf{v}) \leftarrow p(\mathbf{v}) p(v_i = w|\mathbf{v}_{<i}) p(v_i = w|\mathbf{v}_{>i})$
end for
compute $\overrightarrow{\mathbf{h}}_y(\mathbf{v})$ and $\overleftarrow{\mathbf{h}}_y(\mathbf{v})$ by equation (20) and (21)
compute $\mathbf{h}_y(\mathbf{v})$ by equation (12)
compute $p(y|\mathbf{v})$ by equation (13)
$\log p(\mathbf{v}, y) \leftarrow \log p(y|\mathbf{v}) + \frac{1}{2} \log p(\mathbf{v})$

The negative log-likelihood function is reformulated as:

$$\begin{aligned} L(\mathbf{v}, y) &= -\log p(\mathbf{v}, y) \\ &= -\log p(y|\mathbf{v}) - \log p(\mathbf{v}) \\ &= -\log p(y|\mathbf{v}) - \frac{1}{2}\lambda \sum_{i=1}^{D} \log p\left(v_i|\mathbf{v}_{<i}\right) + \log p\left(v_i|\mathbf{v}_{>i}\right), \end{aligned} \tag{15}$$

where the second term is a regularizer term. λ is the adjustable weight. This term helps to find a solution that can also satisfy the unsupervised statistical structure.

As illustrated in Figure 1, deeply mining the inter-correlations between images and text is critical. A two-branch GCN model is used to fuse arbitrary numbers of images and sentences of the social news. The extracted features act as an enhancement of SmiDocNADE, as shown in Figure 2. We follow paper [23] to generate an image graph and a sentence graph by the images and text of a piece of social news. Nodes in two graphs represent images and sentences respectively. The numbers of nodes vary in different graphs due to the arbitrary numbers of images and sentences in the social news. The image graph and the text graph are fed into two separate GCN layer, followed by a fusion layer and another two separate GCN layers.

For the GCN layer update strategy, the output of the $(l+1)^{th}$ graph convolution layer [46] is:

$$H_{l+1} = \sigma(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} H_l \Theta_l), \tag{16}$$

where $H_l \in \mathbb{R}^{N \times C}$ is the input of the $(l+1)^{th}$ convolution layer with N nodes and C feature dimension. σ is the activation function. $\widetilde{A} = A + I_N$, where A is the adjacency matrix, I_N is the identity matrix. $\widetilde{D}$ is the degree matrix of $\widetilde{A}$: $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$. Θ_l is the trainable weighted matrix.

Suppose $H_l^{(I)} \in \mathbb{R}^{N_1 \times C}$ and $H_l^{(S)} \in \mathbb{R}^{N_2 \times C}$ are outputs of the l^{th} convolution layers of image branch and text branch with N_1 image nodes and N_2 sentence nodes. To fuse image data into text branch, image graph centre $\overline{h}_l^{(I)} \in \mathbb{R}^{1 \times C}$ is calculated as:

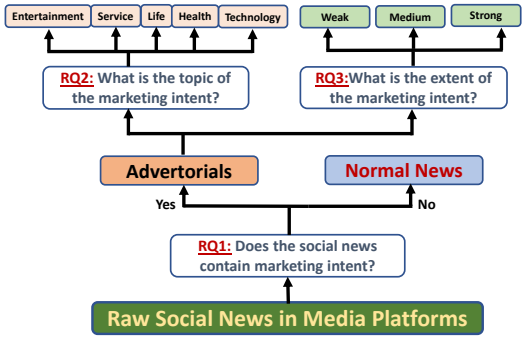


Fig. 3. Logic sequence of three proposed research questions.

$$\overline{h}_l^{(I)} = \frac{\mathbf{1}_{1 \times N_1} H_l^{(I)}}{N_1}. \tag{17}$$

Here $\mathbf{1}_{1 \times N_1}$ is a vector of all ones. The update formula of text nodes in text branch is:

$$H_l^{(S)\prime} = H_l^{(S)} + \mu \overline{H}_l^{(I)}, \tag{18}$$

which is the input of the $(l+1)^{th}$ layer in text branch. Here μ is an adjustive parameter. $\overline{H}_l^{(I)} \in \mathbb{R}^{N_2 \times C}$ in which every row is $\overline{h}_l^{(I)}$. Similarly, the text data are fused into image branch.

Let $\mathbf{e} \in \mathbb{R}^{D_e}$ be the extracted feature vector of a piece of social news by the abovementioned two-branch GCN, where D_e is the feature dimension. Equation (6) and (7) are reformulated as:

$$\overrightarrow{\mathbf{h}}_i(\mathbf{v}_{<i}) = \mathrm{g}(\overrightarrow{\mathbf{c}} + \sum_{k<i} \mathbf{W}_{:,v_k} + G\mathbf{e}), \tag{19}$$

$$\overleftarrow{\mathbf{h}}_i(\mathbf{v}_{>i}) = \mathrm{g}(\overleftarrow{\mathbf{c}} + \sum_{k>i} \mathbf{W}_{:,v_k} + G\mathbf{e}), \tag{20}$$

where $G \in \mathbb{R}^{H \times D_e}$ is the connection matrix. Equation (11) and (12) is reformulated as:

$$\overrightarrow{\mathbf{h}}_y(\mathbf{v}) = \mathrm{g}(\overrightarrow{\mathbf{c}} + \sum_{k \leq D} \mathbf{W}_{:,v_k} + G\mathbf{e}), \tag{21}$$

$$\overleftarrow{\mathbf{h}}_y(\mathbf{v}) = \mathrm{g}(\overleftarrow{\mathbf{c}} + \sum_{k \geq 1} \mathbf{W}_{:,v_k} + G\mathbf{e}), \tag{22}$$

which is also shown in Figure 2.

Minimizing the negative log-likelihood (equation (15)) is achieved by stochastic gradient descent by backpropagation. Algorithm 1 shows the computation of $\log p(\mathbf{v}, y)$. For the topic model, with D and H being the number of words and the size width of hidden layer, the complexity requires $O(DH)$.

IV. Applications and Datasets

A. Applications

As mentioned before, we specifically decompose the research problem about multimodal-based marketing intent analysis into three questions, as shown in Figure 3. In this section, we give clear definitions as follows:

—RQ1: Does the social news contain marketing intent? Because a marketing advertorial is usually designed to pretend as a piece of social news, this question is to distinguish between advertorials and normal social news. For example, is the intent of a piece of social news sharing makeup tips or recommending a brand of lipstick.

—RQ2: What is the topic of marketing intent? It is to mine the marketing topic, which defines a topic distribution of observable textual words and visual words. Targeted marketing relies on accurate marketing content classification so that media platforms can design better marketing strategies about when, where, and to whom the advertorials should be recommended.

—RQ3: What is the extent of the marketing intent? It is to detect the sentiment extent of marketing through the writer's expression. For example, if an advertorial includes detailed prices, contact numbers, or activity instructions, it contains strong marketing intents. This is important because social media may need to balanced the loads of normal social news and commercial advertising to alleviate readers' disgust.

B. Dataset Construction

There is no complete open dataset available for the proposed three questions. In this work, we collect datasets from SOHU's social news dataset [55]. It is a multimodal Chinese dataset containing normal social news and advertorials crawled from media platforms. The content includes current affairs, lifestyle, entertainment, local event information, and so on. Beyond these, some marketing advertorials are concealed in normal social news for the purpose of profit. Each piece of social news contains arbitrary numbers of images and sentences. We randomly selected 80%, 10%, and 10% of samples from datasets DS1, DS2 and DS3 into training, validation, and testing.

1) Dataset for RQ1 (DS1): To evaluate the performance on RQ1, we collect a dataset DS1. We filter the social news with images less than three and more than seven, and also abandon news with few sentences less than six after we filter the commonly used stop words. Totally we get 12375 pieces of tagged social news with 58228 images and 319504 sentences. There are 7258 pieces of positive samples representing marketing advertorials and 5117 pieces of negative samples representing normal news. Figure 4 shows the statistics of DS1, regarding the number of images/sentences per piece of social news. The social news including three images is more than a quarter of the total social news. With the increase in the number of images per piece of news, the number of pieces of social news decreases. In terms of the text information, most of the news includes sentences between 6 to 25. The positive samples (marketing advertorials) and negative samples (normal social news) are already labeled in the original dataset.

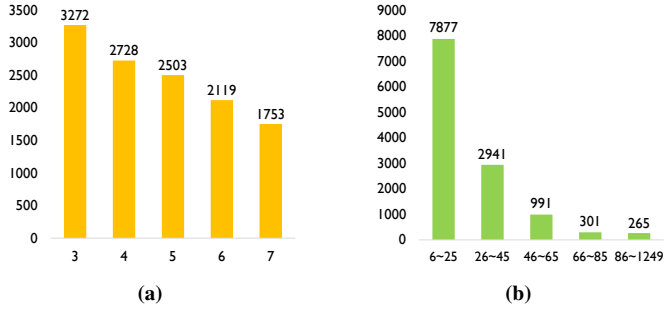


Fig. 4. (a) Statistics of DS1 regarding the number of images per piece of social news. The horizontal axis represents the number of images per piece of social news. The vertical axis represents the number of pieces of social news. (b) Statistics of DS1 regarding the number of sentences per piece of social news. The horizontal axis represents the number of sentences per piece of social news. The vertical axis represents the number of pieces of social news.

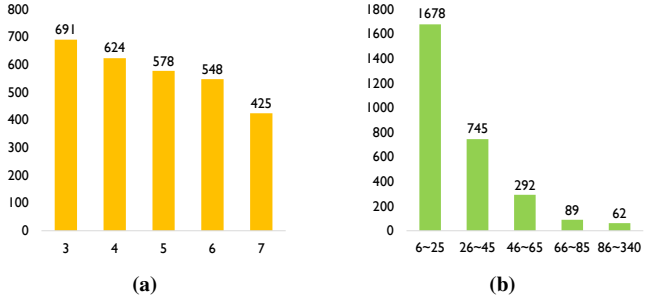


Fig. 5. (a) Statistics of DS2/3 regarding the number of images per piece of social news. The horizontal axis represents the number of images per piece of social news. The vertical axis represents the number of pieces of social news. (b) Statistics of DS2/3 regarding the number of sentences per piece of social news. The horizontal axis represents the number of sentences per piece of social news. The vertical axis represents the number of pieces of social news.

2) Datasets for RQ2 (DS2) and RQ3 (DS3): Two datasets with marketing advertorials labeled by marketing topic and extent were built for training the classifiers to estimate the probabilities of the contents generated under each topic (RQ2) and extent (RQ3). The datasets are labelled by three subjects. Although manual labeling is very expensive and time-consuming, it is more reliable especially for marketing intent inference compared with other automated methods. During the evaluation, we present all images and text of one advertorial at a time and then require the subjects to label the sample with five pre-defined topic labels and three pre-defined extent labels based on their comprehension. The subjects needed to read an advertorial for at least 30 seconds before making the final decision. Finally, we selected 2866 advertorials, including 11754 images and 67978 sentences from DS1 to form 328 health advertorials, 369 life advertorials, 448 service advertorials, 499 entertainment advertorials, and 405 technology advertorials. As shown in Table III, the data is partitioned into three sets: training, validation and testing. Figure 5 shows the statistics of DS2/3, regarding the number of images/sentences per piece of social news. The distribution is similar to DS1.

TABLE III: Statistics of datasets DS2 and DS3.

	Training	Validation	Testing	Total
# Image	9422	1178	1154	11754
# Sentence	54341	7079	6558	67978
# Social news	2292	286	288	2866

V. EXPERIMENTS

In this section, we introduce the details of the experimental setup and the evaluation metrics. we compare and analyze the performance of our proposed method and other competitor methods. We show the detailed performance comparison of accuracy over 5 topics on DS2 and 3 levels on DS3. We also present the results of ablation study and provide qualitative case studies.

A. Experimental Setup

We followed the previous work for vocabulary preparation and construction [39]. For textual data, we firstly delete some useless punctuations and then use 'jieba' word segmentation tool [59] for word segmentation. Then we filter some commonly used stop words. We select 2000 high-frequency textual words into the vocabulary for SmiDocNADE. For image data, we rescale images to make the maximum side of each image be 480 pixels, keeping the aspect ratio. We densely sample 128-dimensional SIFT features. We use four different scales of patch size with 4,6,8,10 pixels and patch step 3 pixels. The extracted features are quantized into 2000 clusters by K-means, which form 2000 visual words. The textual word order and visual word order were determined by simply randomly shuffling following previous works [37], [60], [61]. The hidden state dimension is 2048. To extract the pre-trained features by two-branch GCN, we use Word2Vec [62] model to extract the raw textual features of dimension 256 as node features in the text graph and ResNet [63] model with output dimension 256. We use two graph convolution layers in both image branch and text branch. Parameters are shared between two branches. The output dimension of the convolution layer is 128. The extracted feature dimension is 256. The regularizer weight λ is empirically set as 1. All experiments were run on a Linux workstation with one Nvidia Quadro V100 GPU of 32 GB memory. The training time on the training set of DS1 is about 5 hours.

B. Evaluation Metrics and Baselines

We calculate two metrics to evaluate the performance of methods: accuracy and F1-score. Accuracy$=\frac{TP+TN}{TP+TN+FP+FN}$. F1-score $=\frac{2TP}{2TP+FP+FN}$, where TP and TN are numbers of positive and negative samples that are correctly classified. FP and FN are numbers of positive and negative samples that are falsely classified.

To answer the above three research questions, we compared the performance of the proposed system with several competitors: image only based methods (SIFT [56] and I-GNN [23]), text only based methods (TFIDF [57] and T-GNN [23]), and multi-modal based methods (TBSN [58], TBSAGPool [51], TBCGNN [23], and SupDocNADE [39]). T-GNN, I-GNN and TBCGNN are originally proposed to detect marketing intents in [23]. TBSN, TBSAGPool, and TBCGNN fuse multimodal data in the supervised strategy. SupDocNADE is a supervised multimodal neural autoregressive topic model method. DF is the decision fusion method of images and sentences. The embedded features of images and sentences are firstly fed into two separate prediction models. Then the semantic-level decisions are fused for the final prediction. MMIA is our proposed method and MMIA-DF is the variant of MMIA. For MMIA-DF, SmiDocNADE is enhanced by a new two-branch GNN in which text branch and image branch are fused by a decision fusion strategy. MMIA-R is to use a pre-trained ResNet [63] instead of SIFT to extract visual features for SmiDocNADE.

C. Results on DS1, DS2 and DS3

Table IV shows the comparison results between baselines and our method for three questions. From the table, it is clear to see that the performance of our system is the best compared with the other eight methods in all three tasks, highlighting that a supervised topic model indeed increases the discriminative power of the model. The improvement of accuracy is obvious. The F1-score is slightly better. Accuracy measures the ratio of correct predictions, while F1-score conveys the balance between the precision and the recall. The main goal of our task is to recommend advertorials accurately to target readers. The focus is not to find all advertorials from the social news. Thus the accuracy is the key performance indicator in our task.

In terms of RQ1, we obtain the accuracy of 75.1% and the F1-score of 73.9%. The improvements are +1.9% (from 73.2% to 75.1%) on accuracy and +17.6% (from 56.3% to 73.9%) on F1-score compared with SupDocNADE. The methods of SIFT and I-GNN using only image data are the two worst on both two metrics. This reveals that it is difficult to discovery marketing intent by image only because the latent marketing intents are sometimes more implicit in images. In contrast, text based methods, e.g., TFIDF and T-GNN achieve better performance, even better than some multimodal based methods. The accuracy improvement of TFIDF is 13% compared with SIFT. It also verifies that the text is the major modality in advertorials. Although the f1-score of TF-IDF is very close to the proposed method in RQ1 (73.8 vs 73.9), the accuracy improves a lot (72.1 vs 75.1). With proper modelling, multimodal based methods (e.g., [23], [39] and our proposed method) can obtain better performance by deeply exploring the inter-correlations cross modalities. This confirms that integrating multimodal data is superior for the task of marketing intent recognition on DS1.

In terms of RQ2, from the table, we can see that by incorporating the most information into the model, our method performs the best (69.3% accuracy) in marketing topic inferring. Image only based methods still perform the worst, indicating that marketing topic presented only by images is not clear. The performance gaps between text only methods and image only methods are also larger than those in RQ1. For example, compared with SIFT, TFIDF method improves

TABLE IV: Accuracy and F1-score Comparison.

Method		RQ1		RQ2		RQ3	
		Accuracy%	F1-score%	Accuracy%	F1-score%	Accuracy%	F1-score%
Image-based	SIFT [56]	59.1	55.3	30.6	36.7	58.3	64.0
	I-GNN [23]	64.5	60.4	43.5	46.9	60.3	68.8
Text-based	TFIDF [57]	72.1	73.8	**67.7**	**59.2**	63.9	70.2
	T-GNN [23]	68.9	64.8	64.6	63.8	71.9	70.4
Multimodal-based	TBSN [58]	69.8	63.9	60.1	65.4	**72.2**	**71.1**
	TBSAGPool [51]	71.2	60.2	61.0	61.1	63.2	69.9
	TBCGNN [23]	**74.0**	**60.8**	67.4	57.3	71.3	72.6
	supDocNADE [39]	73.2	56.3	67.1	61.4	72.2	69.6
	DF	70.9	69.8	60.6	59.2	64.2	69.0
	MMIA-DF	73.8	70.2	67.9	58.1	72.6	71.5
	MMIA-R	70.5	62.8	63.9	65.1	66.6	68.5
	MMIA	**75.1***	**73.9***	**69.3***	**69.2***	**74.3***	**72.8***
p-value		4.94e-2	3.67e-2	4.88e-2	2.20e-2	4.55e-2	3.90e-2

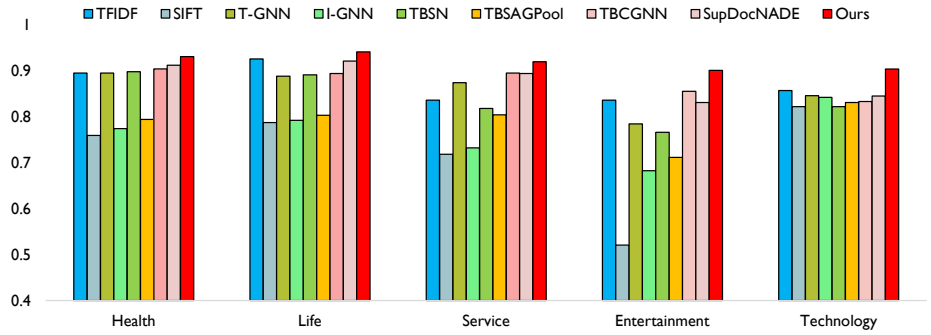


Fig. 6. Detailed performance comparison of accuracy over 5 topics on DS2.

TABLE V: Results of Ablation Study.

Method			RQ1		RQ2		RQ3	
SmiDocNADE	Two-branch GCN	LSTM	Accuracy%	F1-score%	Accuracy%	F1-score%	Accuracy%	F1-score%
✓	-	-	73.5	66.5	68.4	65.4	70.2	69.8
-	✓	-	74.0	60.8	47.4	57.3	71.3	72.6
✓	-	✓	73.7	66.7	68.8	65.7	66.3	70.4
✓	✓	-	**75.1**	**73.9**	**69.3**	**69.2**	**74.3**	**72.8**

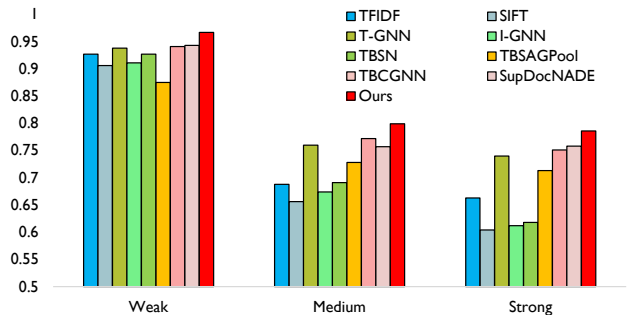
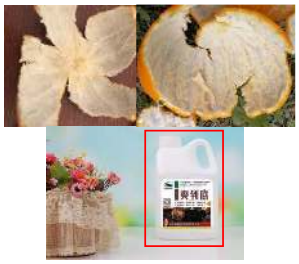


Fig. 7. Detailed performance comparison of accuracy over 3 levels on DS3.

the accuracy by even 37.1%. This is because some advertorials are about virtual products or services, such as entrepreneurship training. Most of the images are scene pictures, which lifts difficulty level for marketing topic classification. Moreover, an image usually includes multiple objects. The unrelated content brings noise into the model. Our proposed method also achieves the best 74.3% accuracy (2% more than the second best) in RQ3. It is not surprising that the accuracy scores of the image-based methods are the lowest, because language is always the major component of marketing emotion expression. It can be concluded that visual modality (images) always expresses marketing emotion in a veiled way.

From the table, it is obvious that MMIA-DF performs worse than our proposed method. This shows that simple decision-level fusion strategy performs worse than feature-



(a)

What mother needs to know: little secret hidden in baby's nails. If your baby has barbed nail, the reason may be unbalanced nutrition. Vitamin deficiency can cause dry skin...

Customizing the wardrobe is not easy! If you do not know these points, your budget can only be floated: What should I do when I make an appointment with the designer to measure the door? ...

What is the cause of mildew on the inner layer of citrus peel? Recent years, suspected internal mold in citrus is increasing. Testing agency test report shows...This view belongs to the Citrus Agricultural Technology Service Co., Ltd.

(b)

Purchase list | The most complete wedding purchase list, the most valued items are here! For your information, Princess Wedding has organized various articles related to the marriage category...

Kaiyi X5 is equipped with a panoramic sunroof to increase daylighting in the car and improve air quality, creating a comfortable and fresh driving environment. X5 sunroof has dual opening mode and...

Powerful | Official statistics come out! Beautiful Yiju City has won 6 crowns in the local property market. Recently, the real estate information network released the 2017 local property market sales data. Beautiful Yiju City...

(c)

【Renault】Every minute counts, a successful life. Hualian 1F Renault Watch Counter, January 1st to March 31st , 50% off some mechanical watches. Original price 1680, current price 588...

Meddsy emergency rescue toolbox is being crowdfunded on Indiegogo, the early bird price is 99 dollars (about 680 yuan of soft sister currency). Welcome to follow Wechat account...

If you also like to grow flowers, just follow our Wechat account: plan de home. Wechat ID: zhiwudejia. Free flower seedlings! Learn tips for growing flowers and green plants every day...

Fig. 8. Examples for marketing intent extent analysis. (a) 'weak' extent; (b) 'medium' extent; (c) 'strong' extent.

Fig. 9. Examples of advertorials that were unsuccessfully predicted.

level fusion strategy in our research questions. Compared with other baselines, both accuracy of MMIA and MMIA-DF are the best in RQ2 and RQ3. After replacing SIFT with ResNet that is pre-trained on ImageNet, both accuracy and f1-score drop. The possible reason could be pre-trained network cannot extract better image features due to the domain gap between images from ImageNet and advertisement images in our dataset. We also conducted t-tests to measure the statistical significance. We compared with the baselines that achieve the highest accuracy (shown in bold text). p-value < 0.05 indicates that the improvements of our method are statistically significant.

Figure 6 and 7 demonstrate the detailed performance comparison of accuracy on the DS2 and DS3 respectively. Our proposed method achieves the best in all five topics in task RQ2 and three extents in task RQ3. Specifically, compared with the worst method of SIFT, the improvements of our method are +38.1% (from 52.1% to 90.2%) on the 'entertainment' topic class and +18.2% (from 60.4% to 78.6%) on the 'strong' extent class. Compared with the second best method of SupDocNADE, the improvements of our method are +5.9% (from 84.6% to 90.5%) on the 'technology' topic class and +4.2% (from 75.7% to 79.9%) on the 'medium' extent class.

Table V shows the results of the ablation study in order to verify how much benefits brought by SmiDocNADE and Two-branch GCN parts. To verify whether a sequential model can achieve better performance, we adopted Long Short-Term Memory (LSTM) instead of GCN. It is observed that without SmiDocNADE or GCN feature, the performance drops from 75.1% to 74% or 73.5% in accuracy for RQ1. Practically, in RQ3, our proposed method can gain 3% more accuracy. It demonstrates that the combination can greatly enhance the discrimination ability. It is obvious that our proposed SmiDocNADE enhanced by the Two-branch GCN outperforms that with LSTM in RQ1, RQ2, and RQ3.

D. Case Studies

Qualitative examples in three marketing intent extents are provided to show what kind of content and sentiment our proposed method has captured. In Figure 8, three marketing intent extents: 'weak', 'medium', and 'strong' are presented respectively. We select three correctly recognized examples for each extent class. Because writers can post arbitrary numbers of images, we choose three or four images for each advertorial.

In Figure 8 (a), marketing information is imperceptibly embedded in the news contents. In particular, the product images of a brand of baby nutritional milk powder are embedded in the first advertorial, while the topic of the text is about the health of baby's nails. The author correlates the baby's nails with nutritional milk powder by the symptoms of nail barbs and added no description. Without images, it is hard for readers to recognize the marketing intent. The middle advertorial gives an official account ID in the bottom right corner of the images without description to attract followers. The third one shares knowledge about the diseases and pests for citrus planting in the name of a company. Moreover, a product image is presented in the content. Figure 8 (b) shows the advertorials with medium marketing extent. The authors' main focus is on describing the product properties and advantages by using the words like 'the most valued', 'comfortable and fresh', and 'won 6 crowns'. Figure 8 (c) shows the examples with strong marketing extent by using clear information related to purchase, as shown in the red words, and activity direction words, such as 'follow the account'. These advertorials convey the marketing intent in a hard way. More examples are available on *https://winlucky15.github.io/MMIA-pages/*.

Figure 9 shows two examples of advertorials that were unsuccessfully predicted by our proposed method. These two examples are both advertorials of two official accounts for brand promotion, while our method failed in recognition. The common characteristic of these two examples is that marketing texts are both embedded in the corners of images. This arouses us to recognise words in the images in the future work.

E. Experiments on MIR Flickr Dataset

To verify the performance of our method on other similar multimodal classification task, we conduct experiments on MIR Flickr Dataset [64]. This is a public dataset which contains 25000 images with textual tags. These images are labeled into 38 categories, such as nature, clouds, water, sky and so on. We follow the previous paper [39] for dataset split and experimental setups. Five baselines are compared: TFIDF, Multiple Kernel Learning SVMs [65], TagProp [66], Multimodal DBM [67], and SupDocNADE [39]. Because MIR Flickr is a multi-lable dataset, we adopt a linear SVM for classification. Moreover, every two nodes in the text graph

TABLE VI: MAP Comparison on MIR Flickr Dataset.

Method	MAP%
TFIDF	38.4
Multiple Kernel Learning SVMs [65]	62.3
TagProp [66]	64.0
Multimodal DBM [67]	65.1
SupDocNADE [39]	65.4
MMIA	**66.3**

are connected, because some images may contain tags less than three. Mean average precision (MAP) is the metric to evaluate the performance. The comparison results are shown in the Table VI. Compared with other five baselines, our proposed method MMIA achieves the best. Specifically, the improvements of our method are +1.2% and +0.9% compared with Multimodal DBM and SupDocNADE.

VI. Conclusions

In this work, we deeply analyze marketing intent behind the social news in media platforms. We propose a novel Multimodal-based Marketing Intent Analysis system (MMIA) to estimate the marketing intent embedded in the multimodal contents. We propose a supervised SmiDocNADE to increase discriminative power for specific tasks and incorporate multimodal knowledge by a two-branch GCN to mine inter-correlations cross modalities. The extensive evaluation demonstrates the advantage of our proposed system.

For the future work, we would like to consider whether some attention mechanisms can be used to re-weight the influence of different words to improve the performance of SimDocNADE. Because during data collection, we find that images always contain multiple entities. Some entities are related to the author's marketing intents, while some entities are not. For text, some textual words are directly related to the marketing content, while some textual words are related to the social news.

References

[1] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. Nielsen, *Reuters institute digital news report 2019*. Reuters Institute for the Study of Journalism, 2019.

[2] M. Stelzner, "2013 social media marketing industry report," https://www.socialmediaexaminer.com/social-media-marketing-industry-report-2013/, 2013.

[3] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," *Marketing Science*, vol. 31, no. 3, pp. 521–543, 2012.

[4] "Australian press council," https://www.presscouncil.org.au/document-search/guideline-advertorials-june-2005/, 2005.

[5] "the global state of digital in 2019 report," https://wearesocial.com/global-digital-report-2019, 2019.

[6] P. Kotler, S. Burton, K. Deans, L. Brown, and G. Armstrong, *Marketing*. Pearson Australia, 2012.

[7] W.-L. Wang, E. C. Malthouse, B. Calder, and E. Uzunoglu, "B2b content marketing for professional services: In-person versus digital contacts," *Industrial Marketing Management*, vol. 81, pp. 160–168, 2019.

[8] H. H. Chang, K. H. Wong, and T. W. Chu, "Online advertorial attributions on consumer responses: materialism as a moderator," *Online Information Review*, 2018.

[9] B. Hendricks, "How to write an advertorial: Layout & guidelines," https://study.com/academy/lesson/how-to-write-an-advertorial-layout-guidelines.html, 2017.

[10] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1423–1432, 2019.

[11] X. Ding and Z. Chen, "Improving saliency detection based on modeling photographer's intention," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 124–134, 2018.

[12] K. Shu, S. Mukherjee, G. Zheng, A. H. Awadallah, M. Shokouhi, and S. Dumais, "Learning with weak supervision for email intent detection," *arXiv preprint arXiv:2005.13084*, 2020.

[13] C. Chen, C. Fu, X. Hu, X. Zhang, J. Zhou, X. Li, and F. S. Bao, "Reinforcement learning for user intent prediction in customer service bots," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1265–1268.

[14] W. Chen and J. J. Corso, "Action detection by implicit intentional motion clustering," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3298–3306.

[15] J. Wang, W. X. Zhao, H. Wei, H. Yan, and X. Li, "Mining new business opportunities: Identifying trend related products by leveraging commercial intents from microblogs," in *Proceedings of the Conference on Empirical Methods in Natural Language*, 2013, pp. 1337–1347.

[16] B. Hollerit, M. Kröll, and M. Strohmaier, "Towards linking buyers and sellers: detecting commercial intent on twitter," in *Proceedings of the International Conference on World Wide Web*, 2013, pp. 629–632.

[17] A. Shishkin, P. Zhinalieva, and K. Nikolaev, "Quality-biased ranking for queries with commercial intent," in *Proceedings of the International Conference on World Wide Web*, 2013, pp. 1145–1148.

[18] L. Wang, M. Ye, and Y. Zou, "A language model approach to capture commercial intent and information relevance for sponsored search," in *Proceedings of the International ACM Conference on Information and Knowledge Management*, 2011, pp. 599–604.

[19] Q. Guo and E. Agichtein, "Exploring searcher interactions for distinguishing types of commercial intent," in *Proceedings of the International Conference on World Wide Web*, 2010, pp. 1107–1108.

[20] A. Ashkan and C. L. Clarke, "Term-based commercial intent analysis," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 800–801.

[21] Y. Wang, S. Liu, S. Li, J. Duan, Z. Hou, J. Yu, and K. Ma, "Stacking-based ensemble learning of self-media data for marketing intention detection," *Future Internet*, vol. 11, no. 7, p. 155, 2019.

[22] X. Liang, C. Wang, and G. Zhao, "Enhancing content marketing article detection with graph analysis," *IEEE Access*, vol. 7, pp. 94 869–94 881, 2019.

[23] L. Zhang, J. Zhang, Z. Li, and J. Xu, "Towards better graph representation: Two-branch collaborative graph neural networks for multimodal marketing intention detection," in *Proceedings of the International IEEE Conference on Multimedia and Expo*, 2020, pp. 1–6.

[24] S. Kim, J.-Y. Jiang, M. Nakada, J. Han, and W. Wang, "Multimodal post attentive profiling for influencer marketing," in *Proceedings of The Web Conference*, 2020, pp. 2878–2884.

[25] M. Mazloom, R. Rietveld, S. Rudinac, M. Worring, and W. Van Dolen, "Multimodal popularity prediction of brand-related social media posts," in *Proceedings of the ACM International Conference on Multimedia*, 2016, pp. 197–201.

[26] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 997–1007, 2017.

[27] Q. You, L. Cao, Y. Cong, X. Zhang, and J. Luo, "A multifaceted approach to social multimedia-based prediction of elections," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2271–2280, 2015.

[28] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1910–1921, 2016.

[29] J. Wang, J. Wang, C. Sun, S. Li, X. Liu, L. Si, M. Zhang, and G. Zhou, "Sentiment classification in customer service dialogue with topic-aware multi-task learning," in *Proceedings of the International Conference of Association for the Advancement of Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9177–9184.

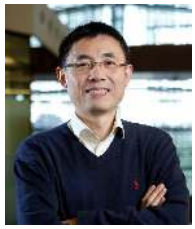
[30] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1231–1243, 2013.

[31] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment

classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2019.
[32] T. Liu, J. Wan, X. Dai, F. Liu, Q. You, and J. Luo, "Sentiment recognition for short annotated gifs using visual-textual fusion," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1098–1110, 2019.
[33] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, 2020.
[34] M. Gong, Y. Gao, Y. Xie, and A. Qin, "An attention-based unsupervised adversarial model for movie review spam detection," *IEEE Transactions on Multimedia*, 2020.
[35] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in *Proceedings of the IEEE International Conference on Data Mining Workshops*, 2013, pp. 440–444.
[36] H. P. Chan, W. Chen, and I. King, "A unified dual-view model for review summarization and sentiment classification with inconsistency loss," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1191–1200.
[37] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *Proceedings of the Conference on Neural Information Processing Systems*, 2012, pp. 2708–2716.
[38] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "Topic modeling of multimodal data: an autoregressive approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1370–1377.
[39] Y. Zheng, Y. Zhang, and H. Larochelle, "A deep and autoregressive approach for topic modeling of multimodal data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 6, pp. 1056–1069, 2015.
[40] P. Gupta, Y. Chaudhary, F. Buettner, and H. Schütze, "Document informed neural autoregressive topic models with distributional prior," in *Proceedings of the Conference of Association for the Advancement of Artificial Intelligence*, vol. 33, 2019, pp. 6505–6512.
[41] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2011, pp. 2407–2414.
[42] D. Valsesia, G. Fracastoro, and E. Magli, "Learning localized representations of point clouds with graph-convolutional generative adversarial networks," *IEEE Transactions on Multimedia*, 2020.
[43] W. Nie, M. Ren, A. Liu, Z. Mao, and J. Nie, "M-gcn: Multi-branch graph convolution network for 2d image-based on 3d model retrieval," *IEEE Transactions on Multimedia*, 2020.
[44] L. Lu, Y. Lu, R. Yu, H. Di, L. Zhang, and S. Wang, "Gaim: Graph attention interaction model for collective activity recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 524–539, 2019.
[45] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan, "A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2020.
[46] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[47] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun, "Graph neural networks: A review of methods and applications," *arXiv preprint arXiv:1812.08434*, 2018.
[48] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.
[49] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the Conference on Neural Information Processing Systems*, 2017, pp. 1024–1034.
[50] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proceedings of the Conference on Neural Information Processing Systems*, 2018, pp. 4800–4810.
[51] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 3734–3743.
[52] H. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li, "Detecting online commercial intention (oci)," in *Proceedings of the International Conference on World Wide Web*, 2006, pp. 829–837.
[53] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[54] W. Chong, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1903–1910.
[55] "Sohu second algorithm competition," https://biendata.com/competition/sohu2018/, 2018.
[56] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein, "Sclera recognition using dense-sift," in *Proceedings of the IEEE International Conference on Intelligent Systems Design and Applications*, 2013, pp. 74–79.
[57] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf* idf, lsi and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.
[58] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407, 2018.
[59] "'jieba' chinese word segmentation tool," https://github.com/fxsjy/jieba, 2020.
[60] S. Lauly, Y. Zheng, A. Allauzen, and H. Larochelle, "Document neural autoregressive distribution estimation," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4046–4069, 2017.
[61] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 29–37.
[62] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the Conference on Neural Information Processing Systems*, 2013, pp. 3111–3119.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[64] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 39–43.
[65] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 902–909.
[66] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the mirflickr set," in *Proceedings of the International Conference on Multimedia Information Retrieval*, 2010, pp. 537–546.
[67] N. Srivastava and R. Salakhutdinov, "Discriminative transfer learning with tree-based priors." in *Proceedings of the International Conference on Neural Information Processing Systems*, vol. 3, no. 4. Citeseer, 2013, p. 8.

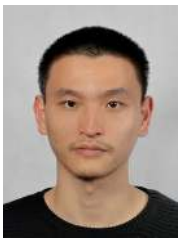
Lu Zhang reveived her M.E. degree from Chinese Academy of Sciences, Beijing, China in 2016. She is currently working toward the Ph.D. degree with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. Her research interests include machine learning and deep learning for multimodal media data analysis. She has published several papers in top journals including TMM.

Jialie Shen is a Reader in Computer Science, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast (QUB), Belfast, United Kingdom. He received his PhD in Computer Science from the University of New South Wales (UNSW), Australia in the area of large-scale media retrieval and database access methods. Dr. Shen worked as a faculty member at Hong Kong, Singapore, Australia and England and researcher at information retrieval research group (Led by Professor Keith van Rijsbergen), the University of Glasgow, Scotland before moving to the QUB. Dr. Shen's main research interests include information retrieval, machine learning, multimedia systems and audio/video analytics. His research has been published or is forthcoming in leading journals and international conferences, including ACM SIGIR, ACM Multimedia, IJCAI, AAAI, IEEE Transactions and ACM Transactions.



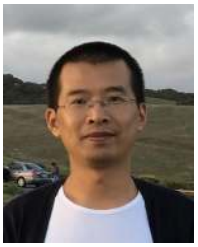
Jian Zhang (Senior Member, IEEE) received his B.S. degree in electronics from East China Normal University, China, his M.S. degree in computer science from Flinders University, Australia, and his Ph.D. degree in electrical engineering from the University of New South Wales (UNSW), Australia. From 1997 to 2003, he was a Principal Research Engineer and the Research Manager of Visual Communications Team with Motorola Australian Research Centre. From 2004 to 2011, he was a Principal Researcher and a Project Leader with Data61, Australia, and a Conjoint Associate Professor with the School of Computer Science and Engineering, UNSW. He is currently an Associate Professor in the Faculty of Engineering and IT and the Director of the Multimedia and Data Analytics Lab at the University of Technology Sydney, Australia. He has authored or co-authored over 200 paper publications, book chapters and 10 approved U.S. patents. His current interests include social multimedia signal processing, large-scale image and video content analytics, retrieval and mining, 3D-based computer vision, and intelligent video surveillance systems. Dr. Zhang was the General Co-Chair and Technical Program Co-Chair of the International Conference on Multimedia and Expo (ICME) in 2012 and 2020 respectively; and the Technical Program Co-Chair and General Co-Chair of the IEEE Conference on Visual Communications and Image Processing in 2014 and 2019 respectively. He was an Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology (2006 – 2015). Currently, he is an Associate Editor for the IEEE Transactions on Multimedia and a Member of Technical Directions Board, IEEE Signal Processing Society.

Jingsong Xu received the B.S. degree in computer science and the Ph.D. degree in pattern recognition from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2014, respectively. He is currently a Research Fellow with the Global Big Data Technologies Center, University of Technology Sydney (UTS), Ultimo, NSW, Australia. His research interests include computer vision, pattern recognition, and machine learning.



Zhibin Li reveived his Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia. He is now a postdoctoral research fellow of the Commonwealth Scientific and Industrial Research Organisation, Australia. His principal research interest includes traffic analysis, factorization models, and theoretical analysis of machine learning. He has published 11 papers in top journals and refereed conference proceedings including NeurIPS2020, KDD2019 and TKDE.

Yazhou Yao is a professor at School of Computer Science and Engineering, Nanjing University of Science and Technology. With the support of the China Scholarship Council, he received the PhD degree in Computer Science, University of Technology Sydney, Australia at 2018. From July 2018 to July 2019, he worked as a Research Scientist at Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include multimedia processing and machine learning.



Litao Yu received the Ph.D. degree from the University of Queensland, Brisbane, QLD, Australia, in 2016. He then worked as a Research Fellow with the QUT Centre of Robotics, Queensland University of Technology, Brisbane, in 2017. He is currently a Research Fellow with the Multimedia Data Analytics Lab (MDAL), Global Big Data Technologies Centre (GBDTC), University of Technology Sydney, Ultimo, NSW, Australia. He is also an Adjunct Research Fellow with the Institute for Integrated and Intelligent Systems, Griffith University, Nathan, QLD, Australia. His research interests include machine learning, multimedia content analysis, and image processing.