# ANALYZING IMAGE TWEETS IN MICROBLOGS

**TAO CHEN**

*(B.Eng (Hons.), East China Normal University)*

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2016

# DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

TAO CHEN

20 January 2016

# ACKNOWLEDGEMENTS

This thesis would not have been possible without the support, direction, and love of a multitude of people.

Foremost, I would like to express my deepest gratitude to my advisor Prof. Min-Yen Kan for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm and immense knowledge. Without his generous and unwavering help, I would not have been able to complete this thesis.

I would also like to thank members of my thesis committee, including Prof. Tat-Seng Chua, Prof. Shengdong Zhao and Prof. Lexing Xie, who have put in the time and effort to review and assess this thesis. I also want to thank Prof. Chew-Lim Tan and Prof. Hwee Tou Ng for their insightful comments to my thesis proposal and graduate research paper.

My sincere thanks also go to Prof. Peng Cui for his generous help and guidance during the start of my research in multimedia. I also appreciate the help from my colleagues including Jin Zhao, Jesse Prabawa Gozali, Ziheng Lin, Junping Ng, Aobo Wang, Jovian Lin, Xiangnan He, Muthu Chandrasekaran, Dongyuan Lu, Kazunari Sugiyama, Hany SalahEldeen, Yi Yu and Yongfeng Zhang.

Although not directly related to this thesis, I would like to thank Google for providing the internship opportunity in NYC. It's a wonderful three-month experience for both work and life. I also thank Google for its generous Anita Borg Memorial Scholarship that covers my tuition fee for one semester.

My appreciation also goes out to my friends who are either in Singapore or far away, for their help and support for my ups and downs. In particular, I thank my fellow PhD friends for their companion and help for accomplishing each milestone. I was also very lucky to have great flatmates in UTown, who made our dormitory like a home.

Finally, I would like to express my heartful thanks to my parents for their love, support, encouragement and understanding all the time. Without them, I would not have the courage to conquer all the difficulties.

# CONTENTS

# ABSTRACT

Social media platforms now allow users to share images alongside their textual posts. These *image tweets* make up a fast-growing percentage of tweets, but have not been studied in depth unlike their text-only counterparts. Most existing studies on image tweets tackle tasks that are originated from text tweet domain, and their main effort is to incorporate generic image features (*e.g.*, low-level features, deep learning features) to improve the performance of using text-only approaches.

In this thesis, we conduct a series of studies to answer four fundamental questions about image tweets: 1) What are the characteristics of image tweets? 2) What are the relationships between the image and text in image tweets? 3) How to model such image-text relationships? and 4) How to interpret the semantics of an image tweet? Answers to these questions will not only help us gain a deep understanding of image tweets and the related user behaviors, but also be beneficial to downstream applications.

To answer the first question, we collect a large corpus of microblog posts from (Western) *Twitter* and (Chinese) Sina *Weibo*. We find 56.0% of posts on Weibo and 14.1% on Twitter are image tweets. We then perform a multipronged analysis of these image tweets from the perspective of image characteristics, user posting behaviors (*e.g.*, temporal, access) and textual contents.

As an image tweet usually consists of two media—an image and its accompanying text, we naturally ask: what are the relationships between these two? Using an appropriate corpus analysis, we identify two key image-text relations for image tweets: *visual relevance* and *emotional relevance*. Considering the practical values of visually relevant image tweets, we build an automated classifier utilizing text, image and social context features to distinguish them from the others, obtaining a macro $F_1$ of 70.5%.

Based on this, a follow-up question arises: can we model these image-text relations and explain how image tweets are generated? To this end, we

develop Visual-Emotional LDA (VELDA), a novel topic model that captures the image-text relations from multiple perspectives (namely, visual and emotional) to model the image tweet generation process. Experiments on real-world image tweets in both English and Chinese and other user generated content show that VELDA significantly outperforms existing methods in the task of cross-modality image retrieval.

Then we turn to the last question: how to interpret the semantics of an image tweet? We show microblog context is the key to understanding image tweets, and devise a context-aware image tweets modeling (CITING) framework to interpret the semantics of image tweets from both intrinsic and extrinsic contexts. To demonstrate the effectiveness of our framework, we focus on the task of personalized image tweet recommendation, developing a feature-aware matrix factorization model that encodes the contexts as part of user interest modeling. Extensive experiments on a large Twitter dataset show our proposed method significantly improves recommendation performance.

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF PUBLICATIONS

Here is the list of published works during my Ph.D. candidature. Works closely related to this thesis are highlighted by (*). The recent research works focusing on image tweets benefit greatly from my research effort committed on general social media analysis and natural language processing.

1. (*) **Tao Chen**, Dongyuan Lu, Min-Yen Kan and Peng Cui (2013). Understanding and Classifying Image Tweets. *In Proceedings of the 21st ACM International Conference on Multimedia (MM'13).*

2. (*) **Tao Chen**, Hany Salaheldeen, Xiangnan He, Min-Yen Kan and Dongyuan Lu (2015). VELDA: Relating an Image Tweet's Text and Images. *In Proceedings of the 29st AAAI Conference on Artificial Intelligence (AAAI'15).*

3. (*) **Tao Chen**, Xiangnan He, Min-Yen Kan (2016). Mining Contextual Text for Image Tweets (under review).

4. Xiangnan He, **Tao Chen**, Min-Yen Kan and Xiao Chen (2015). Review-aware Explainable Recommendation by Modeling Aspects. *In Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15).*

5. Bang Hui Lim, Dongyuan Lu, **Tao Chen** and Min-Yen Kan (2015). #mytweet via Instagram: Exploring User Behaviour across Multiple Social Networks. *In Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15).*

6. **Tao Chen**, Naijia Zheng, Yue Zhao, Muthu Chandrasekaran and Min-Yen Kan (2015). Interactive Second Language Learning from News Websites. *In Proceedings of ACL Workshop on Natural Language Processing Techniques for Educational Applications.*

7. **Tao Chen** and Min-Yen Kan (2013). Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus.*Language Resources and Evaluation*, 47(2)(2013).

8. Aobo Wang, **Tao Chen** and Min-Yen Kan (2012). Re-tweeting from a Linguistic Perspective. *In Proceedings of the NAACL-HLT 2012 Workshop on Language in Social Media.*

# Chapter 1

# Introduction

With improved bandwidth and camera phones, user-generated mainstream social media is no longer solely text but firmly multimedia. *Image tweets*, which we define as user-generated microblog posts that contain an embedded image, are now a staple of user-generated content. While the ability to link images to microblog posts has existed for several years, the difficulty composing such posts made these type of posts a minority. Starting with Sina *Weibo*[1], the dominant microblog platform in China; and later *Twitter*[2] and third-party services such as *Instagram*[3], microblogging platforms now seamlessly include images into their posts.

Such multimedia form of posts attracts larger viewership and prolongs their half-life as compared to their poorer cousins—text-only posts—a claim that has been validated on Weibo [162], and have been found to be 35% more retweetable than text-only tweets on Twitter[4]. No wonder such posts are fast becoming the de facto standard on such microblog platforms. They constitute over 56% of trending posts on Weibo (as reported by Yu *et al.* [152] in 2011), and have seen rapid adoption on Twitter. Figure 1.1 and

[1] `http://weibo.com`
[2] `https://twitter.com`
[3] `http://instagram.com`
[4] `https://blog.twitter.com/2014/what-fuels-a-tweets-engagement`


**Figure 1.1: An example image tweet from Twitter.**


**Figure 1.2: An example image tweet from Weibo.**

1.2 display example image tweets from Twitter and Weibo, respectively. The services of the two platforms are very similar, and they are learning from each other, *e.g.,* "retweet with comment" is an inborn functionality on Weibo (since 2009) and was recently added as functionality within Twitter as well (since April 2015)[5]; while "reply with a photo" feature was first introduced by Twitter in January 2014[6], and then by Weibo in April 2015[7].

[5] `http://techcrunch.com/2015/04/06/retweetception`

[6] `http://appleinsider.com/articles/14/01/29/twitter-updated-with-new-reply-photo-options-voicemail-offers-message-transcription`

[7] `http://cn.engadget.com/2015/04/26/ios-weibo-update`

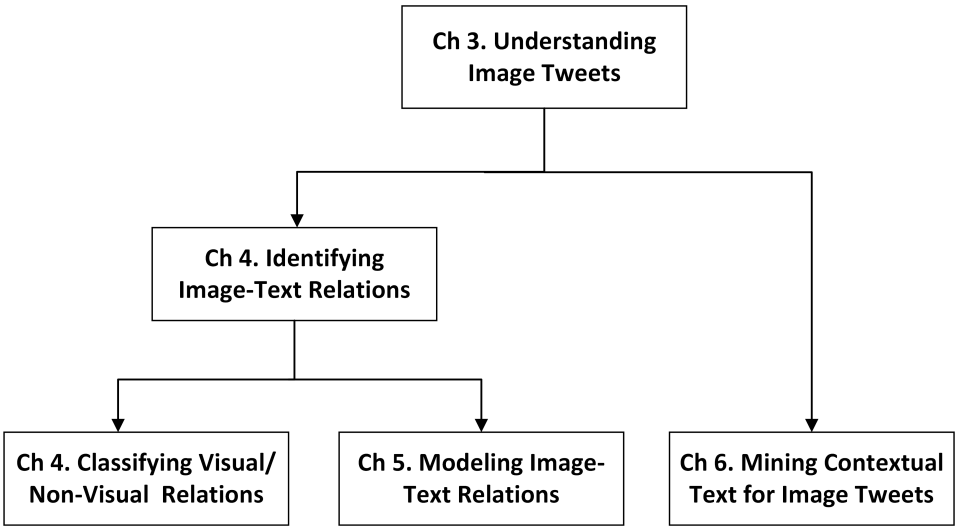



**Figure 1.3: The framework of this thesis.**

## 1.1 Motivation

Unlike their text-only counterparts, image tweets have only been studies in a few works. Most studies tackle tasks that are originated from text tweet domain (*e.g.*, popularity prediction, event detection), and their main contribution is to incorporate generic image features (*e.g.*, low-level features, deep learning features) to improve the performance of using text-only approaches. While helpful, these previous works merely studied the characteristics of image tweets, and many questions about this new class of social multimedia have not been explored. In this thesis, we aim to answer four fundamental questions about image tweets. These answers will not only help us gain a deep understanding of image tweets and the related user behaviors, but also be beneficial to downstream applications. The framework of this thesis is illustrated in Figure 1.3.

**RQ1. What are the characteristics of image tweets?** In contrast to image tweets, the characteristics of text (only) tweets and user behavior surrounding their creating and dissemination have been studied for years. For example, seminal work by Java *et al.* [60] in 2007 has given insights to

a series of works[8]. Since image tweets are a new form of communication, a nature question arises: what are the characteristics of image tweets? Considering the social context that image tweets lie in, we investigate this question from three angles. Firstly, as user-generated images have existed in photo-sharing websites for a few years, *e.g.*, *Flickr*[9] was founded in 2004, are images in microblogs similar to the photos in photo-sharing websites? Secondly, aside from the existence of embedded images, how do image tweets differ from text tweets, in terms of the textual content and various user behaviors? Thirdly, since text tweets and their related microblogging behaviors exhibit significant differences on Twitter and Weibo [39, 40] due to the Western and Chinese culture differences, do image tweets exhibit similar differences? We explore these topics and answer these questions in Chapter 3.

**RQ2. What are the relationships between the image and text?** Though users on microblogs can post an image without accompanying text, it is rare - 97.8% of Twitter image tweets and 95.0% of Weibo image tweets in our dataset (detailed in Chapter 3) have corresponding text. We want to know why people post both image and text and the nature of their correlation. The distinct image-text relations hint at the user's underlying motivations, and are a beneficial source of knowledge for downstream applications. For instance, in image tweet retrieval task, we can group resultant image tweets by their image-text relation. Similarly, to enhance user's experience, we may prioritize the display of image tweets according to the importance of image-text relations in a certain context. This motivates us to build a classifier that automatically recognizes the type of image-text relation given an image tweet. We detail these in Chapter 4.

[8]As of Jan 20, 2016, it has been cited by over 2,000 times according to Google Scholar.

[9]`https://www.flickr.com`

**RQ3. How to model the image-text relationships?** After identifying the key image-text relations from image tweets, a follow-up question arises: can we model these relationships? Furthermore, can we use these models to explain how image tweets are generated? Such an image-text correspondence model is vital for certain real applications. As an image helps to attract readership for microblog posts [162], we can leverage the model to illustrate microblog posts automatically when users do not have proper images at hand. This model is also applicable for cross-modality image retrieval (*i.e.*, given a text, retrieving relevant images from a dataset, or vice verse) and automated image tagging. However, the existing models for image-text relations do not take the unique features of image tweets into consideration, and thus we are motivated to propose a novel model to capture image-text relations for image tweets. We discuss this work in Chapter 5.

**RQ4. How to interpret the semantics of an image tweet?** To utilize such multimedia posts for downstream applications (*e.g.*, personalized news feed, advertising), another fundamental question needs to be addressed: what are the image tweets about? One semantic carrier is the accompanying text which has been studied for years, *e.g.,* representing the semantics by hashtags [72, 82] and latent topics [110, 161]. For the other – the embedded image, existing studies have largely focused on understanding it from its distilled low-level features [139, 58, 9, 10]. However, the gap between these low-level features and the real semantics limits the model fidelity. Other works have leveraged higher level features, *e.g.*, the output from high layers of convolutional neural networks trained for image recognition [19, 156] and the visual objects from an object detector [21].

However, such detected visual objects do not accurately interpret images in the context of microblog. Take the image in Figure 1.4 as an example. The typical visual tags ("child, cute, girl, little, indoor") do not

**Figure 1.4: A poster image for China ends the one-child policy.**

tell the story behind this image - this image was the poster for the story of China abandoning its controversial one-child policy. A photo's semantics depend not only on its pixel values, but also on the context in which the picture was taken and used [59]. In order to properly interpret microblog images, it is mandatory to go from capturing visual properties to modeling context. We present the details in Chapter 6.

## 1.2 Key Contributions

This thesis makes contributions on analyzing image tweets and partically answering four fundamental questions. They are summarized as follows:

1. **Understanding image tweets.** 56.0% of posts on Weibo and 14.1% of Twitter posts contain at least one image. We performed a multi-pronged analysis of these image tweets on Twitter and Weibo from the perspective of image characteristics, user posting behaviors and textual contents. We found images on both platforms are primarily single JPEG formatted pictures, cover a wide range of content variety and are often of low quality.

   On both platforms, image tweets are more likely to be retweeted than text-only tweets and posted primarily by mobile phones during the daytime and weekend. The user's choice of embedding images is highly correlated with the textual topics. Discrepancies also exist

between the two platforms. On Weibo, 37.6% of image tweets are not genuine user generated content (*e.g.,* by-products of using other apps), and their primary purpose is promotion and advertising, leading to a few abberant behaviors on Weibo: *e.g.,* image tweets are extremely actively posted during early morning and 58.8% of image tweets on Weibo embed external URLs, whereas URLs only appear in 8.3% of Twitter image tweets.

2. **Identifying and classifying image-text relations.** Using an appropriate corpus analysis, we identified two key image-text relations for image tweets: 1) visual relevance, *i.e.*, at least one textual word describes part of the image content or the whole image; 2) emotional relevance, *i.e.*, the two medium exhibit the same emotion state (*e.g.*, happy, sad). In additional to these, we observed image and text do not always exhibit strong relevance—neither visually nor emotionally relevant, and the main purpose of including an image is to make the post visually attractive.

   Identifying the image-text relations for image tweets have practical values. In particular, we pay attention to automating the distinction between *visual* and *non-visual* relations. To this end, we crowdsourced human annotations on image-text relation for 4.8K image tweets from Weibo and successfully built an automated classifier utilizing the features from text, image and the social context to distinguish the two classes, obtaining a macro $F_1$ of 70.5%. To encourage further investigation on these topics, we have made the annotated corpus available to the public[10].

3. **Modeling image-text relations**. We develop Visual-Emotional LDA (VELDA), a novel topic model to capture the image-text re-

[10] `http://wing.comp.nus.edu.sg/downloads/imagetweets`

lations from multiple perspectives (namely, visual and emotional), and thus model the generative process of image tweets. Experiments on real-world image tweets in both English and Chinese and other user generated content, show that VELDA significantly outperforms existing methods on cross-modality image retrieval. Even in other domains where emotion does not factor in image choice directly, our VELDA model demonstrates good generalization ability, achieving higher fidelity modeling of such multimedia documents. Moreover, we apply VELDA in a real-world task of automated microblog illustration, using our model to select a relevant image (either visually-relevant, emotionally-relevant or both) drawn from an image collection.

4. **Mining contextual text to uncover the semantics of image tweets.** We have earlier identified the gap between low-level features and the real-world semantics of images. To move from analyzing pixels to understanding the images' context, we propose a **c**ontext-aware **i**mage **t**weets model**ing** (CITING) framework to mine contextual text to model such social multimedia's semantics. We start with the intrinsic context in image tweets: 1) for the text, we enhance hashtags to better represent the topics of images, and 2) for the image, we apply Optical Character Recognition (OCR) to extract text from images. Then we turn to the extrinsic context: 3) using text found on hyperlinked web pages in the tweet; and 4) using text found on search engine result pages when using the image in a query-by-example image search. Mindful that the contextual text from each source differs in quality and coverage, we also propose a series of heuristics to fuse text when multiple channels are triggered. This fusion makes the modeling more accurate and reduces the acquisition cost of the con-

text.

We apply our proposed strategies to user interest modeling, a key application in the microblog domain. To this end, we develope a feature-aware matrix factorization model that encodes the contexts as part of user interest modeling. Extensive experiments on a large Twitter dataset show our proposed contexts significantly improve recommendation performance. To enable comparative studies, we have released the annotated OCR dataset and the image tweets dataset to the public[11].

## 1.3 Organization

In the next chapter, we give a comprehensive overview of the existing works on image tweets. In Chapter 3, we present our work on image tweet understanding, followed by the work of image-text relation identification and classification (Chapter 4), and then we describe the generative model that captures the image-text relations (Chapter 5). We elaborate our strategies of mining contextual texts in modeling the semantics of image tweets in Chapter 6. Finally, we conclude this thesis and discuss the future directions in Chapter 7.

[11] `http://wing.comp.nus.edu.sg/downloads/image-tweet-ocr-rec`

# Chapter 2

# Background

The topic of images in microblogs has only started attracting academic attention recently. What kinds of research have been conducted on image tweets? To this end, we exhaustively searched relevant papers from mainstream multimedia conferences and journals, as well as queried academic search engines (*e.g.*, Google Scholar) via keywords. Comparing to the voluminous number of studies on text tweets, the number of papers about image tweets is relatively small. In this chapter, we give a comprehensive literature review on this topic, in order to help the readers gain an overview about the existing research. The prior works discuss a number of different topics, ranging over 1) analyzing the characteristics of images, 2) classifying the sentiment depicted in image tweets, 3) predicting the popularity of image tweets, 4) detecting multimedia event, 5) identifying fake images, 6) multimedia mining and 7) understanding users. Most of the research have been adapted from research on text tweets; therefore, a common trait for these studies is to exploit visual cues from images to improve the performance.

## 2.1 Image Characteristics

What are microblog images about? A few works attempted to answer this by manually examining a small number of images. Two papers analyzed Twitter images for a specific event, *i.e.*, the *2011 Egyptian Revolution*[1] (581 images, [68]) and the *2012 Gaza Conflict*[2] (243 images, [120]); while another two studied images marked by a specific hashtag, *i.e.* , #guncontrol (290 images, [125]) and #thinspiration (300 images, [43]). Due to the focused topics, these works are unable to give a picture for the general Twitter world. To fill this gap, Thelwall *et al.* [127] randomly sampled 800 Twitter images posted by UK users and US users (400 each) in one week of late 2014, and manually labeled the type, content, purpose and the taken time for the images. The key findings are: 1) the majority (two-thirds) of images are photographs and 15% are screenshots, 2) around 25% of images are of individual person, and 5% are food and drink; 3) except a few advertisement tweets, most tweets do not reveal a clear reason of posting; 4) similarly, it is difficult to infer whether an image is shared shortly after taken, due to the lack of contextual information (*e.g.*, Exif metadata). These findings are consistent with some of our findings, however, the characteristics of Weibo images have not been explored.

## 2.2 Sentiment

Emotion has been established as a key factor in image tweets. In a case study, Stefanone *et al.* [125] identified nearly half of images in a set of 290 Twitter images that talk about gun control issues have emotional solicitations. As such, many researches center on automatically classifying the sentiment of image tweets. Four papers built their sentiment classifier

[1] `https://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011`
[2] `https://en.wikipedia.org/wiki/Operation_Pillar_of_Defense`

using image features only [153, 150, 141, 20], while the majority of works benefitted from both text and image cues, either adopting an early fusion strategy (*e.g.*, concatenate textual and image features in a single vector to build a unified classifier [137, 6]) or a late fusion strategy (*e.g.*, build separate classifiers for each modality and then combine the sentiment scores from the two classifiers for the final prediction [16, 22, 149, 25, 6]).

Regarding to the textual features, most works simply used the surface textual words [16, 22, 25, 26] and a few encoded additional features like word classes (*e.g.,* the number of nouns) [137], and emoticons [26]. In addition, two papers took the advantage of neural networks language models: You *et al.* [149] represented the text via *Paragraph Vector* [73], and Baecchi *et al.* [6] extended the Continuous Bag-of-Words model [91] for sentiment classification task.

For images, mid-level features have been proven to be superior to low-level features [16, 153]. Most papers [22, 137, 25, 26] used the mid-level image features from *SentiBank* visual sentiment concept detectors [16], while Yuan *et al.* [153] constructed their own mid-level features by converting low-level features with classifiers. More recently, the Convolutional Neural Network (CNN) architecture [141, 150, 149, 20] and the Denoising Autoencoder [6] have been used to learn feature representation for images in sentiment classification, and *SentiBank* has been upgraded to *DeepSentiBank* by training the detector with CNN [29]. Among these works, Xu *et al.* [141] first fine-tuned *AlexNet* [70]—a CNN architecture that was pre-trained for object recognition task in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), used the activations from the seventh fully connected (*fc7*) layer neurons as the image representation, and then trained a logistic regression to classify the sentiment of Twitter images. Campos *et al.* [20] adopted another common domain adaptation technique: 1) replace the last fully connected layer (fc8, consisting of 1000 neurons)

of AlexNet with a new, randomly initialized layer consisting of two neurons, 2) initialize the other layers with the values from original AlexNet, and 3) fine-tune the new model on a set of Twitter images with annotated sentiment labels. You *et al.* [150] developed a new CNN architecture of two convolutional layers and four fully connected layers. Inspired by a famous theory from psychoevolution [105]—there are a total of 24 emotions belonging to positive and negative category, You *et al.* [150] specifically constrained the penultimate fully-connected layer to have 24 neurons.

Although comprehensive, these works have not considered the uniqueness of microblog images. For example, meme images are often uploaded to microblog feeds, and such image tweets are more likely to be sentimental rather than neutral. Therefore, the identification of memes will be beneficial for microblog image sentiment classification. On the other hand, some image features, such as artistic principle inspired features [159, 160], have demonstrated their superiority in general image sentiment classification, but have not been utilized for microblog images yet. Though the majority of microblog images are user generated (*i.e.*, not artistic masterpieces), it will still be interesting to know whether art-inspired features are effective for classifying the sentiment of microblog images or not.

With the predicted sentiment labels, Cao *et al.* [22] and Chen *et al.* [25] built interactive demo systems to show the sentimental results from multiple facets, *e.g.*, by geo-location, by topic, by user, and by post. To enable comparative studies, three groups of researchers have released their annotated dataset to the public. Borth *et al.* [16] released a set of 470 positive image tweets and 133 negative tweets from Twitter[3], Yuan *et al.* [154] made a dataset of 1269 Twitter images (769 positive and 500 negative) with sentiment labels to the public[4], and Niu *et al.* [100] 's dataset is slightly

[3] `http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html`
[4] `http://www.cs.rochester.edu/u/qyou/DeepSent/deepsentiment.html`

larger than the other two—consisting of 4,869 Twitter image tweets[5]. In addition to the sentiment prediction, Abdullah *et al.* [1] reported a *Smile Index*—a measure of societal happiness—for a given region on Twitter. To this end, they applied an existing smile detector [56] to Twitter images that contain human faces, and computed the ratio of smile-containing images for each location. They then studied the correlation of happiness with time, economy and public opinion polls. The key findings are: 1) Twitter users are happiest after work (*e.g.*, 9 pm) and on weekends; 2) overall people are happier when making more money; and that the 3) smile index can be used to predict consumer index.

## 2.3 Popularity Prediction

In general, images make microblog posts more popular and survive longer than their text-only counterparts [162]; however, within image tweets, only small portion of them become popular (*e.g.*, being retweeted). What makes an image tweet popular? To answer this, two papers conducted case study by manually analyzing a small set of Twitter images under a specific event and investigating the correlation of an image tweet's retweetability and its image properties. From 290 Twitter images marked by"#guncontrol", Stefanone *et al.* [125] identified four image's psychological properties that are positively correlated to retweeting; namely, that fear and humor appeals, and that *attribute framing* [74] (*i.e.,* presenting an image in a way of highlighting a specific part that authors wish to draw attention to), and positive valences. Similarly, Kharroub *et al.* [68] collected 581 Twitter images that are tagged by one of the top five hashtags (*e.g.*, #jan25, #egypt) used in the 2011 Egyptian Revolution and studied the psychological role

[5] `http://www.mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data`

of the images as part the collective action. They found that emotionally arousing (*i.e.*, negative) violent content is not a positive factor that influence retweetability, in contrast to evolutionary basis discussed in previous work [96].

In another line, researchers attempted to build automatic popularity predictor for general image tweets. From a pure visual perspective, Cappallo *et al.* [23] formulated this as a ranking problem and proposed a latent ranking SVM to model the general popularity of image tweets by incorporating visual features from deep CNNs [70]. Not limited by visual cues, Can *et al.* [21] built a random forest regression model to predict retweet count using a combined model of text (*i.e.*, has hashtag or not), image (color histograms, GIST descriptors, detected objects), and social contextual (*e.g.*, follower count, followee count) features. They also showed that the incorporation of visual features significantly improve the prediction accuracy, and that high-level image features (*i.e.*, visual objects) are more robust than low-level image features.

Unlike the aforementioned two papers studied the general popularity, Bian *et al.* [10] predicted individual user's response (*i.e.,* to retweet or not) to microblog posts. To this end, they identified three key factors that influence a user's reposting decision: namely, her interest with respect to the post's content (both text and image), the social influence from the author of the post, and the epidemic effect (*i.e.*, the chances of the post also retweeted by her followees). They learned optimial weights for the three factors from users' previous retweeting history. By making the predictions over a social network, their model is not only able to predict the diffusion of a post, but also its total retweet count. One important application is to identify viral posts that will be retweeted over certain number of times at their emergence.

## 2.4 Events and Trending Topics

Due to the real-time and news media nature [71], microblogs have become an important venue to discover and follow breaking news and ongoing events. To relieve users from the information overload, the research community has worked on detecting, summarizing, and visualizing events and hot topics. These studies were originally conducted on text tweets, and then adapted to image tweets.

For visual event detection, we classify the prior works into three categories by the medium they utilized. The first type of study (*e.g.*, [67]) does not take the image into account, applying the existing text-based event detection algorithms (*e.g.*, keyword burst) to the textual portion of image tweets. In contrast, another work [49] leveraged the image only, applying near duplicate image detection to determine when an image meme emerges and trends among image tweets. The majority of works [139, 140, 9, 11, 19] model image and text jointly via topic models, and consider each learned topic as an event. Alternatively, these topic models are applicable for sub-events (sub-topics), declaring a hot topic when all input posts are from a particular event (topic). We will discuss these models in depth in Chapter 5.2.

Given an event, a subsequent task is to generate multimedia summaries from a large number of related tweets. One line of research [19, 147, 94, 67] focuses on selecting representative images only, while the other works [9, 11, 88] also generate textual summaries. For the former, the overall goal is to rank candidate images by considering both their relevance to the event and their visual diversity, and then selecting the top ranked images as visual summaries. To this end, graph-based approaches (*e.g.*, random walk) are commonly adopted. To further improve the images' quality, four works [88, 11, 19, 118] carried out preprocessing steps to filter out noisy images—

*e.g.*, memes, screenshots, reaction images—prior to selecting representative images. For the latter—textual summary generation, McMinn *et al.* [88] adopted an existing Twitter text summarization algorithm [121]; Bian *et al.* [9, 11] developed a greedy algorithm to sequentially select top ranked sample text by taking topic coverage, significance (*i.e.*, repost counts) and diversity into consideration; while Schinas *et al.* [118] selected the top posts ranked by their significance and topic coverage.

While most works focused on visual event detection and summarization algorithms, a select few have also developed demo systems to present events and their summaries vividly. Wang *et al.* [139] developed a magazine-style interface, and supplemented the microblog summaries with related web news. Nakaji *et al.* [94] and Kaneko *et al.* [67] projected the representative image tweets to an online map based on tweets' geotags. McMinn *et al.* [88] implemented an interactive interface to visualize events summaries along with event statistics (*e.g.*, the number of users discussing the event), background information about people and organization, links to relevant news articles, and the geo-locations of relevant tweets, etc. Cai *et al.* [18] visualized the events in different views, including evolution graph, timeline and map.

## 2.5 Credibility

Due to the user-generated nature, rumors—a piece of misinformation (false information) or disinformation (deliberately false information)—also spread in microblogs. Such information is misleading and may bring panic to the public during crises. To resolve this, researchers have also spent effort to distinguish deceptive information from trustworthy ones. Earlier works focused on detecting textual rumors [107, 146] and building classifiers leveraging features from tweets (*e.g.,* textual words, linguistic patterns, retweet

count) and users (*e.g.*, has a verified account, the age of account, the number of followers) who author or retweet the post. More recently, researchers turn to image tweets, as they observe multimedia content (*e.g.*, image) is often used in rumor propagation [84]. In 2015, a shared task "Verifying Multimedia Use"[6] was held in the MediaEval Benchmark and a set of 12K image tweets posted for 11 events were been released[7] with manual labels (fake or real) and extracted features from the tweet, user and image.

In contrast to the large number of text variants, images are often reused in rumor dissemination. Therefore, the number of unique fake images in a certain event is small, *e.g.,* eight fake images are identified in *Hurricane Sandy*[8] [46] and 16 in *Boston Marathon Bombing*[9] [15]. Considering this, image features are utilized to expand the ground truth dataset; *i.e.*, propagating labels to image tweets containing near duplicate images [15, 14] of already tagged ones. Following a similar idea, Jin *et al.* [64] grouped image tweets by image, and hypothesized individual tweets in the same group tend to discuss the same topic and thus have a similar credibility score. To be specific, they represented each topic group as the average features of all posts (tweet and user features, similar to previous work on textual rumor detection), and used the majority credibility label as the topic label. Based on this, they trained a topic-level classifier and used its score as additional feature for final, post-level classifier. Experiments show the image information significantly improve the classification performance.

Originating from classic forensic science, image forensics are a set of techniques designed to identify the source of a digital image or to determine whether the content is authentic or modified, without prior knowledge of the image under analysis [104]. As such, Boididou *et al.* [14] incorporated

[6] `http://multimediaeval.org/mediaeval2015/verifyingmultimediause`

[7] `https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2015`

[8] `https://en.wikipedia.org/wiki/Hurricane_Sandy`

[9] `https://en.wikipedia.org/wiki/Boston_Marathon_bombing`

image's forensic features (proposed by [33]) in their rumor image tweet classifier, in addition to the features mined from the tweet and user. The effectiveness of forensic features were proven in their experiment. Besides these two papers, another three papers [46, 15, 90] completely ignore image content or features in training their classifiers.

## 2.6 Search and Mining

Aside from events, image tweets are valuable resources for many mining and search tasks. We observe the search-based mining tasks usually consist of three steps: 1) keyword-based search in microblogs, 2) expanding the initial search results (optional), and 3) refining the final result set via ranking or filtering. Yanai *et al.* [144] collected photos for 100 types of food from Twitter: they started with searching tweets with pre-defined food related keywords, and then applied a general "foodness" classifier and a 100-class food classifier to remove noisy photos. Both classifiers were trained on low-level image features (*i.e.*, HOG patches and color patches coded into a Fisher vector representation); while in more recent work, the images are represented as the last layer output of a CNN [145]. Considering the real-time nature of microblogs, Zhao *et al.* [158] searched celebrity's recent pictures from Weibo. They first obtained a seed set of photos by querying Weibo with celebrity's name, and filtered out noisy photos with a pre-trained face recognizer. They expanded this set by pooling photos posted by active users and incorporating visually similar images, then employing a multi-modal graph based learning method to rank candidate photos by integrating social and visual information. In a similar vein, Gao *et al.* [41] gathered brand-related microblog posts from Weibo. Their seed set was constructed by searching Weibo with brand-related keywords, and refined by a pre-trained logo detector. Based on the seed set, they further mined

posts from the social context (*i.e.*, key users and key locations) and visual context, and then adopted a graph-based approach to filter out noise. In the same context, Qi *et al.* [108] focused on the refinement step, identifying brand-related posts from candidate set. With annotated dataset, they trained visual- and textual-based brand detector to assign a relevance score to candidates. To further rectify these scores, they proposed a graph-based regularization model that takes the visual and textual similarity of posts into consideration.

## 2.7 User-centric Studies

Researchers also exploit image tweets to understand users better. Guntuku *et al.* [45] predicted users' personality (*e.g.*, extraversion, agreeableness) from his/her selfies (*i.e.*, self-portait photos) since selfies are regarded as a medium of self-expression and self-representation. They first transformed low-level image features (*e.g.*, color, GIST) into mid-level representations (*e.g.*, face visibility, is photoshop edited or not) via classifiers, and then built personality classifier based on the mid-level features. In Guntuku *et al.* [45]'s work, selfies are identified by annotators. Unfortunately, manual labeling is time consuming. Given a user and her/his image tweets, can we find his/her selfies automatically? To answer this, Joshi *et al.* [65] made two key observations: 1) selfies often repeatedly appear in user's posts and 2) faces in selfies are usually large. Based on this, they first filtered out non-face photos by face detector, performed visual clustering to the resultant face-photos, ranked each cluster by a weighted combination of average visual similarity and average face size, and finally regarded photos in highly-ranked clusters as selfies. In a pure analysis study, Farci *et al.* [85] analyzed selfies posted by 10 popular politicians on Twitter, aiming to discover how framing effect of a selfie affects the politics of self-representation.

Aside from selfies, generic image tweets are also utilized to learn user's characteristics. Merler *et al.* [89] build a gender predictor based on features from user's profile (*e.g.*, user name, gender predicted from profile picture), posted images (*e.g.*, visual semantics), and text (*e.g.*, n-grams). Given a user and an image tweet, Yang *et al.* [148] predicted the user's emotional response to the post based on the user's historical interests and the emotional influence from social network. Lin *et al.* [78] worked on detecting psychological stress from an image tweet. They proposed a Cross-media Auto-Encoder (CAE) to learn the joint representation of an image tweet with features from text (*e.g.*, number of positive/negative words), image (*e.g.,* color theme, brightness) and the social attention (based on comment, retweet and favorite counts), and then trained a stress classifier based on the joint representation. In a follow-up study, they [79] aimed to predict the stress state for individual users, where the key was to learn a user-level feature representation from all of the user's posts and behaviors in a given period. To this end, they first utilized previous CAE model to obtain the post-level representation, and then proposed a CNN architecture to learn the combined representation for a series of posts. At the end, they trained a deep neural network (DNN) classifier to predict user's stress state using posts' combined representation and user's behavior statistics (*e.g.*, the frequency of social engagement).

In summary, most existing studies on image tweets are originated from text tweet domain. Often, these works focused on discovering effective image features generically (*e.g.* CNN), and synthesizing the information from visual and textual channels. However, many fundamental questions about image tweets, *e.g.*, their characteristics, user behaviors, microblog-specific image features that reveal the image's semantics, have not yet been explored. To fill this gap, this thesis conducted four foundational studies

on image tweets, which benefit these downstream applications.

# Chapter 3

# Understanding Image Tweets

## 3.1 Introduction

Image tweets have existed for a few years, since Weibo's launch in 2009 and later on Twitter's inline support in 2011[1]. However, we do not know them much, *e.g.*, the characteristics of image tweets, except their volume (constituting 56% of hot topic posts in Weibo [152]) and their popularity (retweeted more often and surviving longer than text tweets [162]). There have only been limited studies that have examined image tweets for specific populations—image content related to a specific event/topic [68, 120, 125, 43] or posted by users from specific locations [127].

To bridge this gap, we begin our thesis work by asking "what are the characteristics of image tweets". We collected a large set of 79,293,627 English and 133,352,913 Chinese posts from the public timeline API[2] of Twitter and Weibo in 2014, respectively (detailed in Table 3.1). We name them as Twitter-2014 and Weibo-2014 dataset. Based on these two, we conducted two comparative studies in terms of media (*i.e.*, image versus text tweets) and platform (*i.e.*, Twitter versus Weibo) to uncover the char-

[1] `http://techcrunch.com/2011/08/09/twitter-photo-uploading-now-available-for-100-of-users`

[2] Also known as public stream API on Twitter.

**Table 3.1: Demographics of Twitter and Weibo datasets collected in 2014.**

| | **Twitter-2014** | **Weibo-2014** |
|---|---|---|
| No. of posts | 79,293,627 | 133,352,913 |
| % of image tweets | 14.1% | 56.0% |
| Collection method | Public stream API | Public timeline API |
| Collection period | Apr 15 to Jun 15, 2014 | Nov 7, 2014 to Jan 17, 2015 |

acteristics of image tweets and the underlying user behaviors. To show the evolution of image tweets, we additionally gathered a set of Weibo and Twitter posts from 2012 and conducted similar analysis.

## 3.2 Related Work

As the prominent microblogs, Twitter and Weibo have attracted research to assess their similarities and differences from many aspects, including functionality [28, 152], users behavior [39, 40] and posted content [39, 152]. However, none of these studies focused on image tweets.

While the basic services of the two platforms are similar, Weibo is known notably for its use of rich media (*e.g*, allowing embedded images, videos, music, emoticons and even polls), interactivity (*e.g.*, threaded comments) and for incentivization (*e.g.*, a badge system to encourage users to tweet more) [28, 152, 151, 157]. Due to the popularity of these features, Twitter has also adopted similar functionality (*e.g.*, enabling direct image uploading and threaded conversations).

In a study conducted in 2012 [39], web interfaces were the most popular way to access the microblogging services—contributing 43.1% of Weibo posts and 38.5% of Twitter posts—however, users of the two were active in different time—Weibo users post more on weekends but Twitter users on weekdays. In the same year, in a separate work on information propaga-

tion, the same authors discovered that Twitter users reposted tweets more frequently and much faster than users of Weibo [40].

To answer what kinds of posts are more popular, Yu *et al.* [152] analyzed posts from trending topics and the most influential users in 2011. On Twitter, the majority of influential users were news agencies, thus making trending topics often coinciding with news events. This is in accord with Kwak *et al.* [71]'s claim—Twitter may be better characterized as a news medium rather than a social network. On Weibo, the dominant influential users were unverified accounts, with a strong focus on collecting user-contributed jokes, movie trivia, quizzes and stories. As such, they suggested that Twitter users were more tuned towards news events, while Weibo users were more inclined to share and propagate trivia.

From a linguistic perspective, Gao *et al.* [39] analyzed posts on Twitter and Weibo at the syntactic, sentiment and semantic levels. From the syntactic aspect, the usage of hashtags and URLs on Twitter was 3.2 times and 1.97 times more intensive than on Weibo. Interestingly, Twitter users were more inquisitive (as measured by the percentage of tweets using question marks), while Weibo users tended to exclaim (as similarly measured by tweets with exclamation marks). With respect to semantic concepts, organizations (*e.g.*, companies, institutions) were more likely to be referred to in Twitter posts, while locations and persons appeared more often on Weibo. With respect to sentiment, both platforms posted more positive tweets than negative ones, and Weibo users showed a stronger tendency to publish positive messages than Twitter users.

## 3.3 Image Characteristics

An idiosyncratic factor is that all embedded images on Weibo and Twitter are processed by the uploading agent which imposes certain restrictions and

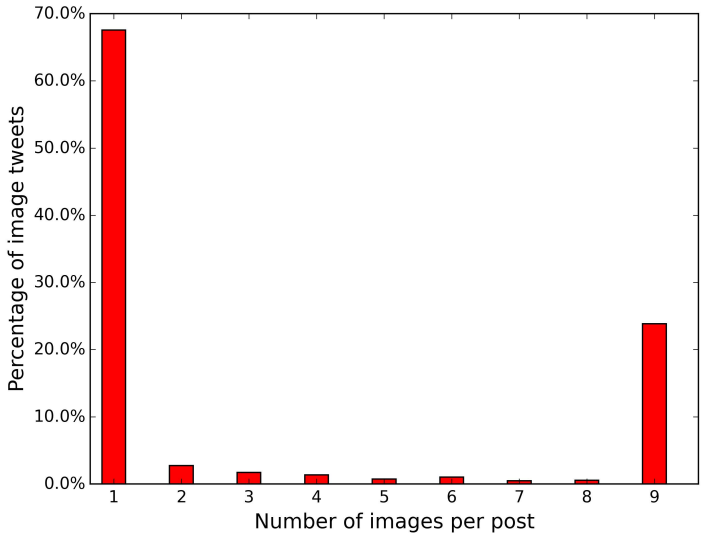



**Figure 3.1: The image number distribution for image tweets in Weibo-2014 dataset.**

post-processing: 1) both platforms allowed only one image per post when image tweets were first introduced, and later have added supported up to 9 images since April 2013[3], and 4 images since March 2014[4] for Weibo and Twitter, respectively; 2) images on both platforms are scrubbed of their Exif metadata; and 3) all images on Weibo (excluding animated GIFs) are converted to JPEG whereas Twitter accepts JPEG, PNG formatted still images and animated GIFs.

In our Weibo corpus, 56.0% are image tweets, of which still images dominate: 97.5% image tweets contain a JPEG formatted picture while 2.5% contain an animated GIF. In Twitter corpus, 14.1% are image tweets, and of which 95.1% of images are JPEG format. On Weibo, 32.4% of image tweets contains more than one image, and a surprisingly 23.9% have 9 images. We plot the distribution of image tweets in terms of image numbers in Figure 3.1. Since our Twitter dataset was collected just after Twitter launched its up-to-four images per post functionality, only 0.017% of Twitter image tweets contain multiple images.

[3] `http://www.techweb.com.cn/internet/2013-04-23/1291915.shtml`
[4] `https://blog.twitter.com/2014/photos-just-got-more-social`

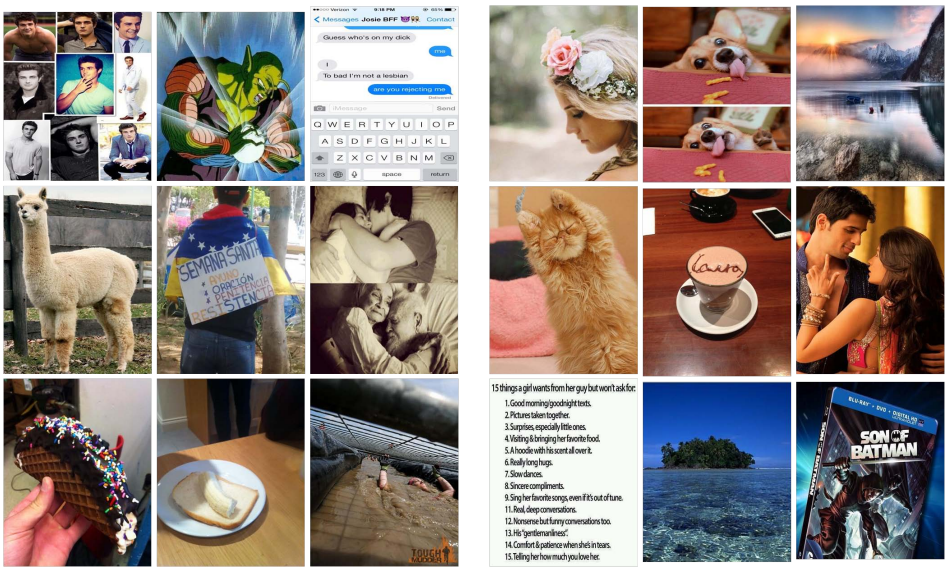



**Figure 3.2: 18 example images from Twitter.**

The proportion of image tweets on two platforms differs a lot (56.0% versus 14.1%). We naturally ask why Twitter users post far less images than Weibo users? We attribute this to two main causes. On one hand, Twitter supports direct image uploading only quite recently (in 2011[5], five years after its launch). In contrast, Weibo had image tweet functionality since its debut (as early as 2009). On the other hand, Twitter faces intense competition from the pure photo-based SNS (*e.g.,* Instagram and Flickr), which winnowed its share of image content; while Weibo does not have such strong competitors in China.

Figure 3.2 and 3.3 list example images from Twitter and Weibo, respectively, which illustrate the variety of images in microblogs: there are photographs of varying quality (both candid and composed) and of varying topics (screenshots, cartoons, digital wallpaper and other forms of decorative images). We thus hypothesize that microbloggers care more about the photo content than quality, as most photos seem to be of low quality, which differs with Flickr [51]. Additionally, we observe a distinct form of image—multi-photo collage—on both platforms. The collage form was originally

[5]`http://techcrunch.com/2011/08/09/twitter-photo-uploading-now-available-for-100-of-users`

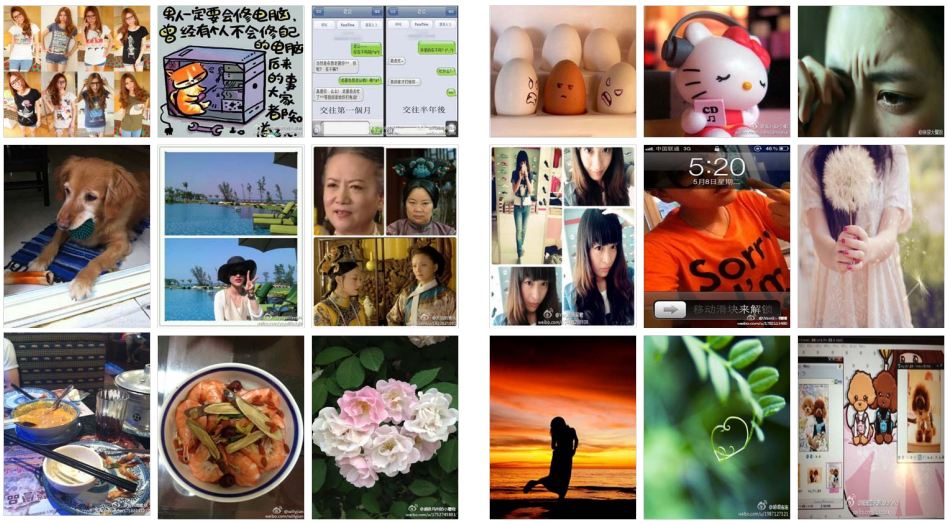



**Figure 3.3: 18 example images from Weibo.**

used for bypassing one-photo-per-post limitation but is also used for different narrative purposes: *e.g.*, to compare objects, and tell stories through an image sequence. The collage form was even encouraged by Weibo in that its web interface provides a toolkit to generate collages automatically.

## 3.4 Access Behavior

We analyzed the devices that people used to post tweets on the two microblogging platforms. For Twitter posts, we manually classified the most frequently used 100 devices into seven categories; namely, mobile client, mobile browser, tablet, desktop browser, desktop client, web client, and third-party web/app (*e.g.*, sharing a post from another website or a by-product activity of using another app). For Weibo posts, we followed the same process, but identified two additional by-products—Weibo game and Weibo activity—that post a tweet automatically for a game or an activity participant. Such by-products are not real, user-generated content by our definition.

Overall, mobile devices predominate on the two platforms, being responsible for 71.0% and 42.6% of posts on Twitter and Weibo, respectively. We

further dissect the devices by post type (*i.e.*, image or text), and display the distributions in Table 3.2. On Twitter, mobile clients generate 75.8 % of image tweets, which is 12.3% higher than its contribution to text tweets. Among them, the official mobile clients, *e.g.*, Twitter for iPhone[6], are the most widely-adopted applications. On Weibo, official mobile applications dominate the market of mobile clients as well; however, the most popular devices for text tweets are not mobile clients, but the traditional desktop browser, accounting for 51.1% of text tweets.

We note that the device of taking pictures and posting image tweets may not always be the same, *e.g.*, taking a photo by a smartphone, but composing the image tweet from a laptop if transferring the photo to the laptop in advance. Are mobile phones still the most popular photo-taking devices? Unfortunately, we cannot answer this question directly since Exif metadata of images has been stripped. Early in 2012, we surveyed 109 Weibo users employed in *Zhubajie*[7] (a Chinese crowdsourcing website) [30] as well as students who were Weibo users within our university. 85% of the respondents self-reported that they primarily took photos by a camera phone, whereas 13.7% used a digital camera. We believe the penetration of camera phones is even higher now, since the market share of smartphones has steadily increased since 2012, when the survey was conducted. This agrees with our hypothesis that microbloggers care more about the photo content than quality since generally cameraphones are unable to capture photos in high quality as professional digital cameras.

Unlike the previous work [39], the desktop browser is no longer the largest contributor except Weibo's text tweets (51.1%), but has been relegated to the second rank, taking up 10.0% of text tweets and 7.7% of image tweets on Twitter, and 14.5% of image tweets on Weibo. Web-based

[6] `https://about.twitter.com/products/iphone`
[7] `http://www.zhubajie.com`

**Table 3.2: The distribution of devices used in Twitter-2014 and Weibo-2014 datasets.**

| Level-1 | Level-2 | Twitter | | | Weibo | | |
|---|---|---|---|---|---|---|---|
| | | Text | Image | Overall | Text | Image | Overall |
| Mobile | Mobile Phone Client | 63.5% | 75.8% | 71.0% | 29.3% | 51.4% | 42.6% |
| | Mobile Phone Browser | 2.2% | 0.8% | | 0.1% | 0.1% | |
| | Tablet | 3.6% | 5.2% | | 0.7% | 0.8% | |
| Non-Mobile | Desktop Browser | 10.0% | 7.7% | 20.2% | 51.1% | 14.5% | 38.1% |
| | Desktop Client | 1.0% | 0.5% | | 0.7% | 0.6% | |
| | Web Client | 5.4% | 5.2% | | 0.7% | 12.0% | |
| By-products | 3rd Party Web/App | 5.3% | 0.3% | 4.6% | 4.2% | 15.7% | 16.0% |
| | Weibo Game | | | | 3.1% | 1.0% | |
| | Weibo Activity | | | | 6.7% | 8.9% | |
| Uncategorized | | 8.9% | 4.5% | 8.3% | 3.4% | 3.0% | 3.1% |

clients, online software that usually schedules posting and manages a user's accounts in multiple social media websites, is the third largest contributor, but has a strong tendency to be used to post image tweets (12.0% of all image tweets versus 0.7% of all text tweets) on Weibo. A manual inspection depicts many of the posts from the web software are advertisements and trivia (*e.g.*, jokes and quotes), and thus often contain an image. Interestingly, this trend is not the case for Twitter, where approximate same proportion of image (5.2%) and text (5.4%) tweets are originated from web clients. Finally, 16.0% of Weibo posts are by-products of using another app, playing a Weibo game or participating a Weibo activity, whereas only 4.6% of Twitter posts are by-products.

## 3.5 Temporal Behavior

By studying the whole Twitter public statues posted from May 2010 to May 2011, Naaman *et al.* [92] shown that Twitter experienced greater volume during the daytime, and less late at night and in early morning (*e.g.*, 4–5 am). In our work, we focus on studying the temporal patterns of image tweets on both Twitter and Weibo. We plot our tweets by posting day and hour in Figure 3.4. The posting time of Twitter has been converted to the

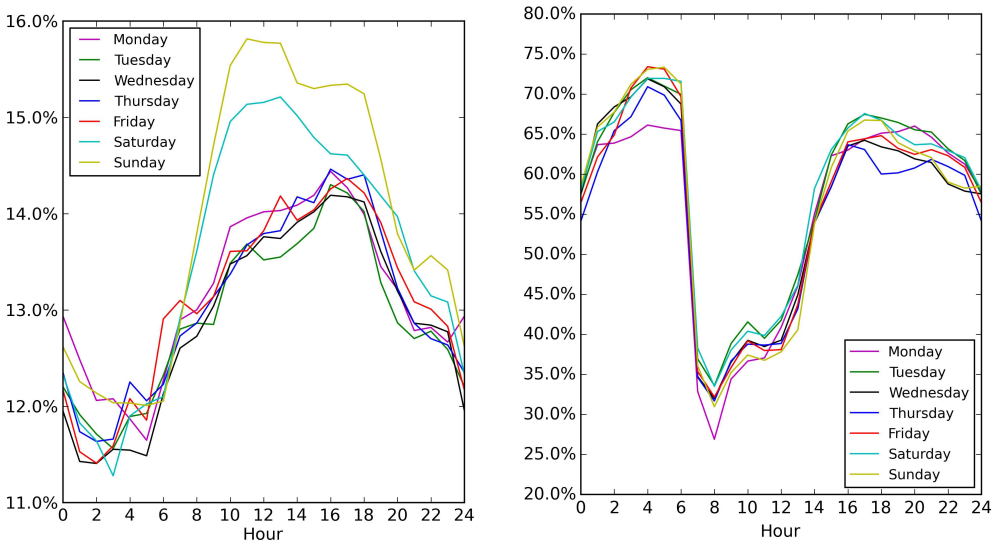



**Figure 3.4: Percentage of image tweets by hour in Twitter-2014 (left) and Weibo-2014 (right) datasets.**

user's local time based on her or his self-reported timezone in profile. Weibo does not provide user's timezone information, therefore, we assume all users are from UTC+08:00, the only timezone used in China. Considering the vast majority of Weibo users live in China, this assumption is not overly presumptuous.

From Figure 3.4, we observe that image tweets are posted more frequently during the daytime and the weekend, and are most actively posted in the afternoon. Such trend is consistent with the overall diurnal activity for Twitter users, as revealed in Naaman *et al.*'s study [93]. We posit that there are more tweet-worthy objects and events during the day, but we have yet to validate this.

Interestingly, Weibo also peaks during 4 am to 5 am, which not only contradicts our intuition but also the phenomenon on Twitter. The first cause comes to our mind is the timezone issue—the night of China may be the daytime of other countries. Does the influence of overseas users become evident during this period? To validate this, we examined the locations of users posted during 4:00 am to 5:59 am. Users from countries

in the day (*e.g.*, US, Canada and European countries) do exhibit a higher tendency to post images tweets (larger than the overall 56.0%), however, China domestic users are still the biggest contributors, authoring 85.6% of all posts during this period.

How could Chinese microbloggers be so active in the early morning? We realize tweets are not always posted by real users, since many web clients support scheduled posts. Our analysis shows during this period 22.8% of the posts were produced by web clients and 97.8% of which contain an image. With a further investigation, we find the amount of software emitted posts and their proportion of image tweets are stable over 24 hours. Consequently, its influence to the ratio of image and text tweets is boosted when the non-scheduled posts are decreased greatly, *i.e.*, during the night and early morning. This the primary cause for the second peak (4 am to 5 am) of posting image tweets on Weibo.

## 3.6 Medium and Reaction

We differentiate among three important types of tweets by user's action, namely, *original posts* (an initial tweet that may elicit retweeting or replying), *retweets* (a re-posting of someone else's tweet), and *replies* (a tweet that responds to another tweet). On Twitter, 47.6% of tweets are original, 28.6% are reposts, and the remaining 23.8% are replies. We are unable to directly report the similar statistics for Weibo for the reason that Weibo's public timeline API (the one we used for data gathering) samples tweets only from the original ones. Instead, we randomly sampled 1 million posts from our large Weibo dataset, and relied on an additional API[8] to retrieve the comment and repost count for each tweet[9]. Via this method,

[8] `http://open.weibo.com/wiki/Statuses/counts/en`

[9] We made the requests in December 2015.

**Table 3.3: The distribution of post types in Twitter-2014 dataset.**

| Proportion | Text | Image | Overall |
|---|---|---|---|
| Original Post | 50.6% | 29.3% | 51.1% |
| Retweet | 22.6% | 66.6% | 28.6% |
| Reply | 26.9% | 5.1% | 20.4% |

**Table 3.4: The distribution of user responses in Weibo-2014 dataset.**

| Proportion | Text | Image | Overall |
|---|---|---|---|
| No interaction | 88.5% | 88.3% | 88.4% |
| Retweeted | 1.7% | 3.5% | 2.7% |
| Replied | 10.6% | 10.2% | 10.4% |
| Retweeted and replied | 0.8% | 2.0% | 1.5% |

the sampled Weibo set consists of 88.4% of tweets which had no user interaction (neither were retweeted nor replied), 1.5% were both retweeted and replied by other users, 10.4% were only replied, and the remaining 2.7% were retweeted. Note it does not mean only 2.7% of all traffic on Weibo are retweets, because some tweets have been retweeted many times. Moreover, we can not make conclusion on which platform's users are more engaged in interaction, due to the two groups of statistics being not strictly comparable.

We further break down the posts by medium, and show the distributions in Table 3.3 and 3.4 for Twitter and Weibo, respectively. As can be seen, image tweets on both platforms were retweeted more often than the text-only posts. On Twitter, 66.6% of image tweets are retweets, while only 22.6% of text tweets are shared from someone else's posts. It implies an image increases the retweetability of a post by almost double. Similarity, on Weibo, the retweetability of image tweets is double that of text posts: 3.5% of image tweets were retweeted but only 1.7% text tweets were retweeted. This confirms the findings from previous work [162, 44, 152].

## 3.7 Content Analysis

A picture is worth a thousand words. Do image tweets still need texts? Though both platforms allow users to post an image without accompanying text, it is rare—95.0% of image tweets on Weibo and 97.8% on Twitter have corresponding text. Note URLs of images on Twitter are not considered as user's input. This leads us to ask the follow-up questions of "how many words is an image worth?" and "does the embedded image make the text succinct?"

To answer these questions, we segmented Weibo (Chinese) posts and similarly tokenized Twitter (English) posts, and then computed the average number of words in the posts (shown in Table 3.5). Microblog specific notations (*e.g.*, hashtag, mention, URL and emoticon), punctuations and digits are regarded as a single token. Both microblogging services have a 140-character length restriction; however, Weibo posts tend to have more words than Twitter posts. The reason is related to the languages—English words consume more characters (4.5 characters per word [103]) than Chinese words (1.59 characters [97]) . Within the platform, Twitter's image tweets are slightly shorter than that of the text tweets (10.7 versus 12.0 average number of words), while Weibo exhibits the opposite—30.5 words on average for image tweets and 23.0 words for text tweets. In this sense, including an image does not always make the text more concise.

Plotting the length distribution by medium in Figure 3.5, we observe that the length of image and text tweets are similar on Twitter, both peaking around ten words, while Weibo's posts exhibit different peaks for image and text tweets. On Weibo, a surprisingly large portion of image tweets are extremely short: 7.2% have two words, and 8.9% have three words. What are these short posts about? To answer this, we examined the text of these posts, and found that a large portion of text are not genuine

**Table 3.5: The average length of posts in Twitter-2014 and Weibo-2014 datasets.**

| No. of words | Text Tweets | Image Tweets |
|---|---|---|
| Twitter | 12.0 | 10.7 |
| Weibo | 23.0 | 30.5 |

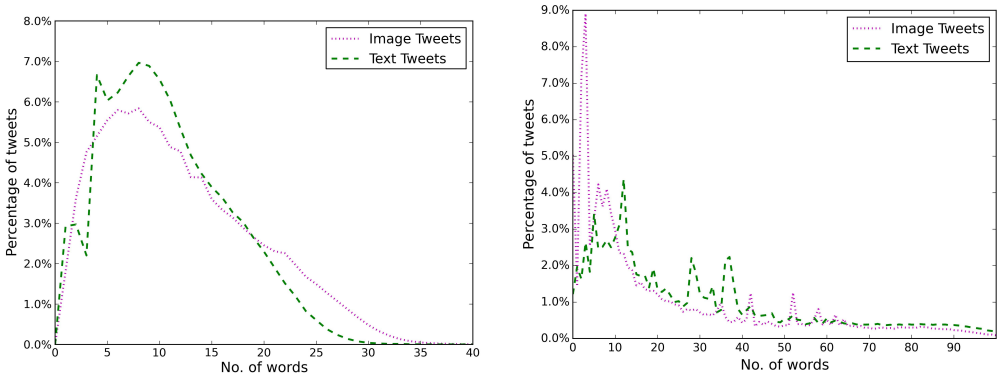



**Figure 3.5: Percentage of text and image tweets by number of words in Twitter-2014 (left) and Weibo-2014 (right) datasets.**

user generated text. For instance, 25.8% of two-word posts are "分享图片" ("share image" in English), which is the default text when users upload images. For three-word posts, 32.5% are "链接:URL" ("Link: URL"), and 29.0% are "地址:URL" ("Address: URL"). These two kinds of text are automatically generated by Weibo when users share an external web page from certain websites. Similar reasons are responsible for the peak at 12 words and a few sudden rises (*e.g.*, 28 words and 36 words) for text tweets on Weibo. For instance, 36.6% of 12-word textual posts are automatically generated by an online game (*i.e.*, by-products) and thus identical. When removing the automatically generated posts, we hypothesize both image and text tweets on Weibo actually exhibit similar word length distribution, peaking around 9 words.

Asides from length of posts, the usage of microblog conventions, *i.e.*, mentions (@username, serving as a conversational purpose), hashtags (marking the keywords or key topics of a post) and URLs (relating to external information sharing), also interest us. Table 3.6 lists the percentage of

**Table 3.6: The usage of microblog conventions in Twitter-2014 and Weibo-2014 datasets. Note URLs of Twitter images and URLs of geolocations on Weibo are not considered in the URL analysis, and "@username" of replies and retweets are excluded in the mention analysis.**

| % of Tweets contain | Twitter | | Weibo | |
|---|---|---|---|---|
| | **Text** | **Image** | **Text** | **Image** |
| Mention | 55.2% | 76.3% | 21.7% | 9.4% |
| Hashtag | 14.9% | 17.8% | 16.3% | 11.6% |
| URL | 18.3% | 8.3% | 41.2% | 58.8% |

posts contain each of the conventions for the two datasets. Overall, Twitter users are more interactive, adopting mentions in 55.2% of text and 76.3% of image tweets, while Weibo users are more prone to share external links, embedding URLs in 41.2% of text tweets and 58.8% of image tweets. In a breakdown analysis by medium, we observe text tweets and image tweets exhibit exactly the opposite tendency on the two platforms, *e.g.*, Twitter's image tweets (76.3%) are more likely to contain mentions than text tweets (55.2%), while this phenomenon is reversed on Weibo (21.7% of text tweets versus 9.4% image tweets). We hypothesize the discrepancy and the high percentage of URLs in Weibo is caused by the non-user generated posts on Weibo, *e.g.*, by-products and scheduled posts, which take a large portion of the total traffic (ref Table 3.2). Most of these posts serve as promotional purpose (*e.g.*, advertisement) and thus often embed an external URL.

At a deeper level, we examine what people tweet about. To this end, we applied latent Dirichlet allocation (LDA; [13]) to a large, ∼1M randomly sampled subset of the Twitter and Weibo datasets, to learn $k = 40$ ($k = 50$, in Weibo) latent topics, where $k$ was tuned on a held-out set. Among these topics, we observe that some exhibit an image to text tweet ratio differing significantly from the overall 14.1% (56.0%). Figure 3.6 and 3.7 lists sample topics with manually-assigned labels for Twitter and Weibo, respectively.

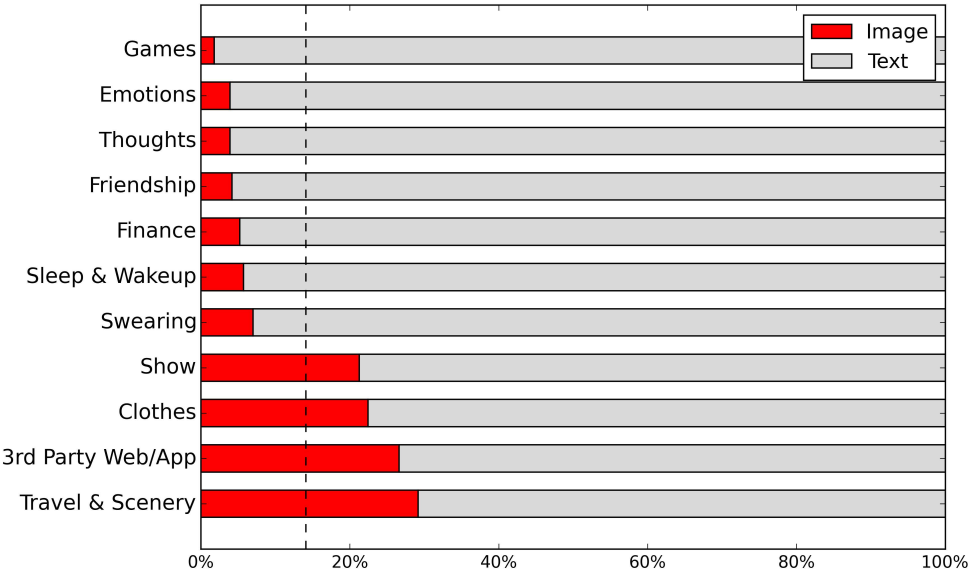



**Figure 3.6: Percentage of image to text posts of Twitter-2014 dataset in skewed topics. The overall percentage of image tweets in the dataset is 14.1%.**

On both platforms, posts on fashion and by-products originating from third party web/app are most often adorned with images, while posts about emotions and everyday routine are mostly text-only. Game-related tweets exhibit differences on the two platforms—often accompanied with images on Weibo, but are mostly pure text on Twitter. Again, this is due to Weibo Games, which automatically generate the majority of game related tweets and often include game's logo in the tweet.

## 3.8 Evolution

Image tweets have existed for a few years but how have they evolved since their birth? To answer this, we conducted parallel analyses on a set of Weibo and Twitter posts collected in 2012. We collected a corpus of 57,595,852 Weibo posts from the public timeline API of Weibo over a period of 7 months in 2012 (hereafter, Weibo-2012), and obtained another set of 332,013,806 English Twitter posts collected by NUS-Tsinghua NExT

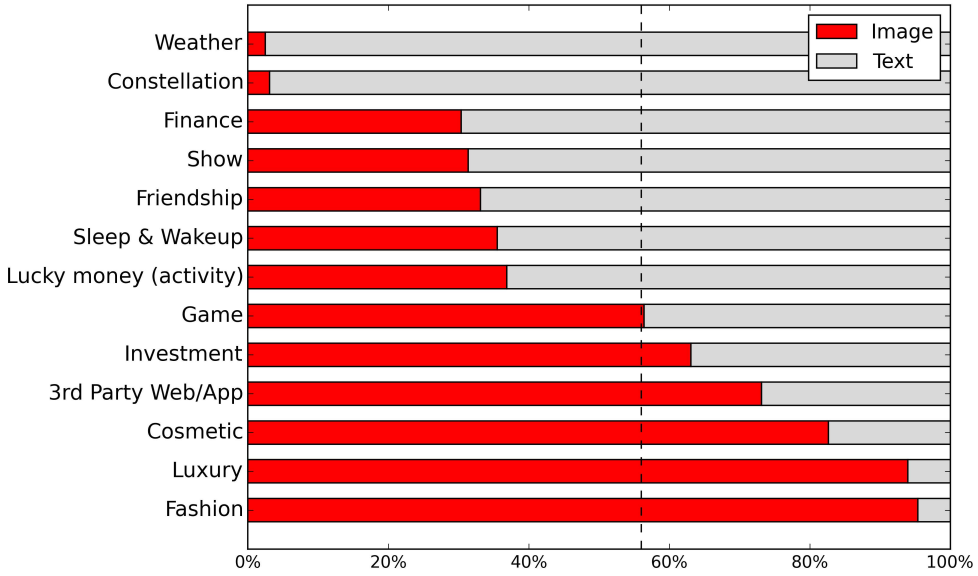



**Figure 3.7: Percentage of image to text posts of Weibo-2014 in skewed topics. The overall percentage of image tweets in the dataset is 56.0%.**

**Table 3.7: Demographics of Twitter and Weibo datasets collected in 2012.**

| | **Twitter-2012** | **Weibo-2012** |
|---|---|---|
| No. of posts | 328,254,527 | 57,595,852 |
| % of image tweets | 3.8% | 45.1% |
| Collection method | Public stream API | Public timeline API |
| Collection period | Jan 1 to Dec 31, 2012 | Apr 1, 2012 to Oct 31, 2012 |

center in the whole year of 2012 (hereafter, Twitter-2012) [31]. Table 3.7 details the demographics of the two datasets. We compare the statistics with their counterparts in 2014 datasets, and summarize the significant changes in the following.

**Percentage of image tweets.** During the two years, the use of image tweets increased for both platforms. On Twitter, the percentage of posts contain an image surges from 3.8% (2012) to 14.1% (2014); while Weibo demonstrated a smaller gain, increasing from 45.1% (2012) to 50.0% (2014). From this, we feel there is still a large room for image tweets to grow in Twitter, but is approaching stability in Weibo.

**Image characteristics.** The most evident change for images is that multi-photo collages were often adopted in 2012, but this use has waned by 2014. We hypothesize many use cases of such collages have been replaced by the multiple-photo-per-post functionality.

**Access behaviors.** On Twitter, slightly more tweets (both image and text) were generated by mobile devices in 2014 (71.0% in total) than in 2012 (65.4%). We see the similar trend for Weibo's image tweets (43.8% in 2012 vs. 52.3% in 2014), but not text tweets (52.3% vs. 30.1%). The reason is that Weibo users have primarily adopted mobile devices to post image tweets: 69.1% of posts by mobile devices are image tweets in 2014, but only 41.3% were in 2012.

**Temporal behaviors.** In Figure 3.8, we plot the percentage of image tweets by hour for the 2012 datasets. When compared to the 2014 datasets (Figure 3.4), Twitter users did not change their temporal behaviors much; but in contrast, Weibo users do exhibit different temporal behaviors between 2012 and 2014. Looking at the posting day, Weibo users in 2012 were more actively in sharing images on weekends, whereas these differences were muted in 2014. Focusing on the posting hour, the anomalous peak in the early morning (still caused by scheduled image tweets) occurs in both years; however, the peak value in 2012 ($\sim$47%) was much smaller than its counterpart ($\sim$75%) in 2014.

**Medium and Reaction.** From our analysis (see Table 3.8), we found more Weibo's posts (88.4% overall)—regardless of medium—were not reposted or commented than posts in 2012 (67.1%). This suggests that Weibo's posts are losing users' attention. We feel the primary reason for this is the increasing number of non-user generated posts, *e.g.*, Weibo activities, scheduled posts, which are not worthy of user interaction. We omit discussion of the results for Twitter in 2014, as they did not change much since 2012.

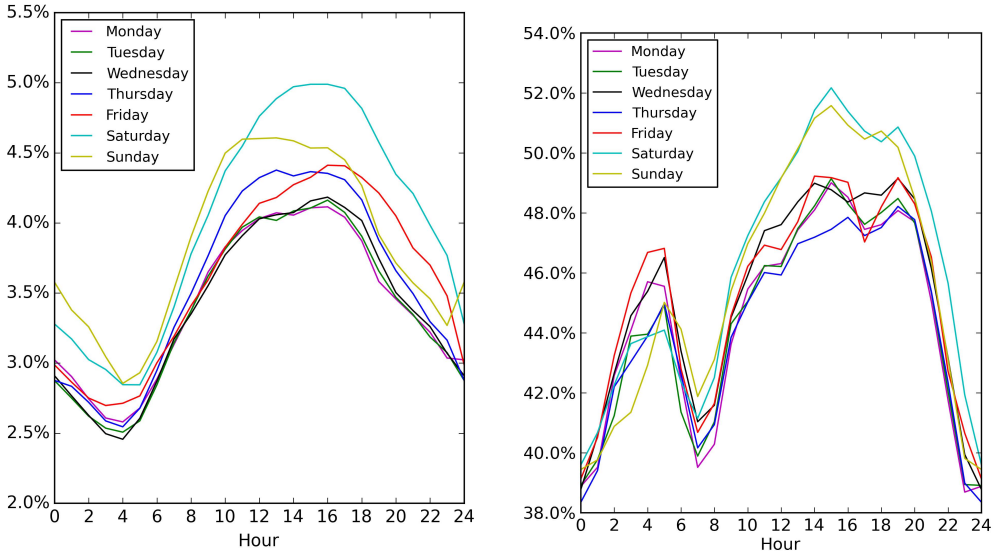



**Figure 3.8: Percentage of image tweets by hour in Twitter-2012 (left) and Weibo-2012 (right) datasets.**

**Table 3.8: The distribution of user responses in Weibo-2012 dataset.**

| % of tweets contain | Text | Image | Overall |
|---|---|---|---|
| No interaction | 69.5% | 64.4% | 67.1% |
| Retweeted | 3.3% | 4.9% | 4.1% |
| Replied | 22.3% | 21.9% | 22.1% |
| Retweeted and replied | 4.9% | 8.9% | 6.7% |

**Content.** On both platforms, the topics that users were interested in were relatively stable over time. For Weibo, we observe users' usage of microblog conventions changed a lot (see Table 3.9). With respect to image tweeets, Weibo users now seek less interactions via mentions (22.6% in 2012 vs. 9.4% in 2014), participate in fewer discussions via hashtags (13.9% vs. 11.6%), and embed much more external URLs (1.5% vs. 58.8%). Again, the primary cause is the non-user generated content. These posts aim at promotion, *e.g.*, attracting users to click the URLs to visit external sites, but not interaction.

**Table 3.9: The usage of microblog conventions in Weibo-2012 dataset.**

| % of tweets contain | Text | Image |
|---|---|---|
| Mention | 14.0% | 22.6% |
| Hashtag | 6.8% | 13.9% |
| URL | 13.7% | 1.5% |

## 3.9 Conclusion

The social Web 2.0 has embraced multimedia with the inclusion of facilities to embed images in microblog posts. Over half (56.0%) of posts on Weibo and 14.1% of Twitter posts contain at least one image. We performed a multipronged analysis of these *image tweets* in Twitter and Weibo from the perspective of image characteristics, user posting behaviors and textual contents.

On both platforms, image tweets 1) have images that are primarily single JPEG-formatted pictures, cover a wide range of content variety and are often of low quality; 2) are more retweetable than their text-only counterparts; 3) posted predominately by mobile phones; and 4) are actively posted in the daytime and weekend. We also found that the choice of medium (*i.e.*, with image or text only) is highly correlated with the topics discussed in the post.

On the other hand, users on the two platforms do exhibit different behaviors of posting image tweets. On Weibo, over 25.6% of image tweets are by-products of using third party applications, participating Weibo activities and playing Weibo games, and 12.0% are scheduled posts by Web clients. Posts from these sources are not genuine user generated content and their primary purpose is promotion, leading to a few abnormal phenomena on Weibo: 1) image tweets are extremely actively posted during early morning; 2) 58.8% of image tweets on Weibo embed external URLs,

whereas URLs only appear in 8.3% of Twitter image tweets.

Applying the same analysis to an older Twitter and Weibo dataset, we document a moderate evolution of image tweets on both platforms: Twitter has significantly increased the proportion of image tweets, but user posting behaviors are relatively stable; and Weibo demonstrated a small gain in terms of the proportion of image tweets, but user behaviors have changed, partially due to the increasing significance of managed social mean via by-product posts and scheduled posts.

Findings in this chapter lay the foundation for this thesis, guiding our subsequent three studies from various perspectives. As we found the majority of image tweets have both image and text, we are motivated to study the relationship between these two modalities (Chapter 4 and Chapter 5). In particular, we are inspired to encode textual topics and posting time as features for image-text relation classification (detailed in Chapter 4.4.2), and use hashtag query to obtain a set of high quality image tweets (detailed in Chapter 5.5.1). Moreover, the analysis in this chapter gives insights on the contexts for image tweet semantic modeling (Chapter 6). For example, we observed many microblog images are synthetic images that contain text, and thus used the embedded text in image as a context. Similarly, hashtags and external pages directed by embedded URLs are identified as the other two contexts.

# Chapter 4

# Identifying and Classifying Image-Text Relations

## 4.1 Introduction

A picture speaks louder than words. It is true that image tweets are retweeted more often than text-only tweets. However, images still needs text—95.0% and 97.8% of image tweets on Weibo and Twitter, respectively, are accompanied by text as described earlier. Why do people post both image and text and what is the nature of their correlation? To answer these questions, we 1) deconstruct the corpus to characterize such posts' image and textual content and the correlation between the two; 2) collect annotations for a subset of these image tweets in the corpus and; 3) build an automated classifier to distinguish two important subclasses of image tweets—*visual* and *non-visual* tweets.

## 4.2 Related Work

As text and image co-occur everywhere (*e.g.*, on the Web, picture books, dictionaries, journals, among others), researchers are interested in exploring

how image and text interact with each other. Most existing work focuses on image-text relations within a specific domain; including education [75, 101, 24], children's literature [119, 98, 123], journalism [131], dictionaries [48], and information design [8]. As a result, the definition of the image-text relations and the taxonomies necessarily vary from one to another. As these domains are not of interest to this thesis, we will not elaborate these works, but examine two studies in depth that target image-text relations for general domains.

From functional perspective, Marsh *et al.* [86] defined image-text relations as how an image is functionally relevant to its text. Their proposal, a hierarchical catagorization scheme with 49 subcategories, further grouped into three major categories (*i.e.*, little relation, close relation, and relevant but goes beyond text) based on closeness, was developed by examining a variety of taxonomies from previous studies and doing a content analysis of 954 image-text pairs from the Web.

In a spirit of mutual interaction (unlike [86] which assumes image is supplementary to text), Martine *et al.* [87] studied the relative importance of image and text (*i.e.*, equal or unequal), as the first kinds of image-text relations. The second proposal, from a logical and semantic angle, is whether one expands upon or repeats the meaning of the other medium.

While insightful, these categorizations are either limited to a specific subject area or predate image tweets and do not cater for the textual content found in social media. Furthermore, neither scheme has been operationalized into an automated classifier. This motivates us to propose a new classification scheme for image-text relations in microblogs.


**Figure 4.1: Image tweets of Weibo with their corresponding text, image and translation. The top two are examples of *visual* tweets, and the bottom two are *non-visual* ones.**

## 4.3 Image and Text Relation

It seems natural to assume that the two mediums should complement each other—an embedded image should present visual highlights of the post, where the text gives contextual description: time, location, event or story. That is, the text and the image are visually related, and as such we deem both media to be of equal standing. We define *visually-relevant* image tweets (*visual* for short) as ones where at least one noun or verb corresponds to part of the image.

In our corpus analysis, we did observe this behavior, but interestingly, there was a surprisingly large proportion of *non-visual* image tweets, where the text and image have little or no visual correspondence. These are hard to detect by just looking at the images themselves: actually, in Figure 3.3 (Chapter 3), the left group of 9 images are from visual image tweets and

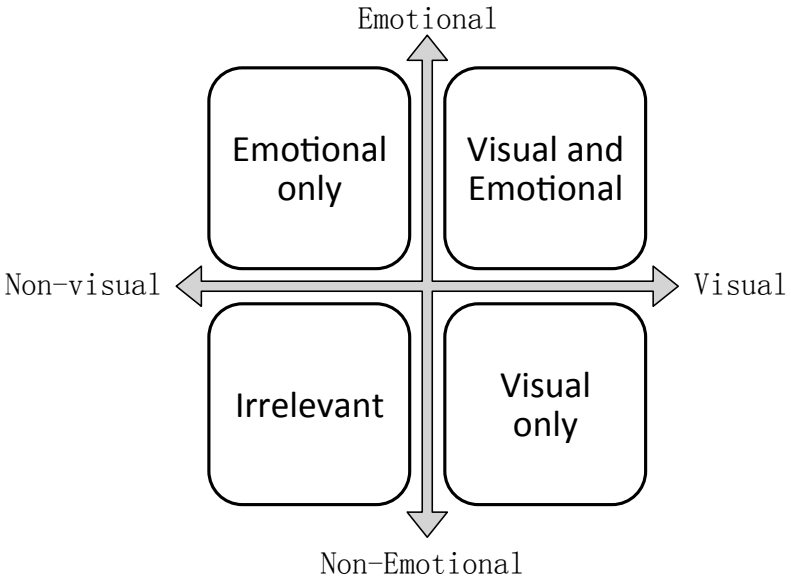



**Figure 4.2: The possible image-text relations.**

the right group of 9 are non-visual image tweets of Weibo. We find the distinction hinges jointly on text and image content together. Figure 4.1 shows two sample visual (top) and two sample non-visual (bottom) posts of Weibo. The motivations for posting images in a non-visual post vary. In the bottom row, the poster embedded an outdoor landscape which has no correspondence to the text, but which may entice readers to view the post.

We notice that a subset of non-visual image tweets that exhibit a consistent characteristic: that of *emotional* relevance. In such tweets, the text and the image share the same emotional state, as in the third example—anger, directed at "蚊子" (mosquitoes). In such cases, the text is the primary medium; the image reinforces the emotional aspects of the text, similar to emoticon use. As visual and emotional relevance are not exclusive, an image and text pair can exhibit one of four possible relations, namely, visual only, emotional only, both visual and emotional, and irrelevant (*i.e.*, neither visual nor emotional). We illustrate the four relations in Figure 4.2.

To test whether these findings are corroborated by actual users, we

**Table 4.1: Distribution of responses to the survey question "What is the primary reason that you insert an image in a tweet?"**

| | |
|---|---|
| A1. A picture is worth a thousand words, which is visually related to the text and makes the text succinct (*visually relevant*). | 29.4% |
| A2. The picture could enhance the emotion of the text. (*emotionally relevant in*). | 66.6% |
| A3. A beautiful picture that is not very relevant to the text, but makes the text more visually attractive (*irrelevant*). | 3.7% |
| A4. Other | 0.9% |

recruited 109 microblog users (62 females and 47 males) from a popular Chinese crowdsourcing site *Zhubajie* as well as students in our university. Respondents were asked to fill out a questionnaire on their image tweet posting behaviors, *e.g.*, *what is the primary photo-taking device* (as we discussed in Chapter 3.3). The results to our most pertinent question—*Why do you embed an image in a tweet?*—are listed in Table 4.1: 66.6% of respondents post images primarily for enhancing their text's emotion, while a much smaller 29.4% did so to provide a visually corresponding artifact as mentioned in the text.

Our survey validates the hypothesis that *emotional correlation is also prominent image-text relation in microblog posts.* motivates us to design a new model that incorporates multiple image-text relation types in next chapter.

## 4.4 Visual/Non-Visual Classification

However, we notice that the distinction between emotional and non-emotional is difficult—our annotation efforts showed that it may be more a continuum than a binary distinction. As such we only consider the binary distinction between visual and non-visual categories here, and leave the exploration of

emotional relevance to next chapter.

The distinction between visual and non-visual has practical value. A text-based image search can utilize embedded images from visual tweets, but not non-visual tweets. For example, the image in the first row of Figure 4.1 would be a suitable image result for the query "sago cream", but a search for "mosquitoes" should not retrieve the image in the third row. The classification may also help automated tagging methods filter out image-text pairs where the relevance assumption does not hold (*i.e.*, non-visual tweets). Finally, as images in visual tweets hold semantic value, social media platforms may choose to prioritize images from visual tweets in loading or in assigning screen real estate for display.

We now turn to the task of making the visual/non-visual distinction automatically via supervised classification. We first construct an annotated dataset via crowdsourcing, then describe the three classes of evidence we employ for machine learning.

### 4.4.1 Dataset Construction

To obtain gold standard annotations, we employed subjects from Zhubajie, a Chinese crowdsourcing website, as well as students at our university, to label a random subset of the image tweets from our large Weibo-2012 dataset (detailed in Chapter 3.8). Subjects were native Chinese speakers and microblog users. We asked subjects to categorize the image-text relation as either visual or non-visual. Each image tweet was annotated by three different subjects, with the simple majority fixing the gold standard. In total, we collected annotations for 4,811 image tweets (hereafter, *Weibo-Rel*) annotated by 72 different subjects. These broke down into 3,206 (66.6%) visual and 1,605 (33.4%) non-visual image tweets. Inter-annotator agreement via Fleiss' $\kappa$ shows substantial ($\kappa = 0.62$) agreement

for the image-text relation classification. To enable future work, we further asked subjects to distinguish emotional from other non-visual tweets, resulting in 519 of the non-visual image tweets tagged as emotional. In other words, 22.6% of images on Weibo are completely irrelevant. However, inter-annotator agreement was not as strong ($\kappa = 0.54$), and annotators found the distinction difficult without contextual evidence. For this reason, we do not build a classifier to distinguish these tweets, but we use them for our later dataset construction in Chapter 5.5.1.

Although our main effort is to collect a sizable annotated image tweets for Weibo, we still want to learn the image-text relation distribution on Twitter. In particular, we are interested in knowing how many image tweets are completely noisy—image and text are neither visually nor emotionally relevant. To this end, we randomly sampled 100 image tweets from our Twitter-2012 dataset (ref Chapter 3.8), and employed 23 workers from Amazon Mechanical Turk[1] to label the relevance of image–text to be either relevant or irrelevant. Each image tweet was annotated by 5 unique workers and the conflicts were resolved by majority voting. The results show 23.0% of Twitter image tweets are noisy, which is very similar to the distribution obtained on Weibo (22.6%).

### 4.4.2 Features

To utilize supervised machine learning, we employ multimedia features that leverage the text, image and social context of an image tweet. Some features are inspired by our earlier comparative analysis in Chapter 3, hypothesizing non-visual tweets tend to have a text tweet nature.

**Text Features.** We preprocess the Chinese text by passing each tweet through a word segmenter, Part of Speech (POS) tagger, and a named entity recognizer (NER). We observed that vocabulary is a good indicator

[1] `http://www.mturk.com/`

of image-text relation: *e.g.*, tweets that mention a physical object and its color exhibit a visual bias. To make the resultant *word* feature more meaningful, we discard stop words and rare words unlikely to re-occur ($freq < 5$). The word features are binary; encoding just the presence (absence) of a word.

We incorporate the learned *topic* from LDA (trained on $\sim$ 1M randomly sampled posts from Weibo-2012 corpus) as another feature. As described earlier in Chapter 3.7, certain textual topics are skewed to be image tweets (*e.g.*, fashion) and we thus hypothesize such topics are more likely to appear in visually relevant image tweets. We also encode *POS density* features (proportion of nouns, verbs, adjectives, adverbs and pronoun within a tweet) based on the intuition that the amount of certain word classes is a good visual relation indicator (*e.g.*, physical objects are usually nouns). We then encode the presence of different classes of *named entities* as another feature. Drilling down more specific *name entities* and visual images are often correlated: celebrity and his photo (like the second row in Figure 4.1), company and its logo or product image, location and the scenery. Hence, the presence of person's name, location name and organization name comprise another three features. Four *microblog-specific* features, the use of @mention, #hashtag, geolocation and URL are the last textual features that we considered.

**Image Features.** As images in the image tweets display a broad spectrum of types, we eschew object detection common in multimedia (TREC-MM) research. We employ *face* detection as an exception, recording the number of faces present, as instances of faces are often the poster herself, friends or family. For the same reason, we also included a composite co-occurrence feature that is activated only when a person's name and face is present. In our dataset, faces were detected in 22.2% of images.

Images with similar content tend to exhibit the same image-text rela-

tion. To capture this, we cluster the images by visual similarity by following Hörster *et al.*' work [54]: we first extract SIFT descriptors [80] from the images as inputs, clustering them to form visual words by building a hierarchical visual vocabulary tree [99]; and then apply LDA to the corpus of images-as-documents' visual vocabulary, aiming to learn $k$ hidden topics[2]. Subsequently, the *image topic* assignment is encoded as a single feature.

**Context Features.** From our earlier analysis in Chapter 3.5, we know that the posting time affects the probability of a tweet containing image or not, *e.g.*, a higher likelihood of being an image tweet during the day. We hypothesize non-visual image tweets are more likely to share similar nature as text tweets, and thus include the hour of the *posting time* as a feature. As people share what they have just seen (visual tweets), we capture whether the *device* used to post the image tweets is mobile or not (*e.g.*, desktops).

Social features round out our set. As discussed eariler (in Chapter 3.6), image tweets are generally more likely to be commented and retweeeted than text tweets . Thus we use the number of *comments* and *retweets* normalized by the number of followers to the author's account as features. We also note that in visual tweets, the *author-replies* to the post herself (usually as a follow up to her reader's comments), so we encode that as another feature. Finally, we use the *follower ratio* (i.e., $\frac{\#followers}{\#followed}$) to differentiate ordinary users from celebrity and organizational accounts.

## 4.5 Experiment

We performed 10-fold cross validation experiments with the Naïve Bayes implementation in Weka 3 [47]. The three sets of features were linearly concatenated into a single vector. Due to the imbalanced distribution (66.6%

[2] $k$ is tuned on a held-out set; $k = 35$ in our case.

**Table 4.2: Feature ablation experimental results with the Naïve Bayes classifier.**

| Class | Features | Macro-$F_1$ (%) |
|---|---|---|
| Text | (1): Words Only (Baseline) | 64.8 |
| | (2): (1) + Microblog-specific | 65.2 |
| | (3): (1) + Named Entities | 65.3 |
| | (4): (1) + Topic | 66.6 |
| | (5): (1) + POS Density | 69.7 |
| Image | (6): (1) + Topic | 65.4 |
| | (7): (1) + Face | 65.7 |
| Context | (8): (1) + Retweets | 60.9 (–) |
| | (9): (1) + Comments | 64.5 (–) |
| | (10): (1) + Replied by Author | 64.7 (–) |
| | (11): (1) + Device | 64.9 |
| | (12): (1) + Follower Ratio | 64.9 |
| | (13): (1) + Posting Time | 65.0 |
| All | (14): (1–7 + 11–13) | 70.5 |

of image tweets are visual), simple accuracy is not an appropriate evaluation metric. Therefore, we report the macro-averaged $F_1$ score, as we feel both classes are equally valuable. The majority baseline (all visual) obtains a macro-$F_1$ score of 40.0 (%).

To understand the impact of each feature class, we start with the best single feature (*words*, $F_1 = 64.8$) and measure the gain (loss) in $F_1$ when adding each feature in turn. The results are shown in Table 4.2. *POS density* turns out to be the second-most useful feature, increasing $F_1$ by 4.9. As a snapshot of content (*e.g.*, noun) and function (*e.g.*, pronoun) words distribution, this feature is effective in identifying non-visual tweets with heavy function word usage (*e.g.*, pure exclamations). Other textual features—*topic*, *named entities* and *microblog-specific*—also lead to small performance increment. The addition of our two image features also make

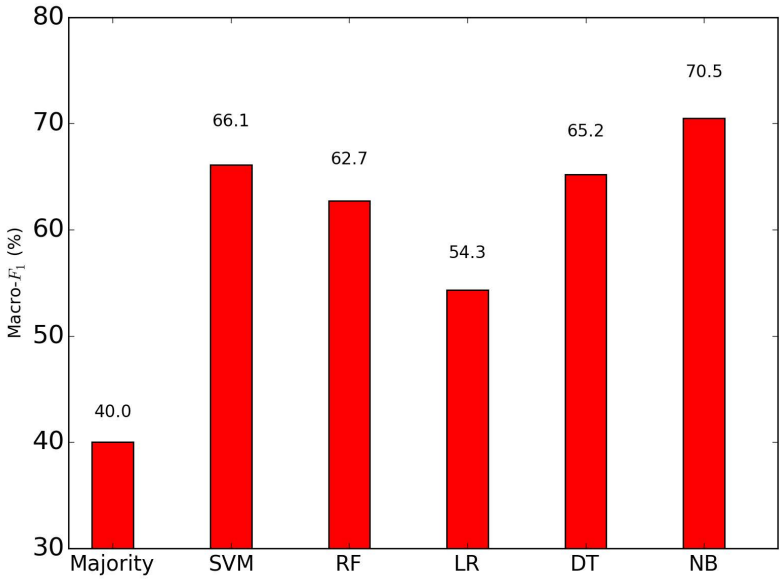



**Figure 4.3: The macro-averaged $F_1$ score of Majority Baseline, Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), Decision Tree (DT), and Naïve Bayes (NB).**

a marginal improvement over the baseline. However, not all the proposed context features are useful. The addition of *posting time*, *device*, and *follower ratio* improve the *word* baseline slightly, while the other three interaction related features (Row 8–10) do not. This may suggest that the interaction behaviors are more influenced by users' social relationships rather than the image-text relation exhibited in image tweets. Our final classifier (Row 14) that combines all features that improved the baseline, achieves an $F_1$ of 70.5.

With the selected features, we then experimented with a few other standard classifiers, including Support Vector Machines (SVM), Random Forests, Logistic Regression, and Decision Trees. All learners used the same 10 folds of dataset. From Figure 4.3, we see Naïve Bayes is the best performing classifier, bettering the second best (*i.e.* SVM) by 4.4 percentage points and the majority baseline by 30.5 points. Thus, we discuss performance and do further experimentation, limiting ourselves to the Naïve Bayes classifier.

**Error Analysis.** We further analyzed the misclassified instances. While *words* are the most discriminative feature, microblog text is relatively short. The brevity of the text sparsifies the *word* feature, giving little information to the classifier. In an extreme case, *e.g.*,"吴氏 宗祠" (ancestral hall of the Wu family), where all the words are rare or out-of-vocabulary, word features are not helpful at all. This partly explains why the *w*ords only baseline plateaus at a $F_1$ of 64.8. The informal language used in microblogs—*i.e.*, neologisms and misspellings—also poses a great challenge to standard natural language processing tools [134, 135]. We have observed many instances where misspellings are processed incorrectly by our word segmentation and named entity recognition tools. One such example is a mispelling of "阿狸" (a cartoon character) as "啊狸" in a visual tweet. The NER tool did not successfully tag this as a named entity. The propagation of this error downstream in our pipeline caused the eventual error.

Besides text features, the inaccuracy of face detection is another source of classification errors. We posit that this is due to the characteristics of images posted with visual tweets (*e.g.*, low photo quality, photo collages). We also observe an inadequacy with our context features. We sampled the feeds and image tweets of some users and realise that users have different tweet posting behaviors. Some users are more inclined to post non-visual than visual tweets, and the inverse is true of others. This is not captured in our proposed context features and we believe that features which consider the behavioral characteristics of users will be very helpful.

## 4.6 Conclusion

In this study, we identified two key image-text relations (*i.e.*, visual and emotioanl relevance) for image tweets. We then made an important distinction about image tweets—the *visually relevant* image tweet—where the

focal point of the tweet is present in both the image and text, complementing each other. In contrast, non-visual tweets use the image as a way of adorning the text in a non-essential manner—*i.e.*, to heighten interest in reading a post. We build an automated classifier leveraging features from text, image, and context evidence sources to achieve a macro $F_1$ of 70.5, an absolute improvement of 5.7% over a text only baseline. To encourage more investigation on these topics, we have made *Weibo-Rel* (the annotated corpus) available to the public[3] to test and benchmark against.

[3] `http://wing.comp.nus.edu.sg/downloads/imagetweets`

# Chapter 5

# Modeling Image-Text Relations

## 5.1 Introduction

From the previous chapter, we learned that microblog images and text correlate for either visual or non-visual (e.g., emotional) purposes—see Figure 5.1 for examples of both—given that the image tweet is not noisy. Given this, can we model such relationships and thus explain the generative process of image tweets? To investigate this question, we start by drawing on the methodology of previous works in multimedia that examined general image–text relations. These methods assume that the image and text are correlated by virtue of a single channel (*i.e.*, a visual channel). Applying a representative method to an image tweet dataset, we find that there is still a large mismatch between an image tweet's image and text.

This chapter's key contribution is to address this modeling gap by introducing Visual–Emotional LDA (VELDA), a novel topic model that captures image–text correlations through multiple evidence sources (namely, visual and emotional, yielding the method's namesake). On experiments with both English (Twitter) and Chinese (Weibo) image tweets and other

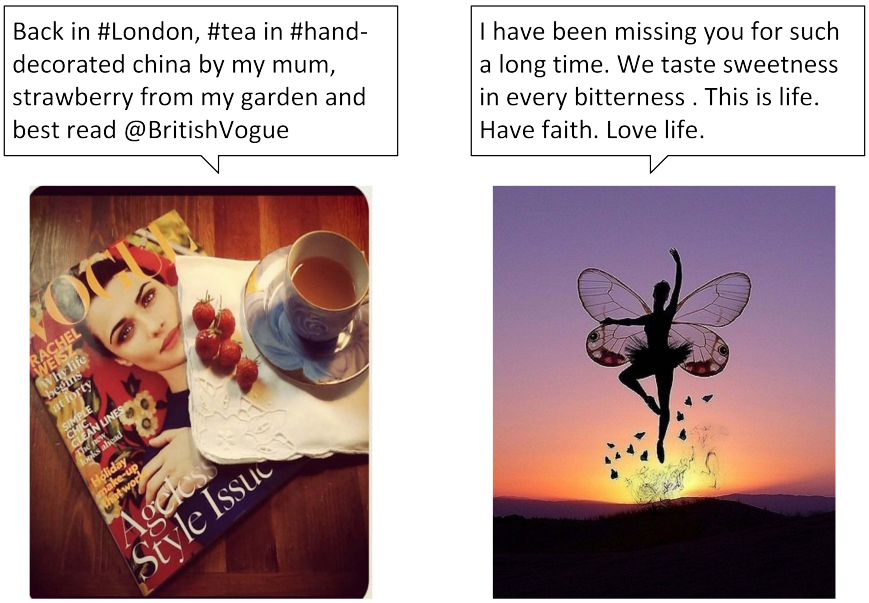



**Figure 5.1: A visually relevant image tweet from Twitter (left) and an emotionally relevant image tweet from Weibo (right).**

forms of user generated content, VELDA yields significantly improved modeling over the other existing methods on cross-modality image retrieval. Even in other domains where emotion does not factor in image choice directly, VELDA demonstrates good generalization ability, modeling these multimedia documents much better than other published methods. Finally, we apply VELDA in a real-world task of automated microblog illustration, using our model to select a relevant image (either visually-relevant, emotionally-relevant or both) drawn from an image collection.

## 5.2 Related Work

The duality of image and text has been a recurring topic of study in the multimedia area. Uncovering the relationship between the two mediums and properly modeling them has been a key area of study. One method for performing this is to map the multimodal data into a common (shared) space such that the distance between two similar objects is minimized. Under this approach, Canonical Correlation Analysis (CCA) [55] and its

extensions are often utilized [111, 122, 34]. CCA finds a pair of linear transformations to maximize the correlations between two variables (*i.e.*, image and text features), jointly reducing the dimensionality of the two spaces that provide heterogeneous representations of the same data. For instance, [111, 34] applied CCA on documents with image and text, where the image is represented as a bag of visual words and the text as a probabilistic topic distribution. They then demonstrated effective cross-modal retrieval, where the query is first transformed to the shared space learned by CCA, and then its nearest neighbors in another modality are returned as the retrieval results.

An alternative method employs probabilistic latent topic modeling to learn the joint distribution of the multi-modal data. These approaches are based on extensions of Latent Dirichlet Allocation (LDA) [13], a generative model that discovers underlying topics that generate the documents and the topic distribution within each document (Figure 5.2). As the original LDA method only applies to single modality, scholars have extended the model to handle multi-modal data. The seminal work of Barnard *et al.* [7] proposed multi-modal LDA (MMLDA) that aimed to capture the association of two modalities at the topic level, assuming the two are generated from the same topic distribution. Later, Blei *et al.* [12] proposed correspondence LDA (hereafter, Corr-LDA; Figure 5.3) to model text and image differently, where the image is assumed as the primary medium and generated first via standard LDA; then, conditioned on image's topics, the text is generated. In this sense, Corr-LDA assumes the topics of the two modalities have a one-to-one correspondence. To relax such constraint, Putthividhy *et al.* [106] proposed a topic-regression multi-modal LDA (tr-mmLDA) to learn a regression from the topics in one modality to those in the other. In real-world scenarios, much of free text may only be loosely associated to an accompanying image (*e.g.*, a Wikipedia article and its cor-

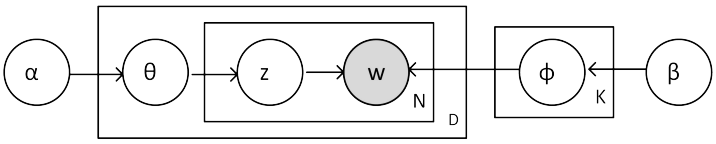



**Figure 5.2: Latent Dirichlet Allocation (LDA). We follow the formalism of Blei *et al.* [13], where plates represent replicates, shaded nodes observations, and unshaded nodes hidden variables or hyperparameters.**

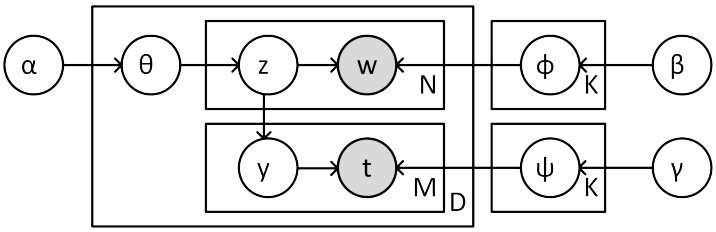



**Figure 5.3: Correspondence LDA (Corr-LDA), where the $N$ plate specifies visual words and the $M$ plate specifies individual words in the text. Note that the variables $Y$ (the topic assignments of textual words) are conditioned on $Z$ (the topic assignments of $N$ visual words).**

responding image), and in some datasets, some documents may lack images or text. To address these shortcomings, Jia *et al.* [62] proposed Multi-modal Document Random Field (MDRF) that connects the documents based on intra- or inter-modality topic similarity. The resultant learned topics are shared across connected documents, encoding the correlations between different modalities. In a separate line of work, multiple modal LDAs have been generalized to non-parametric models [142, 132, 77], which alleviates the need to choose the number of topics *a priori*.

More recently, a few works adapted the existing multimedia topic models to social media domain, in order to detect social events [9, 11, 139, 140, 19, 109] and monitor market competitions [156]. In Corr-LDA, the generation of text is depended on image, Wang *et al.* [139, 140] extended it to allow the image-text dependence to be bilateral; that is, image is also depended on text. Considering the brevity of text, Bian *et al.* [9, 11] restricted an image tweet to be generated from a single topic, instead of a distribution

of multiple topics. With the same basis—one tweet has only one topic, Cai *et al.* [19] additionally incorporated three Twitter specific factors (namely, timestamps, locations and hashtags) into modeling. Observing text is not always visually related to the image, Qian *et al.* [109] added a non-visual topic space for text, and adopted a binary variable to decide whether a textual word is generated from the visual topic distribution (shared by a pair of text and image) or the non-visual topic distribution (exclusive to text). Unlike the aforementioned six papers aiming at detecting events in social media, Zhang *et al.* [156] attempted to discover the latent topics that are competitively shared by multiple brands. As a result, their topic model not only captures the association of image and text, assuming the two modalities are generated from the same document topic distribution, but also learns the brand topic distributions.

Although the prior work is comprehensive, we have found that image tweets can exhibit and be explained from multiple perspectives. Current models assume that the relationship between an image and text can only be attributed to a single (visual) model. Our proposed method extends LDA to cater for this key characteristic in the generative process of image tweets.

## 5.3 Preliminaries

From the related work, we see that existing models assume image and text is correlated from a single (visual) perspective. Do these formalisms model the image-text relations in image tweets well? We discover in this section that the answer in short is "no". To show this, we conduct an initial experiment on image tweets from Weibo, using Correspondence LDA (Corr-LDA).

A good model of image–text relations should be able to help gener-

ate one given the other. In specific, we set the task as cross-modal retrieval: given the text of an image tweet, attempt to retrieve its accompanying image from an image dataset. To test basic image–text model, we first collected a sizable 22+ K (Chinese) image tweet corpus from Sina's Weibo platform (described in more detail in our formal experiments in Section 5.5.1).

For the textual content, we processed the tweets to build a bag-of-words (BoW) model of the text, which required word segmentation and part-of-speech tagging. We further discarded stop words, rare words (with frequency $< 10$) and closed-class words, leaving only open-class words—nouns, verbs, adjectives and adverbs—for the BoW model. This helps to reduce the noise—removing potential words that are irrelevant to the image. We then discarded image tweets with fewer than 5 remaining words.

For the images, we follow standard techniques to compute a similar BoW model of visual words: first, extracting SIFT descriptors per image, 2) running $k$-means to generate $1,000$ clusters from a random sample of 1M SIFT descriptors, and 3) converting each SIFT descriptor into one of $1,000$ visual words. We randomly split the resultant image tweet corpus into a 20K training set and a 2.4K test.

The graphical representation of Corr-LDA is depicted in Figure 5.3. Assuming image is the primary medium, Corr-LDA first generates the image and then generates the text. In particular, the image part is modeled using standard LDA. To generate the text, a topic label $y$ is uniformly sampled from $Z$, the topic assignments of all the visual words, (alternatively, we can say $y$ is sampled from the empirical distribution of the image topics), and then the textual word $t$ is sampled from the textual topic-word distribution $\psi$, a multinomial distribution with Dirichlet prior $\gamma$.

As Corr-LDA has shown promising results in modeling images and visually related words [12] in other domains, we applied Corr-LDA to our

dataset. Under the basic Corr-LDA model, given a text query $T$ containing $N$ terms, $t_1, ..., t_N$, the score for the image $i$ is defined as:

$$score_i = P(T|\theta_i) = \prod_{n=1}^{N} P(t_n|\theta_i) = \prod_{n=1}^{N} \sum_{k=1}^{K} \psi_{k,t_n} \theta_{i,k} \tag{5.1}$$

where $\theta_i$ is the visual topic distribution of the image $i$, and $\psi_k$ is the topic word distribution of the texts. As there is only one ground-truth match for each textual query (*i.e.*, the original image accompanying the post), we use on the position of the ground-truth image in the ranked list. Following previous work [62], an image is considered correctly retrieved if it appears in the top $t$ percent of the image test collection created from its corresponding text in the original post.

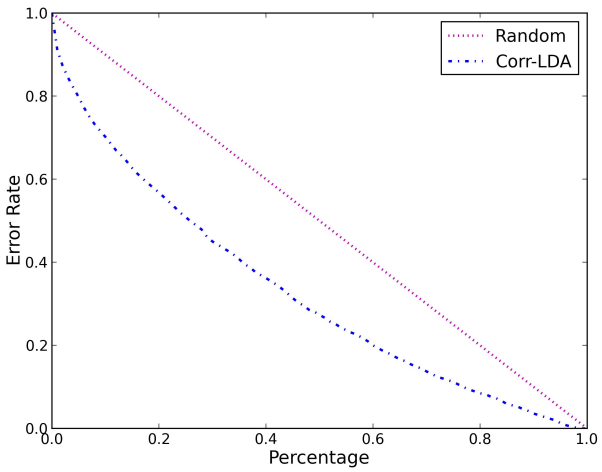



**Figure 5.4: Baseline image retrieval error rate on the Weibo set.**

Empirical tuning of Corr-LDA's parameters yields generally stable performance. For the hyperparameters, we set $\alpha$=1.0, $\beta$=0.1, and $\gamma$=0.1. We set the number of topics $K$ to 30, which minimized the average error rate over all queries. Figure 5.4 plots the results of Corr-LDA and a random baseline, where better systems have a curve that stretches farther towards the bottom left (lower error rate at lower levels of recall). The graph shows that Corr-LDA significantly improves over random selection (which plots a straight line), which agrees with our intuition. However, there is much

room for improvement. Even in the top 10% (10% × 2400 = top 240 images), error rate is close to 70%. Can this basic model be improved?

## 5.4 Visual-Emotional LDA

Thus proper modeling of the image–text relationship — in image tweets specifically — need to account for multiple views; and not just a single visual modality. Higher fidelity modeling then yields improved task performance on cross-modality image retrieval. We propose Visual-Emotional LDA (or VELDA) as a generative model that incorporates the suggested additional emotional aspect. In the following, we first detail VELDA's formulation, and then describe its parameter estimation process.

### 5.4.1 Model Formulation

Figure 5.5 shows the graphical representation of VELDA. In VELDA, each image tweet has three modalities—the textual tweet, and the visual and the emotional view of the image. Similar to other LDA-based methods, we model the three modalities as discrete features, which are referred as textual words, visual words, and emotional words, respectively (details of the feature extraction process described later in Section 5.5.2).

Following Corr-LDA, we correlate image and text in the latent topic level, such that the topic of each textual word corresponds to an image topic; the major difference is that in VELDA, we have two heterogeneous views of an image—visual and emotional. To decide which image view a textual word corresponds to, we introduce a switch variable $r$. When $r = 0$, the textual word is visually related to the image and thus sampling its topic $y$ from the empirical image-visual topic distribution $\tilde{\theta^V}$; likewise $r = 1$ indicates emotional relevance, sampling from the empirical image-emotional topic distribution $\tilde{\theta^E}$. While we could also introduce $r = 2$ to

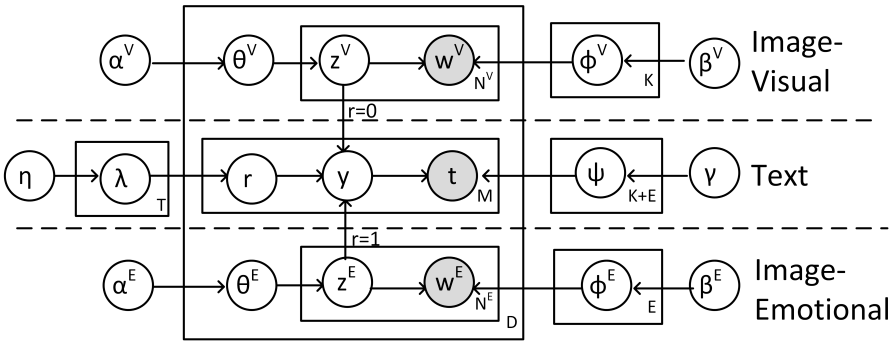



**Figure 5.5: Visual-Emotional LDA's generative model.**

capture attribution to both visual and emotional correlation, the resultant modeling complexity would be changed from linear $(K + E)$ to quadratic $(K \times E)$. To keep the model simple, we did not do so.

Intuitively, the assignment of $r$ should be term-sensitive — some textual words (*e.g.*, a physical object) are more likely to correspond to visual objects within an image, while others tend to reflect the emotion and atmosphere of an image. As such, the switch variable $r$ is personalized for each textual word and sampled from a relevance distribution $\lambda$. Here we randomly initialize the value for all $r$ variables. Alternatively, when external sentiment lexicon is available, we can use such lexicon as prior knowledge to initialize $r$ variables of sentimental words as 1 (*i.e.*,emotionally relevant). We will investigate whether the prior knowledge aids in building a better model in the future work.

The overall generative story is summarized as follows, where specific notations are explained in Table 5.1[1]:

1. For each textual word $t = 1, ..., T$, sample a relevance distribution $\lambda \sim \text{Dir}(\eta)$.

2. For each image–visual topic $k = 1, ..., K$, sample the topic word distribution $\phi^V \sim \text{Dir}(\beta^V)$. Similarly for image–emotional topic $e$ and textual topic $l$.

[1]We use the abbreviation Dir$(\cdot)$, Mult$(\cdot)$ and Unif$(\cdot)$ to denote the Dirichlet, Multinomial and Uniform distribution, respectively.

3. For each image tweet $d = 1, ..., D$, sample its image–visual topic distribution $\theta_d^V \sim \text{Dir}(\alpha^V)$ and image–emotional topic distribution $\theta_d^E \sim \text{Dir}(\alpha^E)$.

   (a) For each visual word $w_n^V$, $n = 1, ..., N_d^V$:

      i. Sample topic assignment $z_n^V \sim \text{Mult}(\theta_d^V)$

      ii. Sample visual word $w_n^V \sim \text{Mult}(\phi_{z_n^V}^V)$

   (b) For each emotional word $w_n^E$, $n = 1, ..., N_d^E$:

      i. Sample topic assignment $z_n^E \sim \text{Mult}(\theta_d^E)$

      ii. Sample emotional word $w_n^E \sim \text{Mult}(\phi_{z_n^E}^E)$

   (c) For each textual word $t_m$, $m = 1, ..., M$:

      i. Sample relevance type $r_m \sim \text{Mult}(\lambda_{t_m})$

      ii. if $r_m = 0$:

         A. Sample a topic $y_m \sim \text{Unif}(z_1^V, ..., z_{N_d^V}^V)$

         B. Sample a word $t_m \sim \text{Mult}(\psi_{k=y_m})$

      iii. if $r_m = 1$:

         A. Sample a topic $y_m \sim \text{Unif}(z_1^E, ..., z_{N_d^E}^E)$

         B. Sample a word $t_m \sim \text{Mult}(\psi_{e=y_m})$

### 5.4.2 Parameter Estimation

In VELDA, we need to infer six sets of parameters: three topic-word distribution ($\phi^V$, $\phi^E$ and $\psi$ for image-visual, image-emotional and text, respectively), two document-topic distribution ($\theta^V$ and $\theta^E$ for image-visual and image-emotional, respectively), and the relevance distribution of textual words $\lambda$. As with LDA, exact inference of the parameters is intractable; so approximate inference is applied.

In this work, we adopt Gibbs sampling to estimate the model parameters, due to its simplicity in deriving update rules and effectiveness in

**Table 5.1: Notations used in VELDA.**

| Symbol | Description |
|---|---|
| $K$, $E$ | number of image-visual and image-emotional topics, respectively. |
| $D$, $T$, $C$, $S$ | number of tweets, unique textual words, unique image-visual words, and unique image-emotional words, respectively. |
| $\alpha^V$, $\alpha^E$, $\beta^V$, $\beta^E$, $\gamma$, $\eta$ | hyperparameters of Dirichlet distributions. |
| $\theta^V$, $\theta^E$ | $D \times K$, $D \times E$ matrices indicating image-visual, image-emotional topic distribution, respectively |
| $\phi^V$, $\phi^E$ | $K \times C$, $E \times S$ matrices indicating image-visual, image-emotional topic-word distribution, respectively |
| $\psi$ | a $(K + E) \times T$ matrix indicating textual topic-word distribution. |
| $\lambda$ | a $T \times 2$ matrix indicating textual word's relevance distribution. |
| $M_{d,t}$, $N^V_{d,c}$, $N^E_{d,s}$ | number of textual words, image-visual words, and image-emotional words in the $d$-th tweet. |
| $M_{d,z}$ | number of textual words in $d$-th tweet that are assigned to topic $z$. |
| $N^V_{d,k}$, $N^E_{d,e}$ | number of image-visual, image-emotional words in $d$-th tweet that are assigned to visual topic $k$, emotional topic $e$, respectively. |
| $M_{t,r}$ | number of times that textual word $t$ is assigned to relevance $r$. |

dealing with high-dimensional data. The basic idea of Gibbs sampling is to sequentially sample all variables from the targeted distribution when conditioned on the current values of all other variables and the data. For example, to estimate the image-visual topic distribution $\theta^V$, we need to sequentially sample its latent topic variable $z^V$. To sample for $z^V_i$ (where $i = (d, n)$ representing the $n$-th word of the $d$-th document), we condition

on the current value of all the other variables:

$$
\begin{aligned}
&P(z_i^V = k | W^V, W^E, T, Z_{-i}^V, Z^E, Y, R) \\
&\quad \propto \frac{N_{k,c,-i}^V + \beta_c^V}{N_k^V + C\beta^V - 1} \cdot (\frac{N_{d,k}^V}{N_{d,k,-i}^V})^{M_{d,k}} \cdot \frac{N_{d,k,-i}^V + \alpha_k^V}{N_d^V + K\alpha^V - 1}.
\end{aligned} \tag{5.2}
$$

Similarly, we can derive the sampling rule for $z_i^E$:

$$
\begin{aligned}
&P(z_i^E = e | W^V, W^E, T, Z^V, Z_{-i}^E, Y, R) \\
&\quad \propto \frac{N_{e,s,-i}^E + \beta_s^E}{N_k^E + S\beta^E - 1} \cdot (\frac{N_{d,e}^E}{N_{d,e,-i}^E})^{M_{d,e}} \cdot \frac{N_{d,e,-i}^E + \alpha_e^E}{N_d^E + E\alpha^E - 1}.
\end{aligned} \tag{5.3}
$$

Next, we sample the latent topics $y$ of the textual words based on the topic assignment of image–visual and image–emotional. Note that for each latent topic $y_i$, there is a switch variable $r_i$ controlling whether it is sampled from image–visual topics or image–emotional topics. If $y_i$ is sampled from image–visual topics, it implies that $r_i$ is sampled to be 0, and vice versa. As such, we need to sample based on the joint distribution of $y_i$ and $r_i$, which leads to:

$$
\begin{aligned}
&P(r_i = 0, y_i = k | W^V, W^E, T, Z^V, Z^E, Y_{-i}, R_{-i}) \\
&\quad \propto \frac{M_{k,t,-i} + \gamma}{M_k + T\gamma - 1} \cdot \frac{M_{t,r=0,-i} + \eta}{M_t + 2\eta - 1} \cdot \frac{N_{d,k}^V}{N_d^V}, \\
&P(r_i = 1, y_i = e | W^V, W^E, T, Z^V, Z^E, Y_{-i}, R_{-i}) \\
&\quad \propto \frac{M_{e,t,-i} + \gamma}{M_e + T\gamma - 1} \cdot \frac{M_{t,r=1,-i} + \eta}{M_t + 2\eta - 1} \cdot \frac{N_{d,e}^E}{N_d^E}.
\end{aligned} \tag{5.4}
$$

Iterative execution of the above sampling rules until a steady state results allows us to obtain the values of the latent variables. Finally, we estimate

the six sets of parameters by the following equations:

$$\begin{aligned}
\theta^V_{k,d} &= \frac{N^V_{k,d} + \alpha^V}{N^V_d + K\alpha^V}, \quad & \theta^E_{e,d} &= \frac{N^E_{e,d} + \alpha^E}{N^E_d + E\alpha^E}, \\
\phi^V_{k,c} &= \frac{N^V_{k,c} + \beta^V}{N^V_k + C\beta^V}, \quad & \phi^E_{e,s} &= \frac{N^E_{e,s} + \beta^E}{N^E_e + S\beta^E}, \\
\psi_{z,t} &= \frac{M_{z,t} + \gamma}{M_z + T\gamma}, \quad & \lambda_{r,t} &= \frac{M_{r,t} + \eta}{M_t + 2\eta}.
\end{aligned} \tag{5.5}$$

### 5.4.3 Discussion

At first glance, VELDA looks complicated, having more parameters than LDA and Corr-LDA. Essentially, it is a well-formed extension of Corr-LDA that adds an emotional view of images and the relevance indicators for textual words. In our experiments, we observed that the larger parameter space does not adversely affect convergence – parameter estimation for VELDA is rather fast, with the Gibbs sampler usually converging within 100 iterations. Meanwhile, distributed computation strategies for Gibbs sampling [138] can also be used on VELDA, making VELDA applicable to large-scale dataset.

One may note that the structure of VELDA — its separation of both the visual and emotional views of images, and the introduction of switch variable $r$ — is generic. Both image views are simply copies of the standard LDA entwined to the text via the switching variable $r$. This means additional views of the image–text relation are easily modeled by simply introducing an additional LDA generative process, adjusting the switching variable and dimension of the textual topics accordingly. Then, the derivations of the existing image parts of the model are unchanged, incurring just additional updating rule(s) for the new factor(s), similar to Equation 5.2 and 5.3.

We posit that VELDA may be applied to other data where capturing multiple relations between channels of information is of import. If the

nature of dataset exemplifies multiple correlations, VELDA can be applied to model its generation. Take news article and its comments as an example. Comments stem from the article but also are derived from the reader's own interest (*e.g.,* represented by his tweets). VELDA may learn the correlations among the three sources: here, comments are similar to our image tweet's text modality; and the article and reader's historical interest are similar to the visual and emotional image views.

## 5.5 Evaluation

We evaluate VELDA in modeling the generation of image tweets against several baseline methods. Although VELDA was conceived to model image tweets, we claim it is also applicable to other related image–text correlation tasks. As such, we investigate how VELDA fares in modeling other general domain image–text pairs. To this end, we collect image tweets from two microblog platforms — Weibo and Twitter — and image-text pairs from Wikipedia and Google. In the following, we describe the collected datasets, our feature extraction process, the evaluation criteria, and conclude by discussing the experiments and their results.

### 5.5.1 Datasets

We collected five image–text datasets. The first two are image tweet collections from (Chinese) Weibo, and (English) Twitter. We also collected three more general image–text datasets: two datasets crawled from Google Images, and Wikipedia Picture of the Day (POTD). Table 5.2 summarizes the demographics for the five datasets.

The first four datasets have a common basis for collection—constructed by a list of queries, so we describe this basis first. Our annotated Weibo-Rel dataset (detailed in Section 4.4.1) contains a collection of 4.8K image

**Table 5.2: Demographics of the five datasets.**

| | **Weibo** | **Twitter** | **G-Zh** | **G-En** | **POTD** |
|---|---|---|---|---|---|
| No. of image–text pairs | 22,782 | 16,427 | 38,806 | 26,903 | 2,524 |
| Text language | Chinese | English | Chinese | English | English |
| Textual vocabulary size | 6,714 | 2,802 | 8,382 | 4,794 | 3,224 |
| Visual vocabulary size | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Emotional vocabulary size | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Avg. no. of textual words | 18.9 | 6.7 | 18.1 | 12.3 | 24.5 |
| Best settings for VELDA | K=100, E=60 | K=40, E=80 | K=80, E=50 | K=100, E=100 | K=40, E=80 |

**Table 5.3: Sample visual and emotional queries, translated from the original Chinese.**

| **Category** | **Example Queries** |
|---|---|
| Visual | birthday, black, visibility, pet, wedding, street, sunset, tree, rain, landscape, Mayday (*a rock band)*, Lakers, Starbucks, The Legend of Sword and Fairy (*a role-playing video game*), lover, bread, beef, cherry, butterfly, sunglasses |
| Emotional | break-up, worry, cry, embrace, give-up, good morning, excuse, love, memory, forget, mood, past, encounter, insist, complain, corner, friendship, dislike, content, constellation |

tweets randomly sampled from Weibo with human image–text relation annotations following our categorization scheme (*i.e.*, visually relevant, emotionally relevant and irrelevant). Though these labels were assigned at the tweet level, only certain words were found to be visual (emotional) indicators.

Based on this, we construct potential visual (emotional) queries by extracting the most frequent textual words from the categorized visual (emotional) image tweets, discarding stop words. In total, the query list consists of 353 words from visually relevant tweets, and 133 words from emotionally relevant tweets[2]. Table 5.3 shows 20 example queries (translated from

[2]We think the ratio of visual and emotional queries are representative to general image tweets, since image tweets in Weibo-Rel dataset were randomly sampled from a large set of public tweets.

their original Chinese) for each group. These queries are consistent with our impression that visual words are largely physical objects, while emotional words are abstract or sentimental.

1. **Weibo**. To collect the Weibo collection, we send each query as a hashtag in Weibo's text-based search interface to obtain up to 1,000 most recent image tweets. For the final dataset, we discarded queries with less than 40 results, randomly sampled 100 image tweets for those with more than 100 results, and further filtered out those failed in following feature extraction procedures, which results in a set of 22,782 image tweets. This dataset was used in our pilot study described earlier.

2. **Twitter**. To test whether image tweets on different platforms and languages exhibit different behaviors, we also examined English image tweets in Twitter. Following a similar pipeline, we constructed a set of 16,427 image tweets from Twitter using the same base queries.As the queries were originally in Chinese, we translated them into English using Google Translate. As the queries are mostly single words, this process generated acceptable translations, according to our spot checks.

The PCOS Factor: Bread: why it can be worse than pure sugar Most people are surprised when they first hear that a white bread has more of a sugar hit than sugar itself! Fluffy soft white bread made with finely...

**Figure 5.6: An example image and its snippet from Google.**

3 & 4. **Google-Zh** and **Google-En**. Image tweets vary greatly in quality for both text and images. We also want to assess VELDA performance on "prominent" images returned from an image search en-

gine. We sent the Chinese and English (translated) text queries to Google Image Search. After filtering out those with non-Chinese (non-English) snippets, we obtained 38,806 and 26,903 image-text pairs for Chinese and English, respectively. Since these are from the general web and are curated by the search engine, we expect these image–text pairs to be somewhat higher in quality than the image tweet counterparts. Figure 5.6 shows an example image and its snippet.

The Lifeboat is Taken through the Dunes, painted by Michael Ancher in oils on canvas in 1883. It is representative of themes the painter often covered, fishermen and other scenes from the Danish port of Skagen.

**Figure 5.7: An example image and its description from Wikipedia's Picture of the Day.**

5. **Wikipedia Picture of the Day**. At the far end of the spectrum for image–text pairing quality is Wikipedia's "Picture of the Day" (POTD) collection, a set of of daily featured pictures accompanied by a short description from Wikipedia[3]. Figure 5.7 shows an example image and its description. Unlike the other four datasets, POTD concentrates on high-quality, manually-curated academic topics, including nature, arts, astronomy, architecture, celebrities, history and science. In this dataset, the most frequent 20 non-stop words are: species, state, united, american, world, australia, plant, common, war, family, native, south, year, water, north, small, bird, female, long, large. We collected the daily pictures and their corresponding descriptions from Nov 1, 2004 to Jun 11, 2014. After removing the unavailable images and animated images and those did not pass the feature extraction procedure, we obtained a total of 2,524 image-text

[3] `http://en.wikipedia.org/wiki/Wikipedia:Picture_of_the_day`

pairs.

We also adopt POTD as it has been used as a dataset in prior work: Jia *et al.* [62]. They used a smaller subset of our collection, where theirs consists of 1,987 image-text pairs from Nov 1, 2004 to Oct 30, 2010.

### 5.5.2 Feature Extraction

We extract textual words from image's textual description, and another two sets of features from image to represent its visual semantic and emotional semantic, respectively. Since VELDA requires all features to be discrete, we represent all three sets of features as bags-of-words.

**Text Features**. For Chinese text, we first pass the text through a Chinese word segmentation program. Then both Chinese and English text are assigned Part-of-Speech (POS) tags. English words are additionally stemmed. We apply a frequency filter to omit words that occur in fewer than 10 (5 in the case of POTD, due to its small size) documents, drop stop words and further discard closed-class words, leaving only open-class words—nouns, verbs, adjectives and adverbs. This helps to reduce the noise by removing words that are potentially irrelevant to the image. We then discard short documents with less than four words. Applying this process resulted in 6714, 2802, 8382, 4794, and 3224 unique words for Weibo, Twitter, Google-Zh, Google-En and POTD datasets, respectively.

**Image's Visual Features**. We adopt the SIFT descriptors to represent the visual semantics of an image. As described in Section 5.3, we adopt the tradition of quantizing SIFT descriptors to yield discrete words by means of a visual codebook learned by $k$-means. To better capture the image characteristics in each dataset, we trained two separate visual codebooks: one for the POTD dataset and another for the four dataset based

on the common basis. Each codebook thus consists of 1000 visual words. Similar to the filtering on text words, we discard short documents (images) that have fewer than 10 visual words.

**Image's Emotional Features**. The feature representation of image emotions has been investigated in many works. Color-based features have proved to be simple yet effective [129, 83, 148]. We adopt 22 color-based features from the state-of-art work [83], summarized in Table 5.4. To turn an image into a bag of emotional words (BoEW), we first segmented each image into patches by a graph-based algorithm [36], and then extracted the 22 color-based features for each patch. Similar to the procedure of constructing visual words, one million emotional patches were randomly sampled to learn 1000 clusters via $k$-means. Finally, each patch is quantized into one of the 1,000 emotional words. As with the visual words, we trained two separate emotional codebooks for images from POTD and images from the other datasets.

The feature representation of image emotions (*a.k.a.*, sentiment or affect) has been investigated in many works. Color-based features have proved to be simple yet effective [129, 32, 83, 61, 148]. We adopt 22 color-based features from the state-of-art work [83], summarized in Table 5.4. To turn an image into a bag of emotional words, we first segment each image into patches by a graph-based algorithm [36], and then extract the 22 features for each patch. Similar to the procedure of constructing visual words, one million emotional patches were randomly sampled to learn 1000 clusters via $k$-means. Finally, each patch is quantized into one of the 1000 emotional words. As with the visual words, we trained two separate emotional codebooks for images from POTD and images from the other datasets.

**Table 5.4: Features used to represent image emotions.**

| Name | Dimension | Description |
|---|---|---|
| Saturation | 2 | Mean and standard deviation of saturation. |
| Brightness | 2 | Mean and standard deviation of brightness. |
| Hue | 4 | Mean hue and angular dispersion, with and without saturation weighted. |
| Color Names | 11 | Amount of black, blue, brown, green, gray, orange, pink, purple, red, white and yellow [130]. |
| Pleasure, Arousal, Dominance | 3 | One set of affective coordinates, calculated from brightness (B) and saturation (S), as follows [129]:<br>Pleasure = 0.69*B + 0.22*S<br>Arousal = -0.31*B + 0.60*S<br>Dominance = -0.76*B + 0.32*S |

### 5.5.3 Experimental Settings

We adopt the cross-modal image retrieval task and the evaluation metric (error rate retrieved by first $t$ percent) as in Section 5.3. Specifically, given a textual query T=$t_1, ..., t_N$, VELDA computes a score for an image $i$ by the following formula:

$$\begin{aligned} score_i =& P(T|\theta_i^V, \theta_i^E) = \prod_{n=1}^{N} P(t_n|\theta_i^V, \theta_i^E) \\ =& \prod_{n=1}^{N} (\lambda_{t_n,0} \sum_{k=1}^{K} \psi_{k,t_n} \theta_{i,k}^V + \lambda_{t_n,1} \sum_{e=1}^{E} \psi_{e,t_n} \theta_{i,e}^E) \end{aligned} \tag{5.6}$$

where $\theta_i^V$ and $\theta_i^E$ are the visual and emotional topic distribution for a test image $i$, and $\lambda$ is the textual word relevance distribution learned during training. Larger scores suggest higher relevance. Note that the marginal probabilities $P(t_n|\theta_i)$ can be pre-computed for each image off-line during learning, such that the score computation is fast, making the real-time retrieval feasible.

Aside from **Corr-LDA**, we further compare VELDA with another state-

of-the-art image–text correlation algorithm, **LDA-CCA** [111]. In this method, two standard LDA models are first trained for texts and visual images individually; *i.e.*, an image-text pair is represented as two independent topic distributions. Then Canonical Correlation Analysis (CCA) projects the two topic distributions to a shared latent space where the correlation between image-text pairs is maximized. For each textual query, the images are ranked to minimize the distance with the query in the shared space.

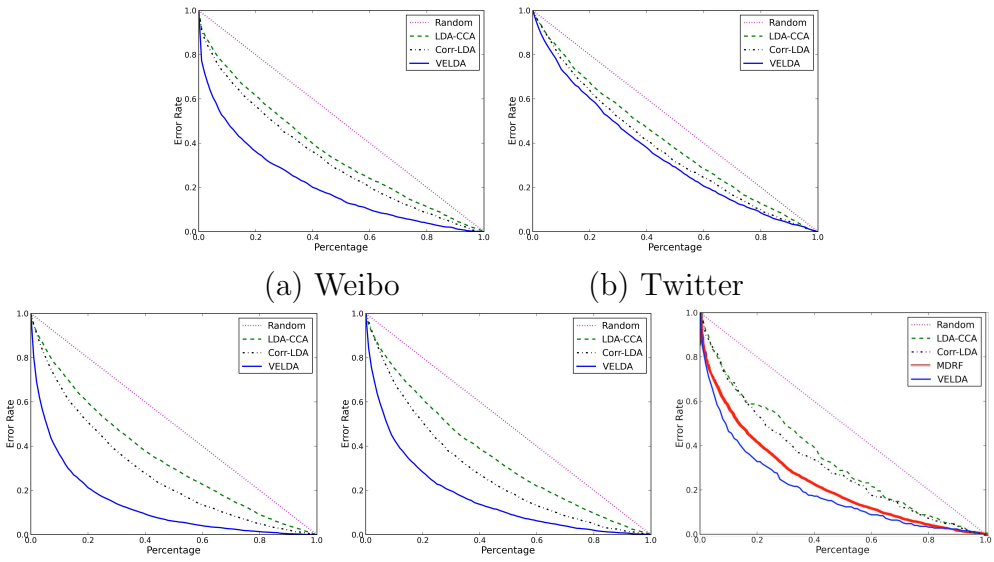
We randomly split each dataset into 90% as training set and the remaining 10% as testing set. Our development testing showed that VELDA operated well over a wide range of hyperparameter settings. As such, we fix the six sets of hyperparameters to relatively standard settings: $\alpha^V$=1, $\alpha^E$=1, $\beta^V$=0.1, $\beta^E$=0.1, $\gamma = 0.1$, and $\eta = 0.5$. We then tune the number of visual topics ($K$) and emotional topics ($E$) in a grid search for each dataset (see Table 5.2 for the detailed settings). We similarly optimize the Corr-LDA and LDA-CCA baselines by searching for their best parameter settings.

### 5.5.4 Results and Analysis

Results on the cross-modal image retrieval tasks on the five datasets are shown in Figure 5.8. Each plot depicts the retrieval errors averaged over all testing queries in a specific dataset for all applicable methods. For all the five datasets, a two-tailed paired $t$-test with threshold 0.001 revealed that the difference between our results and the other methods' is significant.

For POTD, we have additionally overlaid Jia *et al.*'s Multi-modal Document Random Field (MDRF) model results, as given by their paper. This MDRF results are not strictly comparable[4], but we feel are indicative of

[4]We are not able to access Jia *et al.*'s original dataset and the extracted features (`http://www.eecs.berkeley.edu/~jiayq/wikipediapotd`). We use the same ratio of




(a) Weibo (b) Twitter

(c) Google-Zh (d) Google-En (e) Wikipedia POTD

**Figure 5.8: Retrieval error rate by the percentage of the ranked list considered. Curves closer to the axes represent better performance.**

MDRF's performance, and further help to show VELDA's competitive performance.

For all graphs, better performance is equated with lower error rate earlier in the ranked list (curves closer to the bottom left corner). From Figure 5.8, we see that the error rate of VELDA drops dramatically when increasing the retrieval results to first 10%. In particular, more than 20% of ground truth images appear at the very early of the ranked list (*e.g.*, the top 0.8% for Weibo). For concrete comparison, we focus on recall at the top 10% level, reported separately in Table 5.5. Compared to Corr-LDA (the strongest baseline), our proposed VELDA significantly improves retrieval performance by 20.6%, 31.6%, 25.8%, 22.4% for Weibo, Google-Zh, Google-En and POTD dataset, respectively. For the POTD dataset, VELDA outperforms MDRF by around 8%, although not strictly comparable. In this dataset, though emotion is not the primary reason for choosing the images, it might be an implicit factor, *e.g.*, nature related articles prefer

train/test to be as comparable as possible, but our dataset is larger by 537 documents (approximately 1/5 larger).

**Table 5.5: Percentage of images correctly retrieved in the top 10% of the ranked list. The difference between VELDA and any of the two other methods is statistically significant with the two-tailed paired $t$-test ($p < 0.001$).**

| | **Weibo** | **Twitter** | **G-Zh** | **G-En** | **POTD** |
|---|---|---|---|---|---|
| LDA-CCA | 25.6% | 18.8% | 25.0% | 25.6% | 28.8% |
| Corr-LDA | 29.8% | 22.0% | 32.1% | 31.2% | 31.2% |
| VELDA | **50.4%** | **26.6%** | **63.7%** | **57.0%** | **53.6%** |

**Table 5.6: VELDA's performance broken down by query type.**

| | **Weibo** | **Twitter** | **G-Zh** | **G-En** |
|---|---|---|---|---|
| Visual queries | 53.8% | 28.1% | 69.8% | 61.6% |
| Emotional queries | 39.5% | 22.1% | 47.0% | 45.4% |

images that are bright and incur peaceful feeling.

Note the lower performance of all three methods on the Twitter dataset. We attribute this to the brevity of Twitter. As in Table 5.2, each Twitter image tweet has only 6.7 textual words on average (after text processing), far shorter than the other mediums. This passes little textual information to the model, and makes the image-text correlation learning difficult. Even in such sparse data scenarios, VELDA tolerates short text and noise well, bettering Corr-LDA and LDA-CCA by 4.6% and 7.8%, respectively.

We further break down VELDA's performance by query type, as shown in Table 5.6. POTD was not curated by queries and thus not discussed here. We find all the other four datasets show the same trend that VELDA performs better in image-text pairs from visual queries than those from emotional ones. As the query type is a good indicator of the image-text correlation type (visual or emotional), this trend partially implies that learning image-text's emotional correlation is more difficult than the visual correlation.

To apply our VELDA model to other domains, the major parameter

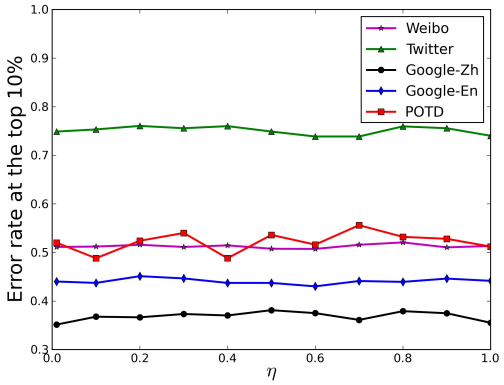



**Figure 5.9: Parameter $\eta$ versus error rate for top 10% retrieval.**

to tune is the hyperparameter $\eta$, which determines the relevance distribution ($\lambda$) of textual words, while other parameters can be easily set with standard LDA heuristic rules. So we further investigate the impact of $\eta$. Theoretically speaking, large (small) $\eta$ makes the relevance distribution $\lambda$ more skewed (balanced). From Figure 5.9, we see that VELDA's performance remains relatively stable for varying values of $\eta$. This insensitivity to $\eta$ shows that VELDA is robust and does not require careful tuning to perform well.

## 5.6 Towards Microblog Illustration

Image tweets may have arrived as a dominant form of social media, but their quality is still worth improving. Many images in such tweets are of poor quality, exhibiting little correlation with the texts in the post. From our previous finding in Chapter 4.4.1, 22.6% of images on Weibo image tweets and 23.0% on Twitter are irrelevant to the text.

When faced with irrelevant images, the readership can be annoyed—we observed respondents' complaints about the irrelevance of such tweets. To maintain and grow readership, there is a need to discover relevant images for which to "illustrate" microblog posts. In fact, microblog authors expend no small effort in sourcing for proper images: of our survey of 109

microblog users, 75 (68%) rely on web image search engines to locate images to accompany their posts, in the case where a self-taken picture is not appropriate.

To validate this use case, we randomly sampled 1000 image tweets from the Weibo and Twitter datasets, sending their images to Google Image Search's query-by-image facility. 59.4% and 38.3% of the images from Weibo and Twitter, respectively, find an exact match among indexed web images (excluding images on both the Weibo and Twitter sites), which corroborates that the discovery/recommendation task as an important problem.

In the literature, such a recommendation task—finding suitable images to illustrate text fragments—is termed *text illustration*. The text illustration task can be formally framed as ranking a set of images based on image-text correlation, given a textual query. This has been explored in a few domains, including children' books [66, 163], news articles [163, 35, 76], textbooks [3] and travelogues [17, 81].

How does VELDA fare as a microblog illustration recommendation agent? Does the error reduction in the previous experiments yield useful recommendations for actual queries? Figure 5.10 shows three example Weibo posts (translated to English) and their top four recommended images by our VELDA. We see for the very visual tweet that mentions many physical objects, *e.g.*, the top example post, our suggested illustrations not only accurately correspond to objects, but also cover a few variety (*e.g.*, capturing three different nuts in the top four illustrations). For the obvious sentimental tweets, *e.g.*, the bottom example post, our recommended images match the emotions of the text well.

We also observe that some pictures may be relevant to the text, but may not be appropriate as an illustration for other users' post. They are either too personal (*e.g.*, the portrait in the third row), too commercial

[Have some #nuts at noon] Nuts such as walnuts, peanuts, sunflower seeds, hazelnuts, cedar nuts and chestnuts, should be part of our daily diet. They are rich in Omega-3 and Omega-6 fatty acids and other essential amino acids and minerals, including carotene, calcium, and iron. These are essential for good health and have anti-aging benefits too.

#Upset I am hungry but I cannot eat now as I have to wait for someone else. What if there is a blackout now? Let me amuse myself by reading up some jokes.

The worries and problems that other people are facing always seem so minuscule and in-significant. However, when you come face to face with the same problems, you will realise that it is not possible to just laugh it off. #painful

**Figure 5.10: Three (translated) Weibo microblog posts, along with VELDA's top 4 suggested illustrations.**

(*e.g.*, the third picture of the first row contains a product's logo) or are aesthetically poor (*e.g.*, picture is blurred).

While the results are encouraging, to build a real-life recommender system, it is necessary to filter such images from the dataset. A static dataset, perhaps compiled by crawling a large set of image tweets, would cover the predominant, stable topics in found in microblog posts, such as daily routines and opinions [60]. For breaking news events and other emergent hot topics, perhaps mining other image tweets in a real time could be a successful method.

## 5.7 Conclusion

Image tweets pair text with image to help convey a unified message. We examine the image–text correlation and its modeling for the purposes of text illustration, in both microblog posts as well as other image–text datasets.

From our second study (Chapter 4), we discover that an image tweet's image and text can be related in different modes, not limited to visual relevance but which can include emotional relevance. A key contribution is our development of Visual-Emotional LDA (VELDA), a topic model variant that captures multiple image–text relations (explored here as visual and emotional modalities). Experiments with VELDA on both English and Chinese image tweets show that VELDA significantly outperforms Correspondence LDA and LDA-based Canonical Correspondence Analysis, two state-of-the-art baselines for modeling the image–text relationship.

VELDA also demonstrates its robustness and generalization, being applicable not only to its intended domain of image tweets but also general image–text datasets. On both search engine image collections as well as the curated Wikipedia "Picture of the Day" dataset, VELDA shows a even larger performance margin over suitable baselines. VELDA's performance brings the possibility of text illustration—finding a suitable image to accompany a text—a real possibility.

# Chapter 6

# Mining Contextual Text for Image Tweets

## 6.1 Introduction

To utilize image tweets for downstream applications, a fundamental question needs to be addressed: *What are the image tweets about?* In Chapter 3, we have partially answered this question from a collective view, and we now turn to each individual image tweet. Automatically interpreting the semantics of images is already difficult, this difficulty is compounded with such social media images, as they are unconstrained by genre, content, audience and quality. As an image tweet consists of two parts—*i.e.* the embedded images and the accompanying text—a natural question is whether we can rely on the text alone to interpret the meaning of the tweet? The answer is clearly no. As shown in Chapter 3.7, 5.0% of image tweets on Weibo and 2.2% on Twitter have no corresponding text at all. Even for those with text, the text and image may exhibit different semantics (*e.g.*, the irrelevant image tweets). Therefore, the image is an indispensable component for image tweet understanding: indeed, "a picture is worth a thousand words".

As we discussed earlier, low-level features do not capture image's semantics well, neither do higher level features like visual objects. It is difficult to relate images to their stories solely from visual tags—recall the poster image for China ends the one-child policy (Figure 6.1, left), and a picture of movie *Fast and Furious 6* also exemplifies this (Figure 6.1, right). In microblog, the semantics of an image are not only reflected on its pixel values, but also shaped by the context in which the picture was taken and used. In order to properly interpret microblog images, it is mandatory to go from capturing visual properties to modeling context.

To bridge this gap, we devise a **c**ontext-aware **i**mage **t**weets model**ing** (CITING) framework (illustrated in Figure 6.2) to enrich the representation of image tweets from both intrinsic and extrinsic contexts. Unlike the previous work, we do not infer the semantics from image's pixels—neither representing an image with low-level features, nor employing visual recognition to images. Instead, we leverage context to exploit text description for the images; stated differently, we attempt to capture "the thousand words" that represents the picture. We start with post's intrinsic contexts, namely, 1) the accompanying text and 2) the image, and then we turn to extrinsic contexts, 3) the external web pages directed by post's embedded URL, and 4) the Web as a whole. Considering the contextual text from each source differs in quality and coverage, we also propose a series of heuristics to fuse text when multiple channels are triggered. This fusion makes the modeling more accurate and reduces the acquisition cost of the context.

A good understanding of social images' semantics benefits many downstream applications, *e.g.,* user interests modeling, retrieval, event detection and summarization. To demonstrate the rationality of the extracted contexts, we focus on the personalized image tweet recommendation task, for which the key is to accurately model users' interests. We develop a generic feature-aware Matrix Factorization (MF) framework to model users' pref-


**Figure 6.1: Two image tweets: (left) China ends the one-child policy and (right) the movie *Fast and Furious 6*. The typical visual tags [2]are "child, cute, girl, little, indoor" and "car, asphalt, road, people, transportation system", respectively.**

erence on features. As users do not explicitly express their dislikes, there is a lack of negative data, which can adversely hurt the learning of user interests [57]. To resolve this, we propose a time-aware negative sampling strategy that samples negative tweets for a user based on how likely the user may see the tweet but has not retweeted it. Lastly, we adopt a pair-wise learning to ranking method to infer users' interests based on our enhanced contexts. We conduct experiments on a large Twitter dataset[1] (hereafter, Twitter-Rec, detailed in Section 6.5.1), showing that our proposed contexts are more effective for users' interests modeling than the textual tweets and visual images themselves, as well as validating the efficacy of our designed recommendation method.

## 6.2 Related Work

In this section, we first review existing studies on understanding image tweets' semantics. As there is no previous work on personalized recom-

[1]Restricted by Weibo API, we are unable to collect user-centric posts from Weibo. We believe our approach is also effective on Weibo since Weibo exhibits similar image characteristics (*e.g.*, not only photographs but also screenshots, synthetic images) and supports similar functions (embedded URLs, hashtags) as Twitter, although the usage of each feature varies across platforms (*e.g.*, URLs are adopted more often on Weibo than Twitter, ref Chapter 3.7).

[2]They are the actual top five tags from Clarifai (`http://www.clarifai.com`), a commercial image recognition system.

mendation of image tweets (to the best of our knowledge), we then review works about general tweet recommendation in the microblog setting.

### 6.2.1 Semantics of Image Tweets

As we discussed earlier, most existing works leverage both the accompanying text and the image to interpret the semantics of an image tweet [21, 9, 10, 11, 139, 140, 156], and one exception is done by Cappallo *et al.* [20] that used image features only. For text, surface textual words are extracted and carried out common pre-processings, *e.g.*, tokenization[3], lowercase, stopwords removal, whereas Can *et al.* [21] discarded actual textual words but encoded a binary feature to indicate the existence of a hashtag.

For image, the existing works followed the multimedia paradigm, attempting to mine the semantics of an image from low-level features, *e.g.*, pixels, color histograms, SIFT descriptors and Speeded Up Robust Features (SURF) [139, 140, 58, 9, 10, 11, 118] or higher level features, *e.g.*, visual objects [21], and the output from the upper layers of CNNs trained for object recognition (hereafter, CNN features) [19, 156, 20]. To be specific, SIFT descriptors and SURF are quantized by means of a visual codebook learned by $k$-means. For CNN features, two papers [19, 20] used them directly (*e.g.*, 4096 dimensional visual features), while Zhang *et al.* [156] carried additional steps to quantize them into discrete features, following a similar approach as the quantization of SIFT and SURF. That is, Zhang *et al.* first constructed $K$ visual clusters by applying $k$-means clustering to randomly sampled CNN features, and then assigned the $r$-nearest visual clusters to each image (*i.e.*, CNN feature), where $r$ is heuristically set to be the number of unique textual words in the accompanying text.

Due to their heterogeneous nature, features from text and image lie in different semantic spaces. To resolve this, multi-modal topic mod-

[3]The corresponding processing for Chinese text is word segmentation.

els [139, 140, 9, 10, 19, 156] were proposed to capture the image and text relations and then project an image tweet to a shared topic space, which we have discussed in Chapter 5.2. However, the representation of an image tweet (*i.e.*, the multimedia document topic distribution) is not directly interpretable for semantics, since these models are unsupervised and the learned topics are latent. Another work by Bian *et al.* [10] directly assigned human readable semantic labels to image tweets via supervised cross-media classification. To save the cost for labeling image tweets, they transferred the knowledge from portal websites (articles with editor assigned category labels) to microblog domain.

Although the most obvious contextual information—accompanying text—has been exploited, these prior works limited their investigation to the textual words of the tweets only. Microblog specific textual features (*e.g.*, hashtags, external URLs) that contain rich contextual information are completely ignored. On the other hand, images are utilized at a shallow level, *i.e.*, using ordinary visual features as those used in the general domain. As we discussed in Chapter 3.3, microblog images exhibit their unique characteristics, *e.g.*, screenshots and image with overlaid text (see Figure 6.1 right). In our work, we aim to fill this gap by leveraging the uniqueness of image tweets for semantic mining.

### 6.2.2 Tweet Recommendation

With the vast amount of tweets, microblog users are now overwhelmed with many uninteresting posts. It is of great necessity to understand users' interest and recommend interesting tweet feeds for users. One line of research [52, 102, 95, 133] attempted to predict the general interestingness (or equivalently, "popularity") of a tweet, in regardless of the identity of an audience. Such prediction task is usually formulated as a classification

problem (*e.g.*, popular or not). To this end, various features have been exploited, such as the explicit features from tweet's textual content (*e.g.*, words, topics and sentiments), contextual meta-data (*e.g.*, posting time), and the author's profile (*e.g.*, the number of followers and followees). The only work that has paid attention to image tweets is done by Can *et al.* [21], which also utilized the shallow image features to build the classifier.

However, a general popular tweet does not necessarily mean it will be interesting to a particular user, since interestingness is subjective and relevant to user's own taste [4]. To generate better recommendations for users, researchers have turned to build personalized models to predict tweet's interestingness. An early work by [128] built a classifier similar to general popularity predictor with additional features from the target user, such as user's retweeting regularity and user–author relations. Later work has formulate it as a typical recommendation problem [27, 53, 37, 155], for which collaborative filtering (CF) is known to be the most effective technique. However, for the microblog platform, CF does not work well for tweet recommendation because of the ubiquitous cold-start problem: most live tweets are newly generated and have never been seen in the training data. To tackle this, existing works incorporate tweet's textual content into collaborative filtering models. Specifically, Chen *et al.* [27] transformed the traditional user–tweet interaction matrix to user–word matrix before applying the matrix factorization method. Following the same idea, Feng *et al.* [37] additionally modelled the user–hashtag interaction, since hashtags can be seen as a good topic indicator. Another work by Hong *et al.* [53] extended the Factorization Machines (FM) to jointly model user–tweet relations and the textual tweet generation.

Despite the fact that many works have studied the tweet recommendation problem, they have primarily focused on the textual tweets. The rich signals in images and contexts have been ignored. To the best of

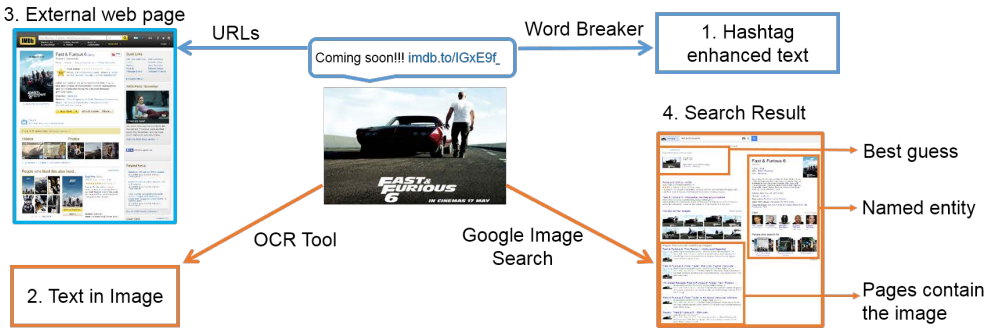



**Figure 6.2: An image tweet's four sources of contextual text. Blue outlines denotes evidence from the text; orange from the image.**

our knowledge, our work is the first to specially consider the personalized recommendation problem with image tweets.

## 6.3 Context-aware Image Tweets Modeling

In this section, we present our CITING framework for image tweets modeling. We first describe the four strategies that construct contexts from different data sources (*cf.* Figure 6.2), and then discuss the rules to fuse the contexts which help to improve text quality and save the acquisition cost.

### 6.3.1 Four Strategies to Construct Contexts

We start with the intrinsic context in image tweets: 1) the textual tweet, and 2) the image itself. Then we turn to the extrinsic context: 3) the external web pages hyperlinked in the tweet, and 4) the whole Web based on search engine.

#### 1. Hashtag Enhanced Text

The most obvious context for a microblog image is its accompanying text which forms the basis. Here we focus on hashtags, which have relatively

high coverage—they are prevalent in image tweets (26.8% have them in our Twitter-Rec dataset). Compared to the textual words of a tweet, hashtags exhibit stronger semantic ties to the post [72]: while we observed that a few hashtags (*e.g.*, #dogphoto) annotate objects present in an image, the majority describe the topic or event of the image (*e.g.*, #itsyourbirthday). In both cases, hashtags are helpful in capturing the semantics of the image. Due to their user-generated nature, hashtags do not exhibit the regularity of controlled vocabulary: people may use different variant hashtags to refer to the same (series of) events (*e.g.*, #icebucket, #ALSIceBucketChallenge; #NewYears2013, #NewYears2014). Gathering hashtags variants can thus help conflate images with common semantics. Observing these variants are often composed with common keywords, we break up hashtags into component words by Microsoft's Word Breaker API[4] [136], finding 14.3% of image tweets utilize multiword hashtags.

**2. Text in Image**

As discussed in Chapter 3.3, images in microblogs are not solely captured by camera, and many of them are software-generated or edited images, *e.g.*, graphics, memes, cartoons and screenshots. We observed that text is often embedded in images not coming directly from a camera source: our manual annotation of 500 randomly-sampled images from Twitter-Rec dataset identified 174 (34.8%) that fall in this category. We term such images as *text-images*, which we further categorize into five subtypes.

In the second column of Table 6.1, we see that one-third of text-images are meme-styled: *i.e.*, a (viral Internet) image overlaid with text (as in Figure 6.3, left). It is impossible to differentiate the semantics of meme-style images from a visual perspective, as many originate from an identical source picture. Figure 6.4 shows two example images. In contrast, the

[4] `https://www.projectoxford.ai/weblm`

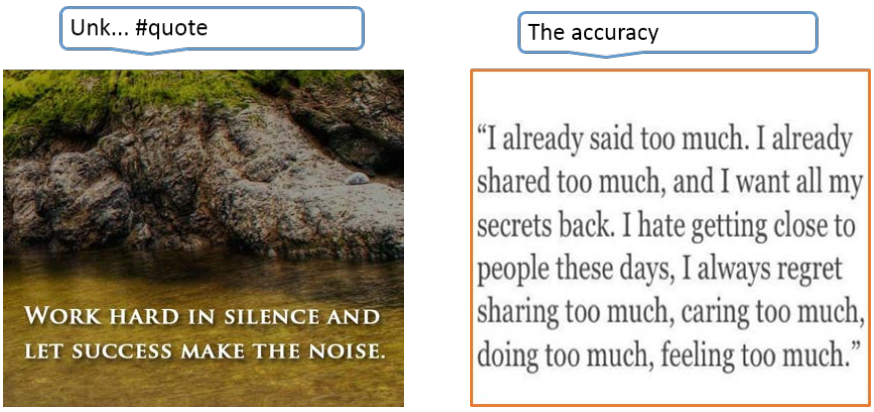



**Figure 6.3: (left) Meme-styled and (right) text-styled image tweets. The tweets' text are given in the callouts.**

embedded captions all but give away the context. In even more text-heavy cases, images can consist purely of text (Figure 6.3, right), accounts for roughly a fifth of text-images. Twitter users sometimes post such pure text-style images to circumvent the 140 character restriction. Screenshots of tweets (8.0%) are also common; we conjecture that the intention of such posts is to achieve the "retweet with comment" feature before Twitter officially supported this function in April 2015. For such tweets that have a strong textual nature, object detectors are close to useless. For the remaining text-images, 16.7% are other synthetic images, and 8.5% are natural photos that contain text in the scene (*e.g.*, road signs). Our findings lead to two key implications: 1) that a large proportion of social media images have a textual aspect, and for posts feature in such images, that 2) the embedded text is an important carrier of its semantics.

As such, we apply the *Tesseract* open source OCR software (version 3.02.02)[5] to recognize text from these images. After further using the vocabulary built by our Twitter posts to filter out noise, 26.4% of the images in our dataset have at least one recognized textual word. As Tesseract is designed for printed text, a natural question to ask is how well does it work for Twitter images? Using our manual annotation as a reference, we

[5]`https://github.com/tesseract-ocr/tesseract`

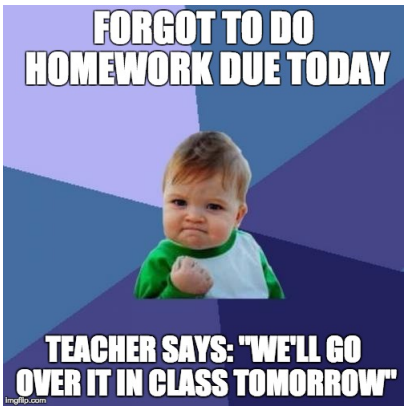



**Figure 6.4: Two meme-styled images have similar visual properties but different embedded captions.**

**Table 6.1: Demographics of the 5 subtypes of text-images and associated Tesseract OCR performance, via its miss rate (false negatives) and macro averaged recall of text words.**

| Category | Manual | Tesseract | |
|---|---|---|---|
| | # (%) | Miss Rate | Avg Recall |
| Text-style | 38 (21.8%) | 10.5% | 0.984 |
| Meme-style | 64 (36.8%) | 42.1% | 0.572 |
| Tweet screenshot | 14 (8.0%) | 7.1% | 0.843 |
| Other synthetic | 43 (24.7%) | 30.2% | 0.500 |
| Natural scene w/ text | 15 (8.6%) | 66.7% | 0.467 |
| Total | 174 | 119 | |

evaluate Tesseract's performance on our 500-image sample set (hereafter, *Twitter-OCR*). Table 6.1's rightmost two columns show its miss rate and the average recall for recognized text. Overall, Tesseract detected text from 119 images, missing 55 images that actually did contain some text. The majority of the misses come from text present in the scene (missed two-thirds) and meme-style text (missed 42.1%). Tesseract performs well on pure-text style images (detected 89.5% of images with some text, and recognized 98.4% of the actual words) and tweet screenshots. The cause of the discrepancy is simple: the more similar the image is to scanned text, the better the performance.

### 3. External Webpages

To provide context as well as to circumvent length limitations, microblog users also embed shortened hyperlinks in their tweets. In our Twitter-Rec dataset, 22.7% of image tweets contain at least one external URL. To the best of our knowledge, URLs in image tweets have not been studied in prior work. What are the external web pages about? How do they correlate to the images?

To answer these questions, we first resolved the hyperlinked shortened URLs and stored the redirected original URLs[6]. We then aggregated the resolved URLs by domain, manually categorizing the top 100 most frequent domains (accounting for 51.8% of URLs) into seven types. Table 6.2 shows the category distribution of the external resources. The majority are news reports, while three other prominent sources are online social networks (15.3%), e-commerce shops (11.9%), and articles (10.9%, *e.g.*, WordPress blogs). YouTube, image aggregators and music links account for the remaining minority (3.9%, 2.3% and 1.8%, respectively).

Interestingly, we also discover that the tweet image often originates from the external resource (82.1% of URL image tweets in our 500 sample set). Often, the image is a key scene in a news event, an item to be sold for online shops, or a portrait of the musician in music links. This suggests the external resource is the original, unsummarized context for such tweets, and thus a reliable source for capturing the image's semantics. We thus apply Boilerpipe [69] to extract the main textual content, then filter out stopwords, and finally use standard $tf.idf$ term weighting to select the top $k$ textual words as features. Considering some pages consist only of the title text (no main text), we use the page's title as another descriptor.

[6]Over half were still accessible, as of 30 September 2015.

**Table 6.2: The categories of 100 most frequent domains in external URLs and Google Image indexed pages. For the 66.0% SNS in that are indexed by Google, 48.0% are Twitter posts, and 40.1% are Pinterest posts.**

| % | News | SNS | Shop | Article | YouTube | Aggregator | Music |
|---|---|---|---|---|---|---|---|
| URLs | **51.7** | 15.3 | 11.9 | 10.9 | 3.9 | 2.3 | 1.8 |
| Google | 5.7 | **66.0** | 3.6 | 0.3 | 2.8 | 18.3 | 1.0 |

### 4. Search Engine as Context Miner

As 85% of Twitter trending topics are news [71] and Internet viral images are popular, many such tweet images have been previously used in other places on the Web, in similar contexts. To obtain these external contexts, we leverage Web search engines, which represent an update-to-date repository for the Web. In our study, we send each image in our Twitter-Rec dataset as a query to Google Image Search (We did this for our Twitter-Rec dataset during the last week of August 2015), then parse the first search engine result page (SERP) to obtain a list of pages that contain the image (including URL, title and snippet). We then follow the links to crawl the actual content of the external pages. In our dataset, a surprisingly large proportion 76.0% of Twitter images have already been indexed by Google.

What are the external webpages about? Following our workflow for tweets' embedded URLs, we also categorize the top 100 domains for such SERP-listed web pages, which accounted for 54.6% of pages. From Table 6.2, we see 66.0% of pages are social networks posts, of which 48.0% originate from Twitter itself. This implies images are re-purposed even in Twitter, and that image re-use is not limited to retweeting. The photo-based Pinterest social network takes up another 40.1% of such posts. The second largest category represents photo aggregators (18.3%, *e.g.*, `imgur.com`), which collect popular images from social networks. The remaining 15% is distributed among the other site types (news sites, e-commerce, YouTube,

music sites and blog sites, representing 5.7%, 3.6%, 2.8%, 1.0% and 0.3%, respectively).

For the query image, Google Image Search occassionally also offers a "best guess" at a short text description. Unlike the tags from traditional visual recognizer, the "best guess" goes from visual description to semantic description, which is actually the best keyword for searching the query image. For Figure 6.1 (right), the best guess is "fast and furious 6" which is spot-on. When the query image is identified as a named entity (*e.g.*, celebrity, movie or landmark), Google also sometimes shows a detailed named entity description in a knowledge graph box (functionality introduced in Google around July 2012). We additionally utilize these sources—the best guess (57.9% of Twitter images) and named entity (8.1%) as image's semantic description when available. In sum, 81.3% of images in our dataset have obtainable contextual text from Google Image Search.

### 6.3.2 Fusing the Text

Image tweets have rich contexts that can be exploited. In our dataset, 89.1% of images have at least one applicable strategy and 39.9% can leverage multiple ones. Therefore, it is important to fuse the text from multiple contexts properly. In the literature, early fusion—fusing the information at feature level—is the most widely used strategy [5].

Following this idea, we could simply fuse text from all sources, however, we find the four sources have large overlaps. To be specific, we survey the overlap among the contextual text sources of external URLs, OCR'ed text and Google Image Search for our dataset in Figure 6.5. As we can see, Google Image Search has large overlaps with external URLs and OCR text. For these overlaps, the other two sources are direct context indicated by image tweet's author, and are believable to provide more accurate semantics

for the image tweet than the SERP-extracted text. Take the two meme-styled images in Figure 6.4 as an example. The best guess description from Google Image Search is "no adulting meme" and "india pakistan match troll", respectively, while none of them reveals the correct semantics as the OCR text does.

As such, instead of merely polling all four sources of contextual text, we can fuse them more opportunistically to cut down computation as well as improve text quality. The tweet's original textual post and enhanced hashtags form the basis for fusion, as they are the most obvious context indicated by the author. We then propose a filtered fusion approach (illustrated in Figure 6.6) to use text obtainable from the other three sources: 1) for an image tweet with an embedded URL, we fuse only the text from its external web page, since the external page is the most accurate and accessible semantic context for the image; 2) for the remainder, we apply OCR on the image and if it contains embedded text, we fuse its OCR text recognized by Tesseract; 3) but if no embedded text is found, we obtain and fuse the SERP-extracted text from Google Image Search. It is worth to note that the fusion strategy helps to reduce the acquisition cost of contexts by 18.0% in our dataset (when treating all API calls as a unit cost), and provides better semantic modelling for image tweets (demonstrated in Section 6.5.3).

## 6.4 Personalized Image Tweet Recommendation

Given the CITING contexts that encode image tweets' semantics, we then apply them to the task of personalized image tweet recommendation. Modelling a user's interest is a fundamental task in social media. To the best of

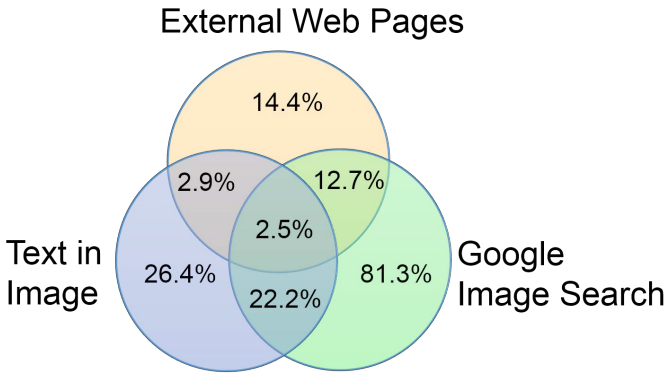



**Figure 6.5: The percentage of image tweets in our dataset that benefit from three major sources and their overlaps. Note although 22.7% of tweets have external URLs, only about two-third (63.2%) were still accessible.**

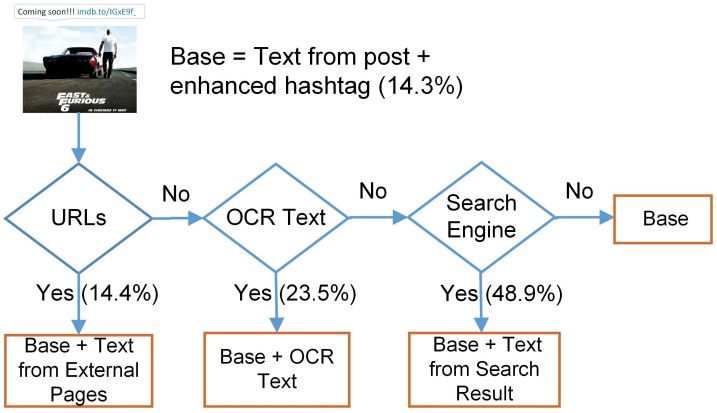



**Figure 6.6: Our rule cascade for fusing text from context sources. The % denotes the coverage of each source alone after fusion.**

our knowledge, this is the first study that learns user's interest from image tweets. To be specific, for a particular user, we aim to model her interest from her previous history, and predict her interest in incoming new image tweets. A direct application is to reorder the image tweets in user's feeds according to their interestingness.

In the following, we first discuss the weakness of traditional collaborative filtering techniques in tweet recommendation, and then detail our proposed feature-aware Matrix Factorization model, a generic method that can incorporate various features for image tweet recommendation.

### 6.4.1 Drawbacks of Collaborative Filtering

Collaborative filtering (CF) is acknowledged to be the most effective and generalizable technique for recommender system [114]. The basic idea is to predict a user's preference based on the histories of other similar users. For example, Matrix Factorization (MF), the most popular CF method, projects users and items into latent space to encode the preference similarity. As CF is designed to operate on the user–item interaction matrix, it represents an item as an ID and thus learns user's preference on item IDs.

We highlight that a key weakness of CF is its inapplicability to new items that have not yet attracted any interaction. This is also known as the *cold start* problem [2]. We show an illustrative example in Figure 6.7, where the rightmost two columns denote two new images tweets, which have never been seen in the training set. In this case, CF will fail to infer users' interest on the new items, and the prediction on new items is no better than random. This phenomenon is even exacerbated in social media like Twitter, due to the medium's strong timeliness and dynamicity. Said differently, old tweets (that are in training set) can quickly become dated and unattractive, while new tweets can be interesting but never appear in training. This makes the traditional CF technique unsuitable for the tweet recommendation task.

### 6.4.2 Feature-aware MF Framework

To overcome the defect of CF, one solution is to go beyond modelling the interaction of user and item IDs to more dense interactions, *e.g.*, their features. As a consequence, although the ID of a new item has not been seen before, we can still effectively infer a user's preference on the new item based on its features (that have been learnt in training). This motivates us to develop a generic model that can capture the interaction with various

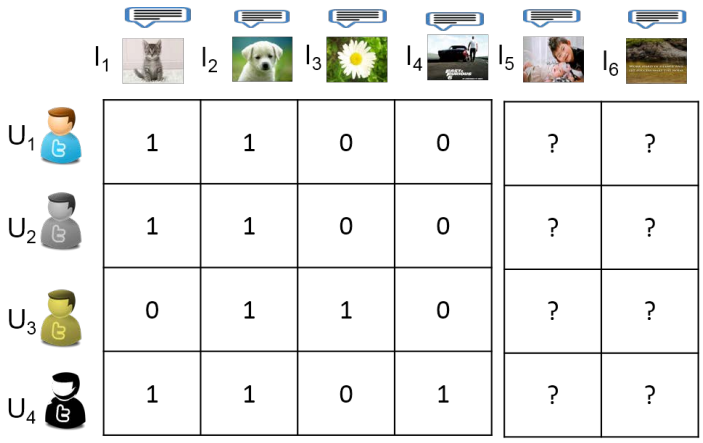



**Figure 6.7: An example of user–item matrix, where 1 denotes the user has retweeted the image tweet and 0 otherwise. The rightmost two columns denote two new items that cause the cold-start problem.**

features for recommendation.

Following the paradigm of factorization machines (FM) [112], in our framework, we transform the user–item interaction matrix to a set of feature vectors (model input) and a vector of interactions (target). As illustrated in Figure 6.8, each row denotes an interaction consists of user ID, item ID, and the various types of contextual features of the item. In our feature-aware MF model, we learn a low-dimensional representation (also termed as "latent vector") for each user and feature. Suppose we have $N$ types of features that represent image tweets, then our model estimates the preference of user $u$ on image tweet $i$ as:

$$\hat{y}_{u,i} = \mathbf{v}_u^T (\sum_{n=1}^{N} \frac{1}{Z_{n,i}} \sum_{f \in \mathbb{F}_{n,i}} \mathbf{q}_f), \tag{6.1}$$

where $\mathbf{v_u}$ and $\mathbf{q_f}$ denote the latent vector for user $u$ and feature $f$, respectively. $\mathbb{F}_{n,i}$ denotes the feature set of item $i$ of the $n^{th}$ feature type, and $Z_{n,i}$ is a normalizer[7] for features.

The model is generic in incorporating any types of features (and com-

[7]Empirically, we find $Z_i = \sqrt{|\mathbb{F}_{n,i}|}$ leads to good result.

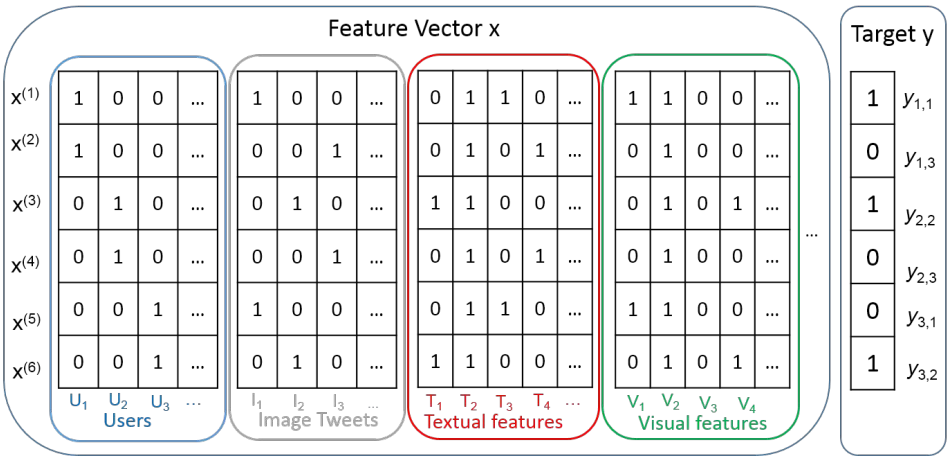



**Figure 6.8: Example feature vectors for the user–item matrix in Figure 6.7. Each row consists of user ID, item ID, and the various features of the item. The rightmost column is the prediction target.**

binations). In the case of image tweets, they can be image features (*e.g.*, visual objects), textual words in tweets and our proposed contexts[8]. We hypothesize that incorporating our proposed contextual features will better capture the rich semantics in image tweets, which will lead to better personalized tweet recommendation.

We point out that the key difference with FM is in the feature interactions considered—FM models the interactions between all pairs of features (including feature pairs in the same feature type), while we only model the interactions between user and item's features. Our design choice is for the model's interpretability. By modelling only the interaction between user and feature, we can interpret user $u$'s preference on feature $f$ as the inner product $\mathbf{v}_u^T \mathbf{q}_f$, which benefits the explainability [50] of our recommendation model.

[8]Note that we exclude item IDs as features into the vector (grayed in Figure 6.8), since most items in test set are not observed in training. Excluding item IDs will favor the prediction of cold-start items and lead to better results.

### 6.4.3 Learning from Implicit Feedback

The objective of tweet recommendation is to provide a user with a personalized ranked list of tweets. From a user's observed behaviors (*e.g.,* retweets), we naturally have explicit positive feedback that represent which tweets users are interested in. These positive tweets should be ranked higher than other negative tweets for the user. This idea fits the pair-wise Learning to Rank (LeToR) framework, and we adopt the *Bayesian Personalized Ranking* [114], which learns the model to maximize the probability that positives should have higher score than negatives for all users:

$$p(.|\mathbf{V}, \mathbf{Q}) = \prod_{u \in \mathbb{U}} \prod_{i \in \mathbb{P}_u} \prod_{j \notin \mathbb{P}_u} \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}), \tag{6.2}$$

where $\mathbb{P}_u$ denotes the positive tweets for user $u$, and $\sigma$ is the sigmoid function that projects the pair-wise difference into probability space. Maximizing the above probability is equivalent to minimize the following loss function:

$$\mathcal{L} = \sum_{u \in \mathbb{U}} \sum_{i \in \mathbb{P}_u} \sum_{j \notin \mathbb{P}_u} \log \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}) + \lambda_1 \left\|\mathbf{v}_u\right\|^2 + \lambda_2 \left\|\mathbf{q}_f\right\|^2, \tag{6.3}$$

where $\|\cdot\|$ denotes the $L_2$ norm for preventing model overfitting, and $\lambda_1$ and $\lambda_2$ are tunable hyper-parameters that control the extent of regularization.

As the number of training instances is very large (all user–item pairs) and there does not exist a closed form solution for model's parameters, learning is usually done by stochastic gradient descent (SGD). In each descent step, the localized optimization is performed on a tuple $(u, i, j)$.

The gradients with respect to each parameter are given as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{v}_u} &= -\hat{e}_{u,i,j} \sum_{n=1}^{N} (\frac{1}{Z_{n,i}} \sum_{f \in F_{n,i}} \mathbf{q}_f - \frac{1}{Z_{n,j}} \sum_{f \in F_{n,j}} \mathbf{q}_f) + \lambda_1 \mathbf{v}_u, \\
\frac{\partial \mathcal{L}}{\partial \mathbf{q}^i_{n,f}} &= -\frac{1}{Z_{n,i}} \hat{e}_{u,i,j} \mathbf{v}_u + \lambda_2 \mathbf{q}^i_{n,f}, \\
\frac{\partial \mathcal{L}}{\partial \mathbf{q}^j_{n,f}} &= \frac{1}{Z_{n,j}} \hat{e}_{u,i,j} \mathbf{v}_u + \lambda_2 \mathbf{q}^j_{n,f},
\end{aligned} \tag{6.4}$$

where $\hat{e}_{u,i,j}$ is defined as

$$\hat{e}_{u,i,j} = \frac{e^{-(\hat{y}_{u,i} - \hat{y}_{u,j})}}{1 + e^{-(y_{\hat{u},i} - y_{\hat{u},j})}}. \tag{6.5}$$

Then we iteratively loop over all the $(u, i, j)$ tuples in the training set, and update the parameters by moving them in the direction of negative gradient weighted by a learning rate until convergence. Learning rate is a key hyper-parameter for SGD that determines the speed of moving towards the optimal values—too large we will skip the optimal solution, while too small we need many iterations to converge. As such, we adopt *Bold Driver* [42], a technique that adjusts learning rate adaptively in each iteration. To be specific, it increases the learning rate by 5% if error rate is reduced since the last iteration; otherwise, resets the parameters to the values of the previous iteration and decreases the learning rate by 50%. In the experiments (detailed in next Section), our model achieves a fast speed of convergence, converging within 30 iterations for most settings.

### Time-aware Negative Sampling

As shown in [113], the way of sampling negative instances plays an important role in the efficacy of pair-wise learner. In recommendation system literature, uniform sampling is most widely used due to its simplicity and acceptable performance [114]. For Twitter, however, due to the large and fast-evolving item space, a uniform sampling will converge slowly and

suboptimally. To tackle this, previous works on text tweet recommendation [155, 27, 53, 128] sampled negative instances from the tweets posted by the target user's followees, rather than from the whole twitter space. In this case, they regarded the tweets that were not retweeted by the user as negative feedback.

However, we argue that a non-retweeted post does not necessarily mean it is disliked by the user. There is a high possibility that the user has not viewed the post at all, due to the overwhelmed information but limited reading time. As such, it is unsuitable to uniformly sample negatives from these non-retweets, and a better solution is to differentiate the sampling based on how likely the user has seen the tweet but disliked it. To achieve this, we reasonably assume that if a user has retweeted a post, she should also have read other tweets (of her followees) that were posted closely to the retweeting time. Thus, we propose a time-aware negative sampling strategy: given a positive image tweet interaction $rt$, we sample non-retweeted image tweets (posted by her followees) in proportion to the time interval between the post and $rt$, that is, posts closer to the time of $rt$ have a higher chance of being selected. Our featured negative sampling method can better assist the pair-wise learning, and we empirically demonstrate its efficacy in Section 6.5.4.

## 6.5 Experiment

We now evaluate CITING, our framework for context-aware image tweets modelling in the task of personalized image tweets recommendation. The goal of our experiment is to answer the following four research questions:

- **RQ 1:** What is the efficacy of the four proposed contexts?
- **RQ 2:** Do the filtered fusion improve model quality?

- **RQ 3:** Can time-aware negative sampling strategy construct better training set than uniform sampling?
- **RQ 4:** Are visual objects sufficient to capture Twitter images' semantics?

### 6.5.1 Experimental Settings

**Dataset.** To construct a dataset, we gathered image tweets from Twitter in a user-centric manner. We first crawled one week of public timeline tweets (8–14 December 2014) which resulted in a set of 5,919,307 tweets, of which 17.2% contained images. From this collection, we randomly sampled 926 users who had at least 100 followees and 100-3000 followers, and posted at least 100 tweets. We placed such requirements to select ordinary and relatively active users, as have similarly done by [128] to construct dataset in a user-centric manner. These 926 users are regarded as target users for our recommendation task.

We then crawled their latest tweets (up to 3,200—limited by the Twitter API), their followee list and further crawled the image tweets published by their followees. In particular, given a user and her retweet $rt$, we sample 10 non-retweeted image tweets according to the time-aware negative sampling strategy in Section 6.4.3. This process results in a dataset of 1,369,133 image tweets (demographics in Table 6.3). To simulate the real recommendation scenario, we adopt a time-based evaluation. For each user, we use her most recent 10 retweets as the test set, giving the rest for training. Note that the user–tweet interaction is extremely sparse: each image tweet is retweeted by 1.22 users on average, and only 31% of the testing tweets have previously been observed in the training set. This validates the sparsity observations in previous works [27, 37].

**Evaluation Metrics.** The objective of tweet recommendation is to

**Table 6.3: Training and test set demographics.**

| | **Users** | **Retweets** | **All Tweets** | **Ratings** |
|---|---|---|---|---|
| Training | 926 | 174,765 | 1,316,645 | 1,592,837 |
| Test | | 9,021 | 77,061 | 82,743 |

rank the candidate tweets such that the interesting tweets are placed at top for the target user. In our case, we mix the testing retweets (*i.e.,* ground-truth) and their negative samples as the candidate tweets for each user. To access the ranking quality, we adopt the average precision at rank $k$ (P@$k$) and Mean Average Precision (MAP) as the evaluation metrics, which have been widely used for the tweet recommendation task [27, 37, 53]. Since users are most interested with the top few recommendations, we report P@$k$ at very top ranks ($k$=1, 3 and 5).

**Parameter Settings**. We tune two regularization parameters ($\lambda_1$ and $\lambda_2$) and the dimensionality of latent factors $K$. We first vary the regularizers until the results are generally stable, and then carefully tune $K$ in a grid search manner (from 10 to 200). We report the performance at $\lambda_1 = 0.05$, $\lambda_1 = 0.01$ and $K = 160$, which shows good results. Similarly, we tune the parameters for other methods, and report their optimal results accordingly. For all the experiments, we set the initial learning rate as 0.01.

### 6.5.2 Utility of Proposed Contexts (RQ 1)

In this subsection, we study the efficacy of our proposed four strategies for context mining. To this end, we add the obtained text from each source to the post's original text separately, and experiment with each of the combined text. For webpages, we separate the title and page content in evaluation, since we find some pages only have title while lacking the main content or vice versa. Observing that some webpages can be very long and only the top few words are most relevant, we use the top 20 words as the page content.

**Table 6.4: Performance of each context source and its coverage. The best single context is the title of Google image search pages.**

| | **P@1** | **P@3** | **P@5** | **MAP** | **Coverage** |
|---|---|---|---|---|---|
| P: Post | 0.359 | 0.325 | 0.287 | 0.275 | |
| P + Hashtag | 0.360 | 0.324 | 0.293 | 0.277 | 14.3% |
| P + OCR text | 0.366 | 0.332 | 0.301 | 0.283 | 26.4% |
| P + URL (title) | 0.374 | 0.326 | 0.294 | 0.278 | 14.2% |
| P + URL (content) | 0.381 | 0.330 | 0.300 | 0.279 | 13.2% |
| P + G (content) | 0.369 | 0.319 | 0.289 | 0.275 | 57.2% |
| **P + G (title)** | **0.388** | **0.344** | **0.308** | **0.288** | 76.0% |
| P + G (guess+NE) | 0.381 | 0.330 | 0.296 | 0.280 | 58.1% |

Table 6.4 shows the performance of each source with its coverage. In general, all context sources show a positive impact on the recommendation performance[9]. Of which, we find that the gains from the two external sources (external URL and Google Image Search) are more significant than the two internal sources (hashtag and OCR text). This validates the usefulness of external knowledge for interpreting images' semantics in social media. The largest improvement is obtained by integrating the titles of Google indexed pages, with a relative 8.1% and 4.7% improvement over using post's text only, in terms of P@1 and MAP, respectively. This is partially owing to the high coverage of Google Image Search over social media images. However, using the actual page content of the Google indexed pages neither improves over titles, nor betters the post's text—even degrading the performance for P@3 and MAP. With a deep analysis, we find this might be caused by the noise introduced by Boilerpipe when extracting the main text from SNS pages and image aggragator sites. These sites make up a large portion in Google's indexed pages (84.3%) but their layouts significantly differ from news and blogs that Boilerpipe was trained on. As a result, Boilerpipe suffers from a high error rate. Thus in our later

[9]Although the P@3 slightly degrades for the source hashtag, other metrics still reveal it as a helpful feature.

experiments, we use all contextual text except the actual content of Google indexed pages.

### 6.5.3 Effectiveness of Context Fusion (RQ 2)

We now evaluate the effectiveness of filtered fusion approach. For comparison purpose, we report the performance of our feature-aware MF model using all context without the filtered fusion (**Non-filtered**) and Post's text only (**Post**). The latter is equivalent to the result by two state-of-the-art models [27, 37] on text tweet recommendation, as the two are special cases of our model when only post's text is considered. To benchmark the performance, we also consider baselines: 1) **Random**: rank image tweets randomly; 2) **Length**: rank image tweets by the number of words in post's text, and the intuition is that longer tweets tend to be more informative and possibly to be more popular [143]; 3) **Profiling**: rank image tweets by the similarity of tweets' text and user's profile, which is constructed from the words of user's historical posts and retweets. To be specific, given a user $u$ and an image tweet $t$, we compute the profile-based similarity score as follows:

$$S_{u,t} = (1 - w) \times cos(posts(u), t) + w \times cos(retweets(u), t),$$

where $cos$ denotes the cosine similarity and $w$ is a tunable parameter for combining the posting and retweeting history.

Table 6.5 shows the results. First, our proposed filtered fusion (CITING, R6) outperforms the three baselines (random, length, profiling) by a large margin. The filtered fusion method also significantly betters the strong baseline of using post's text by 0.06 (16.9% relative improvement) and 0.023 (8.3%) in terms of P@1 and MAP, respectively. When adopting non-filtered fusion approach, the performance slightly drops, *e.g.*, the

**Table 6.5: Performance comparison between our CITING and other approaches. '**' denotes the difference between our method and the other method is statistically significant with $p < 0.01$, and '*' for $p < 0.05$.**

| | **P@1** | **P@3** | **P@5** | **MAP** |
|---|---|---|---|---|
| (1): Random | 0.114** | 0.115 | 0.115 | 0.156** |
| (2): Length | 0.176** | 0.158 | 0.150 | 0.173** |
| (3): Profiling | 0.336** | 0.227 | 0.197 | 0.202** |
| (4): Post | 0.359* | 0.325 | 0.287 | 0.275** |
| (5): Non-filtered | 0.413 | 0.352 | 0.319 | 0.296 |
| (6): **CITING** | **0.419** | **0.355** | **0.319** | **0.298** |

P@1 drops from 0.419 to 0.413. Although not statistically significant, it indicates that our heuristic filtered fusion approach achieves comparable results while saving acquisition costs of the contextual text by 18.0%. These experimental results evidence the effectiveness of our fusion approach and the feature-aware MF model.

### 6.5.4 Importance of Negative Sampling (RQ 3)

To gain more insights into the tweet recommendation task, we now assess the effect of the negative sampling strategy. We compare with the uniform sampling strategy, which is a commonly used strategy by previous works in tweet recommendation [155, 27, 128].

To this end, we constructed a new dataset by uniformly sampling negative image tweets from our training set and pair with the positive image tweets. We then trained our feature-aware MF on this new dataset, using our proposed filtered contexts, and evaluated the method in the same way. Experimental result shows the time-aware sampling strategy significantly betters the random sampling by 0.017 (4.2% relative improvement) and 0.006 (2.1%), for P@1 and MAP, respectively. We conducted a one-tailed

paired $t$-test for both P@1 and MAP, showing that both $p$ values are smaller than 0.05. This validates our time-aware negative sampling strategy is effective in constructing training set of higher quality, and thus aid to learn a better user interest model.

### 6.5.5 Insufficiency of Visual Objects (RQ 4)

We now validate our claim at the outset that annotating visual objects without context does not fare well for social media interpretation. First, we applied *GoogLeNet* [126], the winning system in the recent ILSVRC 2014, to classify the visual objects for our Twitter images. GoogLeNet is trained on 1.2 million Flickr images with 1000 object categories, and each category corresponds to a node in ImageNet/WordNet. The pre-trained model is provided by Caffe [63]. We take the top five labels as the description for each image and conduct the same experiment. We see that prediction using just visual objects does perform worse (P@1= 0.221, MAP= 0.211; Table 6.6), due to its literal description of the images. Our CITING context significantly outperforms visual objects by 89.2% of relative improvement and 40.9% in terms of P@1 and MAP, respectively. This shows the contextual text does capture image tweets' semantics much better.

For comprehensive purpose, we further experiment with the combination of text and visual objects[10], to see whether the incorporation of visual cues could further boost the recommendation performance. As shown in Table 6.6 (R3), the integration of visual objects with post's text slightly improves over post's text 5.6% (relative improvement) and 1.8%, for P@1 and MAP, respectively, while our CITING context still significantly betters such combination by relatively 10.5% and 6.2%. This further validates our contextual text is able to capture semantics of image tweets better. Unlike

[10]The two are considered as two types of features in our feature-aware MF model.

**Table 6.6: Performance of using visual objects.**

| | **P@1** | **P@3** | **P@5** | **MAP** |
|---|---|---|---|---|
| (1): CITING | 0.419 | 0.355 | 0.319 | 0.298 |
| (2): Visual objects (V) | 0.221 | 0.205 | 0.192 | 0.211 |
| (3): Post's text + V | 0.379 | 0.325 | 0.293 | 0.280 |
| (4): CITING + V | 0.425 | 0.350 | 0.313 | 0.298 |

the previous combination, the incorporation of visual objects does not lead to a stable improvement for contextual text: P@1 is slightly improved by 0.006 (1.4%), and MAP remains the same, while the other two metrics drops. This suggests the new and useful information brought by visual objects is very limited, and such visual cues might have already been largely captured by our contextual text (*e.g.*, some best guess descriptions from Google Image Search have visual objects information).

### 6.5.6 Case Studies

While macro-level empirical analysis is useful, it is also instructive to examine individual users and actual posts to better understand the effectiveness of our proposed filtered contextual text. To this end, we examine a few users whose recommendations have a large performance gain when using our proposal. In Figure 6.9, we show such a typical user (refer as *User 1*) and four of her retweets in test set that are enriched by our contextual text. As a consequence, the average recommendation precision of our approach (0.592) significantly outperforms the approach of using visual objects (0.226) and using post's text (0.443). In an even more successful case (shown in Figure 6.10), 9 out of 10 retweets for *User 2* obtained contextual text from our approach. The average precision is boosted from 0.423 (using visual objects) and 0.319 (using post's text) to 0.728 (our approach).

Taking a closer look at the these image tweets, we find a few of them

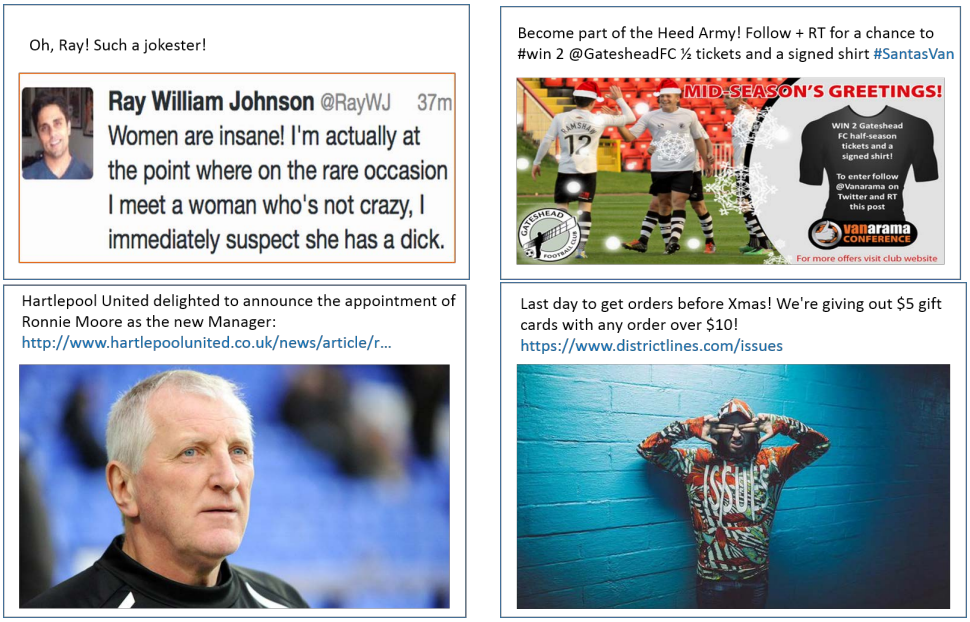



**Figure 6.9: 4 image tweets from *User 1*'s retweets in testing set benefit from our proposed contextual text. For this user, the average precision of using visual objects, post's text, and our proposal are** 0.226, 0.443 **and** 0.592**, respectively.**

trigger multiple context mining channels. Some have both embedded URL and overlaid text in image (see Figure 6.10: the top leftmost and the bottom rightmost). A further investigation shows the external html pages redirected by the URLs contain richer and more relevant information than the overlaid text. This validates our text fusion strategy 1—giving a higher priority of using text from URL than OCR text—is effective. Another image tweet (Figure 6.9, top leftmost) has both overlaid text and search result from Google Image, however, the search result only indicates the image is a quote, but does not reveal its deep semantics as OCR text does. In this case, OCR text is more reliable source than search result, validating our text fusion strategy 2.

## 6.6 Conclusion

Understanding the semantics of social media images is fundamental research. Compared to traditional vision research on stock photo images,

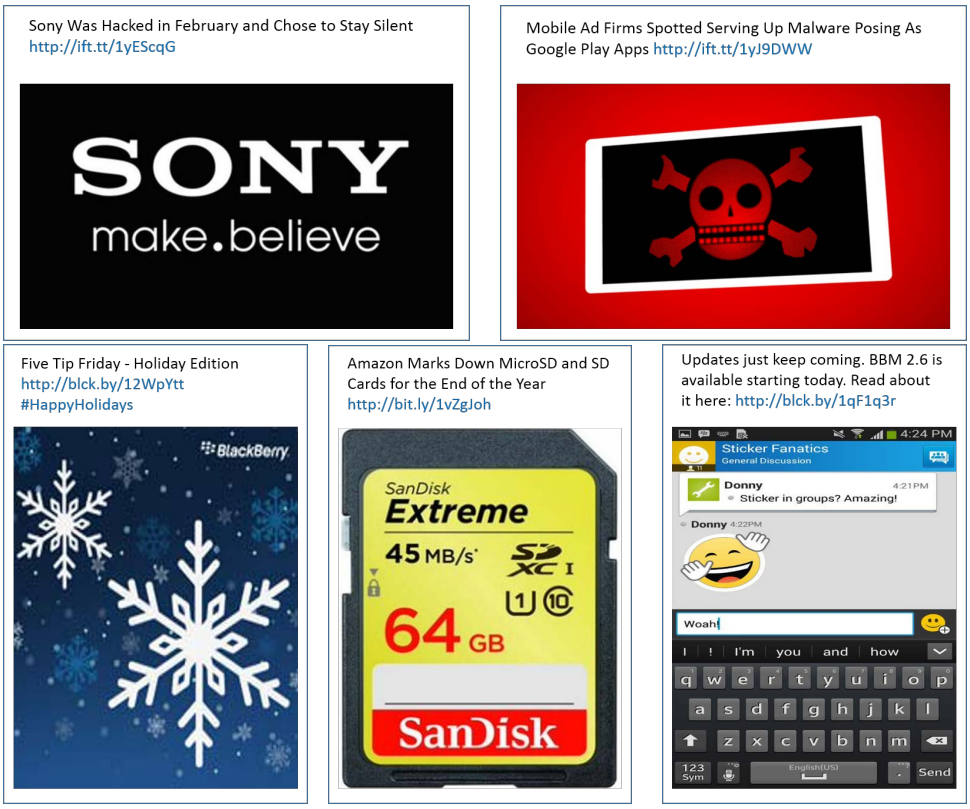



**Figure 6.10: For** *User 2*, **9 out of 10 image tweets (here we show 4) in test set benefit from our contextual text. For this user, the average precision of using visual objects, post's text, and our proposal are** 0.423, 0.319 **and** 0.728, **respectively.**

we have shown that social media images vary much more widely and need to be understood within their context of mention and that visual features (*e.g.*, visual objects) are insufficient for this. However, the context of most social media images are accessible, through accessing appropriate sources. We mine contextual text for an image tweet both intrinsically (*i.e.*, the image and its accompanying text) and extrinsically (*i.e.*, via hyperlinked web pages and the indexed Web). Through our proposed heuristic context acquisition and fusion, the representation of such images is significantly more accurate as evaluated on the task of personalized image tweet recommendation.

We have done an analysis of the coverage and efficacy of acquiring textual context of social images, but there is still much that can be improved here. To spur additional research on social media images, we have released

our large image tweets dataset, as well as our annotated corpus of 500 sample images with their manually-recognized text[11]. In particular, future work can adapt OCR to better acquire text within the images, as current OCR fares poorly on meme-style images and graphics.

[11] http://wing.comp.nus.edu.sg/downloads/image-tweet-ocr-rec

# Chapter 7

# Conclusion and Future Work

A picture is worth a thousand words. Images have been widely adopted as a means for conveying information in microblog platforms. Unlike its text-only counterpart, existing research on image tweet is very limited, and our understanding of image tweets is still shallow. To fill this gap, we conducted a series of studies to answer four fundamental questions about image tweets: 1) What are the characteristics of image tweets? 2) What are the relationships between the image and text? 3) How to model image-text relations? and 4) What is an appropriate method to interpret the latent semantics of an image tweet?

## 7.1 Main Contributions

This thesis makes the following main contributions in analyzing image tweets:

1. We present a complete picture of existing studies (up to Jan 2016) on image tweets, which may serve as a reference for future research in this area.

2. We uncover the image characteristics, the user's behaviors (*e.g.*, access, temporal, reaction), and textual content of image tweets in both

Western (Twitter) and Chinese (Weibo) microblog.

3. We identify two key image-text relations from image tweets, namely, visual relevance and emotional relevance, and build a classifier to automatically distinguish the visual and non-visual relation.

4. We develop Visual-Emotional LDA (VELDA), a novel topic model to capture the image-text correlation from multiple perspectives (*i.e.*, visual and emotional), and thus model the generative process of image tweets.

5. We propose a context-aware image tweets modeling (CITING) framework to mine intrinsic and extrinsic contextual text to uncover an image tweet's semantics, and apply these strategies to user interest modeling, a key application in microblog.

To enable comparative studies, we have released the following datasets to the public:

6. *Weibo-Rel*—4.8K Weibo image tweets with human annotated image-text relations (described in Chapter 4.4.1)[1].

7. *Twitter-OCR*—500 Twitter images with the embedded text recongized by human (described in Chapter 6.3.1)[2].

8. *Twitter-Rec*—1.3 million image tweets for 926 Twitter users' retweeting history (described in Chapter 6.5.1)[3].

## 7.2 Future Work

Image in microblogs is not singleton. It is accompanied by post's text and extended context, and involves a series of users during its lifespan,

[1] `http://wing.comp.nus.edu.sg/downloads/imagetweets`
[2] `http://wing.comp.nus.edu.sg/downloads/image-tweet-ocr-rec`
[3] `http://wing.comp.nus.edu.sg/downloads/image-tweet-ocr-rec`

*e.g.*, creators who make it, adopters (may be the same person as creator) who use it to compose image tweet, disseminators who repost the image tweet, and commenters who express opinions about the image tweet. It embodies the social networked communication in a multimedia manner. As such, research on image tweets is multidisciplinary. It includes several areas such as multimedia, natural language processing, user modeling, search and filtering, social networks, social science, data mining and machine learning, among others. In this thesis, we have answered four foundational questions about image tweets. Still, there are several extensions for our work, and also some open questions that could be addressed for further research.

1. **Deep learning image features for modeling image-text relations.** In our VELDA model, we adopt quantized SIFT descriptors and 22 color-based features to present image's visual and emotional view, respectively. Recent developments in deep neural network approaches have greatly advanced the performance of various visual tasks, such as image recognition [117] and image sentiment classification [29, 150, 149]. Aside from using such architectures as an end-to-end system, the very last fully connected layers from the neural networks (aka deep learning features) can be utilized as a feature representation for other models. As such, we would like to incorporate deep learning features into our VELDA model, and explore whether a better feature representation could further improve the performance of VELDA in modeling image-text relations.

2. **Learning the importance of each context.** In our CITING framework, we heuristically fuse contexts based on their textual quality, and then all textual words are modeled with equal importance under our feature-aware Matrix Factorization (MF) model. We feel it might be even better to differentiate the importance of textual

words by their contextual source. This goal can be easily achieved by extending our feature-aware MF model. To be specific, our MF framework will regard each context as different type of features and weight feature's latent factor by their contextual importance. Such importance scores are parameters in our model and could be automatically learned in training phase.

3. **Other contexts for image's semantics mining.** We are aware of microblog image's context is not limited to the four that we have proposed. Previous works [38, 124] have pointed out photo's metadata (*e.g.*, time and location) are useful in understanding image's content and topics. For instance, a picture taken in a famous landmark is very likely to capture the scene of the landmark. Unfortunately, all the mainstream microblog platforms (*e.g.*, Twitter, Weibo) have scrubbed the Exif metadata from uploaded images. But we feel the posting time and geotag of image tweets can be used as an approximation, especially for those generated by mobile devices. The other context that we can exploit is the identities of microblog users related to the post—*e.g.*, the author, the retweeter, and the commenter of an image tweet. These users execute certain behaviors based on their own interests, and thus we may be able to infer the semantics of an image tweet from related users' historical interests.

4. **Images in conversation.** In microblogs, images are not solely used in a post but also in a reply. As we mentioned at the very beginning of this thesis, both Twitter and Weibo allow users to embed images in a reply/comment. We term such posts as *photo comments*. Text-based conversations have existed in microblogs since their origin, and this form of dialog has been studied in several works, *e.g.*, model conversations and discover speech acts in posts and their replies [115],

automatically generate replies to microblog posts [116]. It will be very interesting to investigate photo comments and the multimedia conversation. Why people include an image in chatting? What are the characteristics of photo comments? What kinds of text tweets or image tweets are more likely to trigger photo comments? If both the original post and the comment contain an image, what are the correlation of the two images? By answering questions, we will also gain a deep understanding of human communication.

# Bibliography


[1] S. Abdullah, E. L. Murnane, J. M. Costa, and T. Choudhury. Collective Smile: Measuring Societal Happiness from Geolocated Images. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '15, pages 361–374, 2015.

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June 2005.

[3] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Enriching Textbooks with Images. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1847–1856, 2011.

[4] O. Alonso, C. C. Marshall, and M. Najork. Are Some Tweets More Interesting Than Others? #HardQuestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '13, pages 2:1–2:10, 2013.

[5] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Systems*, 16(6):345–379, 2010.

[6] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo. A Multimodal Feature Learning Approach for Sentiment Analysis of Social Network Multimedia. *Multimedia Tools and Applications*, pages 1–19, 2015.

[7] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I.

Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 3:1107–1135, Mar. 2003.

[8] P. Berinstein. Moving Multimedia: The Information Value in Images. *Searcher*, 5(8):40–46, 1997.

[9] J. Bian, Y. Yang, and T.-S. Chua. Multimedia Summarization for Trending Topics in Microblogs. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 1807–1812, 2013.

[10] J. Bian, Y. Yang, and T.-S. Chua. Predicting Trending Messages and Diffusion Participants in Microblogging Network. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 537–546, 2014.

[11] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua. Multimedia Summarization for Social Events in Microblog Stream. *IEEE Transactions on Multimedia*, 17(2):216–228, Feb 2015.

[12] D. M. Blei and M. I. Jordan. Modeling Annotated Data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 127–134, 2003.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[14] C. Boididou, S. Papadopoulos, D. Dang-Nguyen, G. Boato, and Y. Kompatsiaris. The CERTH-UNITN Participation @ Verifying Multimedia Use 2015. In *Proceedings of the MediaEval 2015 Workshop*, September 2015.

[15] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman. Challenges of Computational Verification in Social Multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 743–748, 2014.

[16] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Pro-*

*ceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 223–232, 2013.

[17] M. Bressan, G. Csurka, Y. Hoppenot, and J.-M. Renders. Travel Blog Assistant System (TBAS): An Example Scenario of How to Enrich text with Images and mages with Text using Online Multimedia Repositories. In *VISAPP Workshop on Metadata Mining for Image Understanding*, MMIU '08, 2008.

[18] H. Cai, Z. Tang, Y. Yang, and Z. Huang. EventEye: Monitoring Evolving Events from Tweet Streams. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 747–748, 2014.

[19] H. Cai, Y. Yang, X. Li, and Z. Huang. What Are Popular: Exploring Twitter Features for Event Detection, Tracking and Visualization. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 89–98, 2015.

[20] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou. Diving Deep into Sentiment: Understanding Fine-tuned CNNs for Visual Sentiment Prediction. In *Proceedings of the 1st International Workshop on Affect and Sentiment in Multimedia*, ASM '15, pages 57–62, 2015.

[21] E. F. Can, H. Oktay, and R. Manmatha. Predicting Retweet Count using Visual Cues. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 1481–1484, 2013.

[22] D. Cao, R. Ji, D. Lin, and S. Li. A Cross-media Public Sentiment Analysis System for Microblog. *Multimedia Systems*, pages 1–8, 2014.

[23] S. Cappallo, T. Mensink, and C. G. Snoek. Latent Factors of Visual Popularity Prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 195–202, 2015.

[24] R. Carney and J. Levin. Pictorial Illustrations Still Improve Students' Learning from Text. *Educational Psychology Review*, 14:5–26, 2002.

[25] C. Chen, F. Chen, D. Cao, and R. Ji. A Cross-media Sentiment Analytics

Platform For Microblog. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 767–769, 2015.

[26] F. Chen, Y. Gao, D. Cao, and R. Ji. Multimodal Hypergraph Learning for Microblog Sentiment Prediction. In *2015 IEEE International Conference on Multimedia and Expo*, volume ICME '15, pages 1–6, June 2015.

[27] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative Personalized Tweet Recommendation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 661–670, 2012.

[28] S. Chen, H. Zhang, M. Lin, and S. Lv. Comparision of Microblogging Service between Sina Weibo and Twitter. In *2011 International Conference on Computer Science and Network Technology (ICCSNT)*, volume 4, pages 2259 –2263, 2011.

[29] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1410.8586*, 2014.

[30] T. Chen and M.-Y. Kan. Creating a Live, Public Short Message Service Corpus: the NUS SMS Corpus. *Language Resources and Evaluation*, 47(2):299–335, 2012.

[31] T.-S. Chua, H. Luan, M. Sun, and S. Yang. NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content. *IEEE MultiMedia*, 19(3):81–87, 2012.

[32] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in Visual Information Retrieval. *IEEE MultiMedia*, 6(3):38–53, 1999.

[33] V. Conotter, D.-T. Dang-Nguyen, M. Riegler, G. Boato, and M. Larson. A Crowdsourced Data Set of Edited Images Online. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, CrowdMM '14, pages 49–52, 2014.

[34] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the Role of Correlation and Abstrac-

tion in Cross-Modal Multimedia Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, March 2014.

[35] D. Delgado, J. Magalhaes, and N. Correia. Assisted News Reading with Automated Illustration. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1647–1650, 2010.

[36] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[37] W. Feng and J. Wang. Retweet or Not?: Personalized Tweet Re-ranking. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 577–586, 2013.

[38] C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Bringing Order to Your Photos: Event-driven Classification of Flickr Images Based on Social Knowledge. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 189–198, 2010.

[39] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu. A Comparative Study of Users' Microblogging Behavior on Sina Weibo and Twitter. In *Proceedings of International Conference on User Modeling and Personalization*, UMAP '12, pages 88–101, 2012.

[40] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu. Information Propagation Cultures on Sina Weibo and Twitter. In *Proceedings of Web Science 2012*, WebSci '12, 2012.

[41] Y. Gao, F. Wang, H. Luan, and T.-S. Chua. Brand Data Gathering From Live Social Media Streams. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 169–176, 2014.

[42] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale Matrix Factorization with Distributed Stochastic Gradient Descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 69–77, 2011.

[43] J. Ghaznavi and L. D. Taylor. Bones, Body Parts, and Sex Appeal: An

Analysis of #thinspiration Images on Popular Social Media. *Body Image*, 14:54 – 61, 2015.

[44] W. Guan, H. Gao, M. Yang, Y. Li, H. Ma, W. Qian, Z. Cao, and X. Yang. Analyzing User Behavior of the Micro-blogging Website Sina Weibo During Hot Social Events . *Physica A: Statistical Mechanics and its Applications*, 395(0):340 – 351, 2014.

[45] S. C. Guntuku, L. Qiu, S. Roy, W. Lin, and V. Jakhetiya. Do Others Perceive You As You Want Them To?: Modeling Personality Based on Selfies. In *Proceedings of the 1st International Workshop on Affect and Sentiment in Multimedia*, ASM '15, pages 21–26, 2015.

[46] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy. In *2nd International Workshop on Privacy and Security in Online Social Media*, PSOSM '13, pages 729–736, 2013.

[47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: an Update. *SIGKDD Explorations Newsletter*, 11(1):10–18, Nov. 2009.

[48] M. Hancher. Illustrations [in "The Century Dictionary"]. *Dictionaries*, 17:79–115, 1996.

[49] J. S. Hare, S. Samangooei, D. P. Dupplaw, and P. H. Lewis. Twitter's Visual Pulse. In *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval*, ICMR '13, pages 297–298, 2013.

[50] X. He, T. Chen, M.-Y. Kan, and X. Chen. TriRank: Review-aware Explainable Recommendation by Modeling Aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1661–1670, 2015.

[51] R. Herrema. Flickr, Communities of Practice and the Boundaries of Identity: a Musician Goes Visual. *Visual Studies*, 26(2):135–141, 2011.

[52] L. Hong, O. Dan, and B. D. Davison. Predicting Popular Messages in Twitter. In *Proceedings of the 20th International Conference Companion*

*on World Wide Web*, WWW '11, pages 57–58, 2011.

[53] L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization Machines: Modeling User Interests and Predicting Individual Decisions in Twitter. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 557–566, 2013.

[54] E. Hörster, R. Lienhart, and M. Slaney. Image Retrieval on Large-scale Image Databases. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 17–24, 2007.

[55] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3-4):321–377, 1936.

[56] D. Hromada, C. Tijus, S. Poitrenaud, and J. Nadel. Zygomatic Smile Detection: The Semi-Supervised Haar Training of a Fast and Frugal System: A Gift to OpenCV Community. In *Proceedings of 2010 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, RIVF '10, pages 1–5, Nov 2010.

[57] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *Proc. of ICDM*, 2008.

[58] K. Ishiguro, A. Kimura, and K. Takeuchi. Towards Automatic Image Understanding and Mining via Social Curation. In *2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 906–911, 2012.

[59] R. Jain and P. Sinha. Content Without Context is Meaningless. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1259–1268, 2010.

[60] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, 2007.

[61] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can We Understand Van Gogh's Mood? Learning to Infer Affects from Images in Social

Networks. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 857–860, 2012.

[62] Y. Jia, M. Salzmann, and T. Darrell. Learning Cross-modality Similarity for Multinomial Data. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2407–2414, 2011.

[63] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

[64] Z. Jin, J. Cao, Y. Zhang, and Y. Zhang. MCG-ICT at MediaEval 2015: Verifying Multimedia Use with a Two-Level Classification Model. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[65] D. Joshi, F. Chen, and L. Wilcox. Finding Selfies of Users in Microblogged Photos. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14, pages 33–34, 2014.

[66] D. Joshi, J. Z. Wang, and J. Li. The Story Picturing Engine—a System for Automatic Text Illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):68–89, 2006.

[67] T. Kaneko and K. Yanai. Event Photo Mining from Twitter using Keyword Bursts and Image Clustering . *Neurocomputing*, 172:143 – 158, January Kaneko2016.

[68] T. Kharroub and O. Bas. Social Media and Protests: An examination of Twitter images of the 2011 Egyptian Revolution. *New Media and Society*, 2015.

[69] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate Detection Using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, 2010.

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification

with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, NIPS '12, pages 1106–1114, 2012.

[71] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, 2010.

[72] D. Laniado and P. Mika. Making Sense of Twitter. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web*, ISWC '10, pages 470–485, 2010.

[73] Q. V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning*, ICML '14, pages 1188–1196, 2014.

[74] I. P. Levin, S. L. Schneider, and G. J. Gaeth. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects . *Organizational Behavior and Human Decision Processes*, 76(2):149 – 188, 1998.

[75] J. Levin. Pictorial Strategies for School Learning: Practical Illustrations. In M. Pressley and J. Levin, editors, *Cognitive Strategy Research*, Springer Series in Cognitive Development, pages 213–237. Springer New York, 1983.

[76] Z. Li, M. Wang, J. Liu, C. Xu, and H. Lu. News Contextualization with Geographic and Visual Information. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 133–142, 2011.

[77] R. Liao, J. Zhu, and Z. Qin. Nonparametric Bayesian Upstream Supervised Multi-modal Topic Models. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 493–502, 2014.

[78] H. Lin, J. Jia, Q. Guo, Y. Xue, J. Huang, L. Cai, and L. Feng. Psychological Stress Detection from Cross-media Microblog Data using Deep Sparse Neural Network. In *2014 IEEE International Conference on Multimedia and Expo*, ICME '14, pages 1–6, July 2014.

[79] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng. User-level Psychological Stress Detection from Social Media Using Deep Neural

Network. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 507–516, New York, NY, USA, 2014. ACM.

[80] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[81] X. Lu, Y. Pang, Q. Hao, and L. Zhang. Visualizing Textual Travelogue with Location-relevant Images. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 65–68, 2009.

[82] Z. Ma, A. Sun, Q. Yuan, and G. Cong. Tagging Your Tweets: A Probabilistic Modeling of Hashtag Annotation in Twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 999–1008, 2014.

[83] J. Machajdik and A. Hanbury. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 83–92, 2010.

[84] L. Manikonda, R. Venkatesan, S. Kambhampati, and B. Li. Evolution of fashion brands on Twitter and Instagram. *eprint arXiv:1512.01174*, 2015.

[85] M. O. Manolo Farci. Hybrid Content Analysis of the Most Popular Politicians' Selfies on Twitter. *Networking Knowledge*, 8(6), Nov 2015.

[86] E. E. Marsh and M. D. White. A Taxonomy of Relationships Between Images and Text. *Journal of Documentation*, 59:647–672, 2003.

[87] R. Martinec and A. Salway. A System for Image–text Relations in New (and Old) Media. *Visual Communication*, 4(3):337–371, 2005.

[88] A. J. McMinn, D. Tsvetkov, T. Yordanov, A. Patterson, R. Szk, J. A. Rodriguez Perez, and J. M. Jose. An Interactive Interface for Visualizing Events on Twitter. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1271–1272, 2014.

[89] M. Merler, L. Cao, and J. R. Smith. You are What you Tweet...Pic! Gender Prediction based on Semantic Analysis of Social Media Images. In

*2015 IEEE International Conference on Multimedia and Expo*, ICME '15, pages 1–6, June 2015.

[90] S. E. Middleton. Extracting Attributed Verification and Debunking Reports from Social Media: MediaEval-2015 Trust and Credibility Analysis of Image and Video. In *Proceedings of the MediaEval 2015 Workshop*, September 2015.

[91] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at International Conference on Learning Representations*, 2013.

[92] M. Naaman, J. Boase, and C.-H. Lai. Is it Really about Me?: Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192, New York, NY, USA, 2010. ACM.

[93] M. Naaman, A. Zhang, S. Brody, and G. Lotan. On the Study of Diurnal Urban Routines on Twitter. In *International AAAI Conference on Web and Social Media*, ICWSM'12, 2012.

[94] Y. Nakaji and K. Yanai. Visualization of Real-World Events with Geotagged Tweet Photos. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops*, ICMEW '12, pages 272–277, 2012.

[95] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 8:1–8:7, 2011.

[96] J. E. Newhagen and B. Reeves. The Evening's Bad News: Effects of Compelling Negative Television News Images on Memory. *Journal of Communication*, 42(2):25–41, 1992.

[97] J.-Y. Nie, J. Gao, J. Zhang, and M. Zhou. On the Use of Words and N-grams for Chinese Information Retrieval. In *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*,

IRAL '00, pages 141–148, 2000.

[98] M. Nikolajeva and C. Scott. *How Picturebooks Work*. Children's literature and culture. Garland Pub., 2001.

[99] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '06, pages 2161–2168, 2006.

[100] T. Niu, S. Zhu, L. Pang, and A. El Saddik. Sentiment Analysis on Multi-View Social Data. In Q. Tian, N. Sebe, G.-J. Qi, B. Huet, R. Hong, and X. Liu, editors, *MultiMedia Modeling*, Lecture Notes in Computer Science, pages 15–27. 2016.

[101] J. Peeck. Increasing Picture Effects in Learning from Illustrated Text . *Learning and Instruction*, 3(3):227 – 238, 1993.

[102] S. Petrovic, M. Osborne, and V. Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *International AAAI Conference on Web and Social Media*, ICWSM '11, 2011.

[103] J. Pierce. *An Introduction to Information Theory: Symbols, Signals & Noise*. Dover Books on Mathematics Series. Dover Publications, 1980.

[104] A. Piva. An Overview on Image Forensics. *ISRN Signal Processing*, 2013.

[105] R. Plutchik. Emotions: A General Psychoevolutionary Theory. *Approaches to Emotion*, 1984:197–219, 1984.

[106] D. Putthividhy, H. Attias, and S. Nagarajan. Topic Regression Multi-modal Latent Dirichlet Allocation for Image Annotation. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '10, pages 3408–3415, 2010.

[107] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor Has It: Identifying Misinformation in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, 2011.

[108] S. Qi, F. Wang, X. Wang, J. Wei, and H. Zhao. Live Multimedia Brand-related Data Identification in Microblog . *Neurocomputing*, 158:225 – 233,

2015.

[109] S. Qian, T. Zhang, C. Xu, and J. Shao. Multi-modal Event Topic Model for Social Event Analysis. *IEEE Transactions on Multimedia*, PP(99), 2015.

[110] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing Microblogs with Topic Models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10, 2010.

[111] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A New Approach to Cross-modal Multimedia Retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 251–260, 2010.

[112] S. Rendle. Factorization Machines. In *Proceedings of 2010 IEEE International Conference on Data Mining*, ICDE'10, pages 995–1000, 2010.

[113] S. Rendle and C. Freudenthaler. Improving Pairwise Learning for Item Recommendation from Implicit Feedback. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 273–282, 2014.

[114] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, 2009.

[115] A. Ritter, C. Cherry, and B. Dolan. Unsupervised Modeling of Twitter Conversations. In *Proceedings of The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '10, pages 172–180, 2010.

[116] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven Response Generation in Social Media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 583–593, 2011.

[117] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Ima-

geNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[118] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, and P. A. Mitkas. Visual Event Summarization on Social Media Using Topic Modelling and Graph-based Ranking Algorithms. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 203–210, 2015.

[119] J. Schwarcz. *Ways of the Illustrator. Visual Communication in Children's Literature.* American Library Association, 1982.

[120] H. Seo. Visual Propaganda in the Age of Social Media: An Empirical Analysis of Twitter Images During the 2012 Israeli-Hamas Conflict. *Visual Communication Quarterly*, 21(3):150–161, 2014.

[121] B. P. Sharifi, D. . Inouye, and J. K. Kalita. Summarization of Twitter Microblogs. *The Computer Journal*, 57(3), 2014.

[122] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized Multiview analysis: A Discriminative Latent Space. In *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'12, pages 2160–2167, 2012.

[123] L. Sipe. Revisiting the Relationships Between Text and Pictures. *Children's Literature in Education*, 43:4–21, 2012.

[124] E. Spyrou and P. Mylonas. A Survey of Geo-tagged Multimedia Content Analysis within Flickr. In *Artificial Intelligence Applications and Innovations*, pages 126–135. Springer Berlin Heidelberg, 2014.

[125] M. Stefanone, G. Saxton, M. Egnoto, W. Wei, and Y. Fu. Image Attributes and Diffusion via Twitter: The Case of #guncontrol. In *2015 48th Hawaii International Conference on System Sciences*, HICSS '15, pages 1788–1797, Jan 2015.

[126] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In

*Proceedings of IEEE Computer Vision and Pattern Recognition*, CVPR '15, 2015.

[127] M. Thelwall, O. Goriunova, F. Vis, S. Faulkner, A. Burns, J. Aulich, A. Mas-Bleda, E. Stuart, and F. D'Orazio. Chatting through Pictures? A Classification of Images Tweeted in one Week in the UK and USA. *Journal of the Association for Information Science and Technology*, 2015.

[128] I. Uysal and W. B. Croft. User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2261–2264, 2011.

[129] P. Valdez and A. Mehrabian. Effects of Color on Emotions. *Journal of Experimental Psychology: General*, 123(4):394–409, 1994.

[130] J. van de Weijer, C. Schmid, and J. Verbeek. Learning Color Names from Real-World Images. In *Proceedings of the 20th IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '07, pages 1–8, 2007.

[131] J. van der Molen. Assessing Text-picture Correspondence in Television News: The Development of a New Coding Scheme. *Journal of Broadcasting and Electronic Media*, 45(3):483–498, SUM 2001.

[132] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized Multi-Modal Topic Model. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, UAI '12, pages 843–851, 2012.

[133] A. Wang, T. Chen, and M.-Y. Kan. Re-tweeting from a Linguistic Perspective. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 46–55, 2012.

[134] A. Wang and M.-Y. Kan. Mining Informal Language from Chinese Microtext: Joint Word Recognition and Segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 731–734, 2013.

[135] A. Wang, M.-Y. Kan, D. Andrade, T. Onishi, and K. Ishikawa. Chinese Informal Word Normalization: an Experimental Study. In *Proceedings of*

*International Joint Conference on Natural Language Processing*, IJCNLP '13, pages 127–135, 2013.

[136] K. Wang, C. Thrasher, E. Viegas, X. Li, and B.-j. P. Hsu. An Overview of Microsoft Web N-gram Corpus and Applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, HLT-DEMO '10, pages 45–48, 2010.

[137] M. Wang, D. Cao, L. Li, S. Li, and R. Ji. Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, ICIMCS '14, pages 76–80, 2014.

[138] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang. PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications. In *Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management*, AAIM '09, pages 301–314, 2009.

[139] Z. Wang, P. Cui, L. Xie, H. Chen, W. Zhu, and S. Yang. Analyzing Social Media via Event Facets. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1359–1360, 2012.

[140] Z. Wang, P. Cui, L. Xie, W. Zhu, Y. Rui, and S. Yang. Bilateral Correspondence Model for Words-and-Pictures Association in Multimedia-Rich Microblogs. *ACM Transaction on Multimedia Computing, Communications and Applications*, 10(4):34:1–34:21, July 2014.

[141] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li. Visual Sentiment Prediction with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1411.5731*, 2014.

[142] O. Yakhnenko and V. Honavar. Multi-Modal Hierarchical Dirichlet Process Model for Predicting Image Annotation and Image-Object Label Correspondence. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, SDM '09, pages 283–293, 2009.

[143] R. Yan, M. Lapata, and X. Li. Tweet Recommendation with Graph Co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for*

*Computational Linguistics*, ACL '12, pages 516–525, 2012.

[144] K. Yanai and Y. Kawano. Twitter Food Photo Mining and Analysis for One Hundred Kinds of Foods. In *Proceedings of the 15th Pacific-Rim Conference on Advances in Multimedia Information Processing*, PCM '14, pages 22–32, 2014.

[145] K. Yanai and Y. Kawano. Food Image Recognition using Deep Convolutional Network with Pre-training and Fine-tuning. In *2015 IEEE International Conference on Multimedia Expo Workshops*, ICMEW '15, pages 1–6, June 2015.

[146] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic Detection of Rumor on Sina Weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 13:1–13:7. ACM, 2012.

[147] Y. Yang, P. Cui, Z. Vicky, W. Zhu, and S. Yang. Emotionally Representative Image Discovery for Social Events. In *Proceedings of the 4th ACM International Conference on Multimedia Retrieval*, ICMR '14, 2014.

[148] Y. Yang, P. Cui, W. Zhu, and S. Yang. User Interest and Social Influence Based Emotion Prediction for Individuals. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM'13, pages 785–788, 2013.

[149] Q. You, J. Luo, H. Jin, and J. Yang. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1071–1074, New York, NY, USA, 2015. ACM.

[150] Q. You, J. Luo, H. Jin, and J. Yang. Robust Image Sentiment Analysis using Progressively Trained and Domain Transferred Deep Networks. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI '15, pages 381–388, 2015.

[151] L. Yu, S. Asur, and B. Huberman. Artificial inflation: The real story of trends and trend-setters in sina weibo. In *2012 International Confernece on Social Computing (SocialCom) and 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT)*, pages 514–519, 2012.

[152] L. Yu, S. Asur, and B. A. Huberman. What Trends in Chinese Social Media. In *Proceedings of the 5th SNA-KDD Workshop*, SNA-KDD'11, 2011.

[153] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: Image Sentiment Analysis from a Mid-level Perspective. In *Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 10:1–10:8, 2013.

[154] J. Yuan, Q. You, and J. Luo. Sentiment Analysis Using Social Multimedia. In A. K. Baughman, J. Gao, J.-Y. Pan, and V. A. Petrushin, editors, *Multimedia Data Mining and Analytics*, pages 31–59. Springer International Publishing, 2015.

[155] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern. Predicting Information Spreading in Twitter. In *Computational Social Science and the Wisdom of Crowds Workshop*, 2010.

[156] H. Zhang, G. Kim, and E. P. Xing. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1425–1434, 2015.

[157] L. Zhang and I. Pentina. Motivations and Usage Patterns of Weibo. *Cyberpsychology, Behavior, and Social Networking*, 15(6):312–317, June 2012.

[158] N. Zhao, R. Hong, M. Wang, X. Hu, and T.-S. Chua. Searching for Recent Celebrity Images in Microblog Platform. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 841–844, 2014.

[159] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring Principles-of-Art Features For Image Emotion Recognition. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 47–56, 2014.

[160] S. Zhao, H. Yao, Y. Yang, and Y. Zhang. Affective Image Retrieval via Multi-Graph Learning. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 1025–1028, 2014.

[161] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and Traditional Media Using Topic Models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR '11, pages 338–349, 2011.

[162] X. Zhao, F. Zhu, W. Qian, and A. Zhou. Impact of Multimedia in Sina Weibo: Popularity and Life Span. In *Joint Conference of 6th Chinese Semantic Web Symposium (CSWS'12) and the First Chinese Web Science Conference (CWSC'12)*, 2012.

[163] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock. A Text-to-picture Synthesis System for Augmenting Communication. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, AAAI'07, 2007.