Understanding and Classifying Image Tweets*

Tao Chen[1] Dongyuan Lu[1] Min-Yen Kan[1,2] Peng Cui[3]
[1]Web IR / NLP Group (WING), National University of Singapore
[2]NUS Interactive and Digital Media Institute
[3]Beijing Key Laboratory of Networked Multimedia, Tsinghua University
{taochen, ludy, kanmy}@comp.nus.edu.sg cuip@tsinghua.edu.cn

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.

We study a large corpus of image tweets in order to uncover what people post about and the correlation between the tweet's image and its text. We show that an important functional distinction is between visually-relevant and visually-irrelevant tweets, and that we can successfully build an automated classifier utilizing text, image and social context features to distinguish these two classes, obtaining a macro F_1 of 70.5%.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

image tweets; analysis; microblog; classification

1. INTRODUCTION

With improved bandwidth and camera phones, 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 China's Sina *Weibo* and later *Twitter* and third-party services such as *Instagram*, microblogging platforms now seamlessly include images into their posts.

There have been preliminary works that describe image tweets: as early as 2011, Yu *et al.* [10] reported that 56.43% of all microblog posts in Sina Weibo were image tweets. Zhao *et al.* [11] discovered that such image tweets were retweeted more often and survived longer than text-only posts. We may thus posit that users compose image tweets to attract and retain the interest of their readership.

While helpful, these findings only just start to answer the many questions about this new class of social communication. What types of images do users embed? Do the images distinctly differ from images on image- and photo-sharing websites (e.g., *Flickr*)? Do the textual contents of image tweets differ from posts that are text-only?

To answer these questions, we have collected a corpus of image tweets from Sina Weibo. Our contributions are in: 1) deconstructing the corpus to characterize such tweets' image and textual content and the correlation between the two; 2) collecting annotations for a subset of these image tweets in the corpus and; 3) building an automated classifier to distinguish two important subclasses of image tweets – *visual* and *non-visual* tweets.

2. WHAT ARE IMAGE TWEETS?

To answer these questions, we collected a corpus of tweets comprising of both image and text-only tweets. Over a period of 7 months in 2012, we sampled posts from the public timeline API of Weibo, accumulating a dataset of 57,595,852 tweets. To analyze the tweets in more depth, we further manually annotated a small ∼5K subset of the corpus[1].

Image Characteristics. To date, the images in image tweets have been studied in only a few works. Ishiguro *et al.* [4] show that predicting the number of views of an image tweet using social curation evidence (i.e. favoriting and explicit listing) outperforms using image features. Wang *et al.* [9] designed a joint text and image topic model to detect the onset of new events from image tweets.

An idiosyncratic factor is that all embedded images in Weibo are processed by the Weibo uploader which imposes certain restrictions and post-processing: 1) only one image is allowed per post[2]; 2) images are scrubbed of their EXIF metadata; and 3) all images (excluding animated GIFs) are converted to JPEG.

*This research is supported by the National Natural Science Foundation of China, No. 61003097; International Science and Technology Cooperation Program of China, No. 2013DFG12870, and by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502203 .

[1]Annotated corpus available at: `http://wing.comp.nus.edu.sg/downloads/imagetweets/`

[2]Note that since April 24 2013, Weibo has supported multiple images per post and displayed them in an album style.

In our corpus, 45.1% are image tweets, of which still images dominate: 97.5% image tweets contain a JPEG formatted picture while 2.5% contain an animated GIF. Figure 1 gives 18 examples which illustrate the variety of images in image tweets: there are photographs of varying quality (both candid and composed) and of varying topics, screenshots, cartoons, digital wallpaper and other forms of decorative images. Our manual inspection of our annotated corpus finds 69.5% of the images are natural photographs (including digitally edited ones), 13.2% are synthetic, and the remaining 17.4% are multi-photo collages. The collage form bypasses Weibo's one-photo-per-post limitation and are used for different narrative purposes: e.g., to compare objects, and tell stories through an image sequence.

With respect to posting habits, a survey of our annotators (who are users of Weibo) reveal that 85% self-reported that they use their camera phone as their photo-taking device, whereas 13.7% used a digital camera. This accords with our hypothesis that Weibo users care more about the photo content than quality, as most photos seem to be of low quality, which differs with Flickr [3].

Image-Tweets versus Text Posts. We attempt to uncover differences between image tweets and text-only ones by answering "when", "what" and "why". Plotting our tweets by posting hour (in Figure 2 (a)), we observe that image tweets are posted more frequently during the daytime. We posit that there are more tweet-worthy objects and events during the day, but we have yet to validate this.

For "what", we applied latent Dirichlet allocation (LDA; [1]) to a large, ∼1M subset of the whole dataset, to learn $k = 50$ latent topics, where k was tuned on a held-out set. Among these 50 topics, we observe that some exhibit an image- to non-image tweet ratio differing significantly from the average 45.1%. Figure 2b lists sample topics with manually-assigned labels: advertisements and posts on fashion, travel and food are adorned with images, while posts about emotions and everyday routine are mostly text-only.

As to "why", many studies on text-only tweets have already been done. The motivation in posting can be summarized as either societal (daily chatter, conversations) or informational (sharing information, reporting news) in nature [5]. From the distribution of LDA topics, we see the preference of posting image tweets or text tweets is correlated with the content. For example, advertisement tweets tend to include a product image to make it more informative; whereas tweets about the everyday routine – tweets whose topics are about work or sleeping, for example – are prone to be text-only. This answers "why" from a collective standpoint, so next we drill down to investigate the relation of the image and text in individual tweets.

3. IMAGE AND TEXT RELATION

Though users can post an image without accompanying text, it is rare – 99.1% of our image tweets have corresponding text. We want to know why people post both image and text and the nature of their correlation.

Two previous studies have attempted to answer this for the general domain. Marsh and White [6] identified 49 relationships, grouping them into three major categories based on the closeness of the relationship. Martinec and Salway [7] studied text-image relations from two perspectives: status (in terms of relative importance) and logico-semantic (one expands or repeats the same meaning of another). While

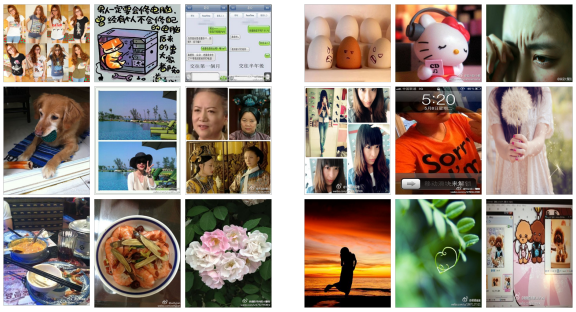
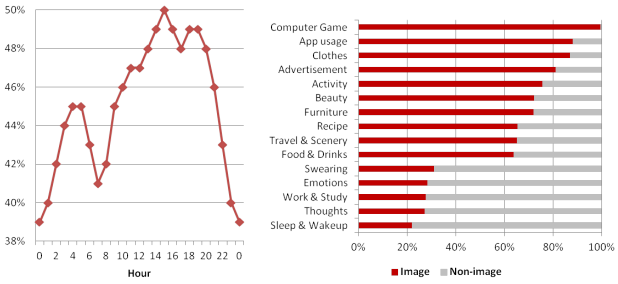


Figure 1: 18 Example Images from Image Tweets.



(a) (b)

Figure 2: (a) % of image tweets by hour. (b) % of image to non-image tweets in skewed topics.

insightful, these categorizations 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.

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 1, the left group of 9 images are from visual image tweets and the right group of 9 are non-visual image tweets. We find the distinction hinges jointly on text and image content together. Figure 3 shows two sample visual (top) and two sample non-visual (bottom) tweets. The motivations for posting images in a non-visual tweet 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 mosquitoes). In such cases, the text is the primary medium; the image reinforces the emotional aspects of the text, similar to emoticon use. However, we notice that the distinction

Figure 3: Image tweets with their corresponding text, image and translation. The top two are examples of *visual* tweets, and the bottom two are *non-visual* ones.

between emotional and general non-visual tweets 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, although we feel it is interesting.

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. E.g., the image in the first row of Figure 3 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.

4. VISUAL/NON-VISUAL CLASSIFICATION

We now turn to the task of making this 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.

Dataset Construction. To obtain gold standard annotations, we employed subjects from *Zhubajie*[3], as well as students at our university, to label a random subset of the image tweets. 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 3 different subjects, with the simple majority fixing the gold standard. In total, we collected annotations for 4,811 image tweets (also used in our manual analysis in Section 2) annotated by 72 different subjects. These broke down into 3,206 (66.6%) visual and 1,605 (33.4%) non-visual image tweets[4].

To utilize supervised machine learning, we employ multimedia features that leverage the text, image and social context of an image tweet.

[3]`http://www.zhubajie.com`, a crowdsourcing website.

[4]To enable future work, we further asked subjects to distinguish emotional from other non-visual tweets. However, inter-annotator agreement was not as strong ($\kappa = 0.54$), so we do not discuss this further.

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 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 as another feature. We also encode *POS density* features (proportion of nouns, verbs, adjectives, adverbs and pronoun within a tweet), as well as the presence of different classes of *named entities*. These features are useful as visual tweets mention concrete objects, people's names (E.g, the second row in Figure 3), places and products. We also use four *microblog-specific* features that test for the presence of @mentions, #hashtags, geolocation coordinates and URLs.

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 are detected in 22.2% of images.

Images with similar content tend to exhibit the same image-text relation. To capture this, we cluster the images by visual similarity. Following the bag-of-visual-words methodology, we first extract SIFT descriptors from the images as inputs, clustering them to form visual words by building a hierarchical visual vocabulary tree [8]. We then apply LDA to the corpus of images-as-documents' visual vocabulary, aiming to learn k hidden topics[5]. Subsequently, the *image topic* assignment is encoded as a single feature.

Context Features. From our earlier analysis, we know that the posting time affects the probability of a tweet being visual or not; for this reason, we include the hour of the *posting time* as a feature. As people share what they have just seen (visual tweet), we capture whether the *device* used to post the image tweets is mobile or not (e.g. desktops).

Social features round out our set. Weibo readers can post comments on the post. In our dataset, 46.3% of the image tweets have at least one comment, and 21.5% have been reposted; as opposed to 33.7% and 16.2% for text-only tweets. 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.

5. EXPERIMENT

We performed 10-fold cross validation experiments with the Naïve Bayes[6] implementation in Weka3 [2]. The three sets of features were linearly concatenated into a single vector. Due to the imbalanced distribution (66.6% of image

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

[6]Experimenting with other learners (e.g., SVM, Logistic Regression) yielded worse results.

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 0.40.

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 1. *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 words 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 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 do not. Our final classifier (Row 14) that combines all features that improved the baseline, achieves an F_1 of 70.5.

Table 1: Experimental Results and Feature Analysis.

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

We further analyzed the misclassified instances. While *words* are the most discriminative feature, microblog text is relatively short (tweets on Weibo are limited to 140 characters). 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. 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 in 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.

6. CONCLUSION

The social Web 2.0 has embraced multimedia with the inclusion of facilities to embed images in microblog posts. We performed a multipronged analysis of these *image tweets* from visual, textual and social context perspectives. We discover that images from image tweets demonstrate a wider spectrum of image types as compared with image-sharing websites, and that the tweets differ from text-only ones in terms of their topical content and posting time.

We make 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 the annotated corpus available to the public to test and benchmark against.

We have further identified that non-visual tweets are often associated emotionally, and will classify such emotionally relevant image tweets in future work. We also plan to conduct similar study in other platforms (e.g., Twitter), and compare the findings with Weibo.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[3] R. Herrema. Flickr, communities of practice and the boundaries of identity: a musician goes visual. *Visual Studies*, 26(2):135–141, 2011.

[4] K. Ishiguro, A. Kimura, and K. Takeuchi. Towards automatic image understanding and mining via social curation. In *ICDM*, 2012.

[5] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD*, 2007.

[6] E. E. Marsh and M. D. White. A taxonomy of relationships between images and text. *Journal of Documentation*, 59:647–672, 2003.

[7] R. Martinec and A. Salway. A system for image–text relations in new (and old) media. *Visual Communication*, 4(3):337–371, 2005.

[8] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[9] Z. Wang, P. Cui, L. Xie, H. Chen, W. Zhu, and S. Yang. Analyzing social media via event facets. In *MM*, 2012.

[10] L. Yu, S. Asur, and B. A. Huberman. What Trends in Chinese Social Media. In *SNA-KDD*, 2011.

[11] X. Zhao, F. Zhu, W. Qian, and A. Zhou. Impact of Multimedia in Sina Weibo: Popularity and Life Span. In *CSWS and CWSC*, 2012.