VELDA: Relating an Image Tweet's Text and Images

Tao Chen,[1] **Hany M. SalahEldeen,**[2] **Xiangnan He** [1] **Min-Yen Kan,**[1,3] **Dongyuan Lu**[1]
[1]School of Computing, National University of Singapore
[2]Department of Computer Science, Old Dominion University
[3]NUS Interactive and Digital Media Institute, Singapore
{taochen, xiangnan, kanmy, ludy}@comp.nus.edu.sg hany@cs.odu.edu

Abstract

Image tweets are becoming a prevalent form of social media, but little is known about their content – textual and visual – and the relationship between the two mediums. Our analysis of image tweets shows that while visual elements certainly play a large role in image-text relationships, other factors such as emotional elements, also factor into the relationship. We develop Visual-Emotional LDA (VELDA), a novel topic model to capture the image-text correlation from multiple perspectives (namely, visual and emotional).

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.

1 Introduction

Smartphones with cameras have morphed traditional text-only user generated content into multimedia. *Image tweets* – microblog posts which embed images – enable a viewer to see the world through someone else's eyes. Their multimedia form attracts larger viewership and prolongs their half-life as compared to their poorer cousins – text only posts – in Sina *Weibo* (Zhao et al. 2012), the dominant microblog platform in China, and have been found to be 35% more retweetable than text-only tweets in Twitter[1].

These posts are fast becoming the *de facto* standard on such microblog platforms. They constitute over 45% of overall traffic in Weibo (Chen et al. 2013) and have seen rapid adoption in Twitter. How about image-only tweets? While a picture may be worth a thousand words, image-only posts are still rare: over 99% of tweets with images are also accompanied by text (Chen et al. 2013).

A natural set of questions emerge that our work attempts to address. Why do people post image tweets? What is the nature of the relationship between the text and image? And, given the answers to these questions, can we design a model that explains how image tweets could be generated?

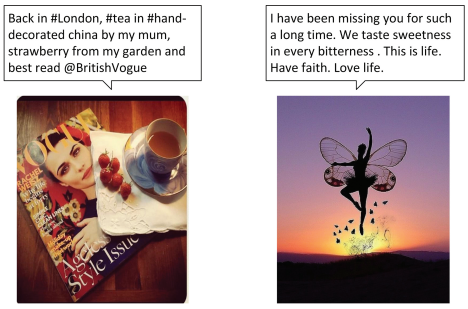


Figure 1: A visually relevant image tweet from Twitter (left) and an emotionally relevant image tweet from Weibo (right).

Through a corpus analysis, Chen et al. (2013) identified image and text in the microblog can be either visually or emotionally correlated. To validate this, we surveyed image tweets authors, and discovered images are used for visual purpose (*i.e.*, both image and text mention the same physical object) as well as for enhancing the emotion of the post (see Figure 1 for both examples). However, current multi-modal models only capture a single factor in modeling image-text correspondence.

Our paper'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 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 existed methods. Finally, we apply VELDA in a real-world task of automated microblog illustration, selecting a relevant image from an image collection.

2 Related Work

The duality of image and text has been a recurring topic of study in the multimedia area. Uncovering and model-

[1]https://blog.twitter.com/2014/what-fuels-a-tweets-engagement

ing the relationship between the two mediums has been a key area of study. One method is to map the multimodal data into a shared space such that the distance between two similar objects is minimized. Under this approach, Canonical Correlation Analysis (CCA) (Hotelling 1936) and its extensions are often utilized (Rasiwasia et al. 2010; Sharma et al. 2012). CCA finds a pair of linear transformations to maximize the correlations between two variables (*i.e.*, image and text), jointly reducing the dimensionality of the two heterogeneous representations of the same data.

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) (Blei, Ng, and Jordan 2003), a generative model to discover underlying topics that generate the documents and the topic distribution within each document. The seminal work of Barnard et al. (2003) proposed multi-modal LDA 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 and Jordan (2003) proposed correspondence LDA (hereafter, Corr-LDA; Figure 2) 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, Attias, and Nagarajan (2010) proposed a topic-regression multi-modal LDA 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, and in some datasets, some documents may lack images or text. To address these shortcomings, Jia, Salzmann, and Darrell (2011) 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 (Yakhnenko and Honavar 2009; Virtanen et al. 2012; Liao, Zhu, and Qin 2014), which alleviates the need to choose the number of topics *a priori*.

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 (*e.g.*, visual) model. Our proposed method extends LDA to cater for this key characteristic in the generative process.

3 Visual-Emotional LDA

To understand how image tweets are generated, we first turn to their authors. Prior work by Chen et al. (2013) identified three image-text correlations, namely, visually relevant, emotionally relevant and irrelevant (*e.g.*, noisy image tweets). To test whether their findings are corroborated by actual users (their study was limited to corpus analysis), we recruited 109 *Weibo* users (62 females and 47 males) from the popular Chinese crowdsourcing site *Zhubajie*[2]. Respondents were asked to fill out a questionnaire on their image posting behavior. In the question of "Why do you embed an image in a tweet?", 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 a prominent image–text relation in microblog posts*[3].

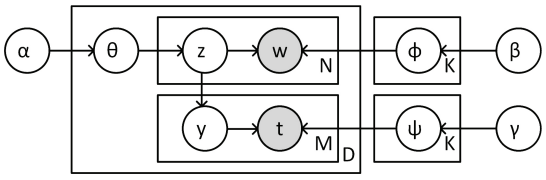


Figure 2: Correspondence LDA (Corr-LDA), where the N plate specifies visual words and the M plate specifies individual words in the text. Y, the topic assignments of textual words, are conditioned on Z, the topic assignments of N visual words.

To achieve high fidelity modeling of the image-text relationship in image tweets, we must account for multiple channels. We propose Visual-Emotional LDA (VELDA) as a generative model that incorporates the suggested emotional aspect in modeling image tweets. In the following, we first detail VELDA's model formulation, and then describe its parameter estimation process.

3.1 Model Formulation

Figure 3 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 topic modeling methods, we model the three modalities as discrete features, which are referred as textual words, visual words, and emotional words, respectively (the feature extraction process is detailed later in Section 4.2).

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 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.

[2]http://www.zhubajie.com

[3]While the survey was limited to Chinese *Weibo* users, we believe this correlation also holds for other microblog sites and cultures.

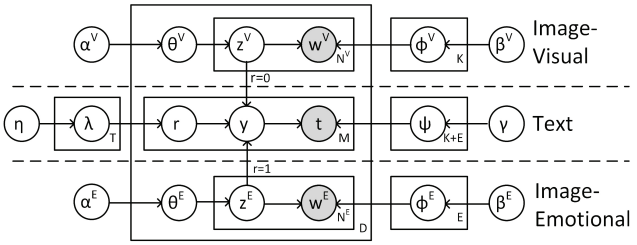


Figure 3: Visual-Emotional LDA's generative model.

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 λ. The overall generative story is summarized as follows, where specific notations are explained in Table 1[4]:

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.
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 $t_m \sim \text{Mult}(\psi_{e=y_m})$

3.2 Parameter Estimation

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

We adopt Gibbs sampling to estimate the model parameters, due to its simplicity in deriving update rules and effectiveness in dealing with high-dimensional data. The basic

[4]We use the abbreviation Dir(·), Mult(·) and Unif(·) to denote the Dirichlet, Multinomial and Uniform distribution, respectively.

Table 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.
α^V, α^E, β^V, β^E, γ, η	hyperparameters of Dirichlet distributions.
θ^V, θ^E	$D \times K$, $D \times E$ matrices indicating image-visual, image-emotional topic distribution, respectively
ϕ^V, ϕ^E	$K \times C$, $E \times S$ matrices indicating image-visual, image-emotional topic-word distribution, respectively
ψ	a $(K + E) \times T$ matrix indicating textual topic-word distribution.
λ	a $T \times 2$ matrix indicating textual word's relevance distribution.
$M_{d,t}$, $N_{d,c}^V, N_{d,s}^E$	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.
$N_{d,k}^V, N_{d,e}^E$	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.

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 θ^V, we need to sequentially sample its latent topic variable z^V. To sample for z_i^V (where $i = (d, n)$ representing the n-th word of the d-th document), we condition on the current value of all other variables.

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

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

$$P(z_i^E = e|W^V, W^E, T, Z^V, Z_{-i}^E, Y, R) \propto \frac{N_{e,s,-i}^E + \beta_s^E}{N_k^E + S\beta^E - 1} \cdot (\frac{N_{d,e}^E}{N_{d,e,-i}^E})^{M_{d,e}} \cdot \frac{N_{d,e,-i}^E + \alpha_e^E}{N_d^E + E\alpha^E - 1}. \quad (2)$$

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 that states whether it is sampled from image–visual topics or image–emotional topics. If y_i is sampled from image–visual topics, it implies that the sampled r_i was 0, and vice versa. As such, we need to sample based on the joint distribution of y_i and r_i, which leads to:

$$\begin{aligned} &P(r_i = 0, y_i = k|W^V, W^E, T, Z^V, Z^E, Y_{-i}, R_{-i}) \\ &\propto \frac{M_{k,t,-i} + \gamma}{M_k + T\gamma - 1} \cdot \frac{M_{t,r=0,-i} + \eta}{M_t + 2\eta - 1} \cdot \frac{N_{d,k}^V}{N_d^V}, \\ &P(r_i = 1, y_i = e|W^V, W^E, T, Z^V, Z^E, Y_{-i}, R_{-i}) \\ &\propto \frac{M_{e,t,-i} + \gamma}{M_e + T\gamma - 1} \cdot \frac{M_{t,r=1,-i} + \eta}{M_t + 2\eta - 1} \cdot \frac{N_{d,e}^E}{N_d^E}. \end{aligned} \quad (3)$$

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:

$$\theta^V_{k,d} = \frac{N^V_{k,d} + \alpha^V}{N^V_d + K\alpha^V}, \quad \theta^E_{e,d} = \frac{N^E_{e,d} + \alpha^E}{N^E_d + E\alpha^E}, \quad \phi^V_{k,c} = \frac{N^V_{k,c} + \beta^V}{N^V_k + C\beta^V},$$
$$\phi^E_{e,s} = \frac{N^E_{e,s} + \beta^E}{N^E_e + S\beta^E}, \quad \psi_{z,t} = \frac{M_{z,t} + \gamma}{M_z + T\gamma}, \quad \lambda_{r,t} = \frac{M_{r,t} + \eta}{M_t + 2\eta}. \tag{4}$$

3.3 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. VELDA's calculation is also compatible with distributed computation strategies for Gibbs sampling (Wang et al. 2009), making VELDA applicable to large-scale data.

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 r. 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. The derivations of the existing image parts of the model are unchanged, just incurring additional updating rules for any new factors.

4 Evaluation

We evaluated 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 investigated how VELDA fares in modeling other general domain image–text pairs. To this end, we collected image tweets from two microblog platforms — Weibo and Twitter — and image-text pairs from Google and Wikipedia. In the following, we describe the collected datasets, our feature extraction process, the evaluation criteria, and conclude by discussing the experiments and their results.

4.1 Datasets

We collected five image–text datasets (see Table 2). The first four have a common basis for collection – constructed by a list of queries, so we describe this basis first. Previous work by Chen et al. (2013) released a collection of 4K image tweets curated from Weibo with human image–text relation annotations following their 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 their work, 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, we obtain 353 words from visually relevant tweets (*e.g.*, bread, sunset), 133 words from emotionally relevant tweets (*e.g.*, worry, love), and use each single word as a query.

Table 2: Demographics of the five datasets.

	Weibo	**Twitter**	**G-Zh**	**G-En**	**POTD**
Size	22,782	16,427	38,806	26,903	2,524
Text language	Chinese	English	Chinese	English	English
Best settings for VELDA	K=100, E=60	K=40, E=80	K=80, E=50	K=100, E=100	K=40, E=80

1. **Weibo**. We sent each query as a hashtag in Weibo's 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 kept those with 40 to 100 results, which resulted in a set of 22,782 image tweets.
2. **Twitter**. Following the same pipeline, we constructed 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. Our spot checks show the translated words are acceptable.

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 engine. We sent the Chinese and English (translated) text queries to Google Image Search. We obtained 38,806 and 26,903 images and their associated text snippet 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 tweets.

5. **Wikipedia POTD**. At the high end of the quality spectrum is the "Picture of the Day" (POTD) collection, a set of of daily featured pictures accompanied by a short description from Wikipedia[5]. Unlike the other four datasets, POTD concentrates on high-quality, manually-curated academic topics. We collected the daily pictures and their corresponding descriptions from 1 Nov 2004 to 11 Jun 2014, obtaining a total of 2,524 image-text pairs.

4.2 Feature Extraction

We extract textual words from image's textual description, and another two sets of features from the image to represent its visual semantics and emotional semantics, respectively. We adopt current best practices to generate features for each modality. 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

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

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.

Visual Features. We adopt the standard Scale-Invariant Feature Transform (SIFT; Lowe 2004) descriptors to represent the visual semantics of an image and follow 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 datasets based on the common basis. Each codebook thus consists of 1000 visual words.

Emotional Features. 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 (Valdez and Mehrabian 1994; Colombo, Del Bimbo, and Pala 1999; Machajdik and Hanbury 2010; Jia et al. 2012; Yang et al. 2013). We adopt 22 color-based features from the state-of-art work (Machajdik and Hanbury 2010), summarized in Table 3. To turn an image into a bag of emotional words, we first segment each image into patches by a graph-based algorithm (Felzenszwalb and Huttenlocher 2004), 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 3: Features used to represent image emotions.

Name	Dim.	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 (van de Weijer, Schmid, and Verbeek 2007).
Pleasure, Arousal, Dominance	3	One set of affective coordinates, calculated from brightness and saturation, following (Valdez and Mehrabian 1994).

4.3 Experimental Settings

A good model of image–text relations should be able to help generate one given the other. 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. Specifically, given a textual query T = $t_1, ..., t_N$, VELDA computes a relevance score for an image *i* by the following formula:

$$\begin{aligned} score_i &= P(T|\theta_i^V, \theta_i^E) = \prod_{n=1}^{N} P(t_n|\theta_i^V, \theta_i^E) \\ &= \prod_{n=1}^{N} (\lambda_{t_n,0} \sum_{k=1}^{K} \psi_{k,t_n} \theta_{i,k}^V + \lambda_{t_n,1} \sum_{e=1}^{E} \psi_{e,t_n} \theta_{i,e}^E) \end{aligned} \quad (5)$$

where θ_i^V and θ_i^E are the visual and emotional topic distribution for a test image *i*, and λ is the textual word's relevance distribution learned during training.

As there is only one ground-truth match for each textual query (*i.e.*, the original image accompanying the post), we consider the position of the ground-truth image in the ranked list as an evaluation metric. Following (Jia, Salzmann, and Darrell 2011), we consider an image as correctly retrieved if it appears in the top t percent of the image test collection created from the text in the corresponding post.

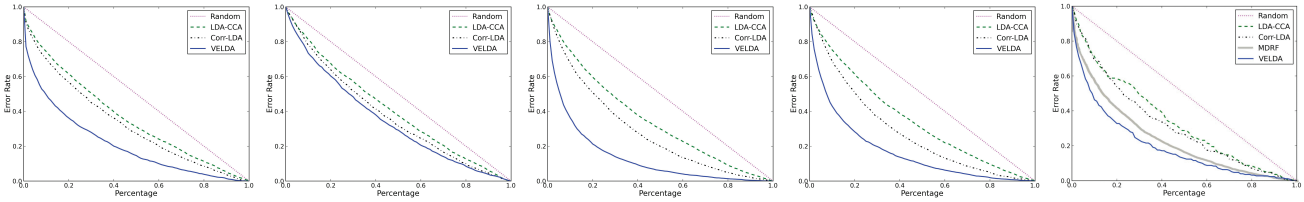
We compare VELDA against a **random** baseline and two state-of-the-art image–text correlation algorithms – **Corr-LDA** (Blei and Jordan 2003) and **LDA-CCA** (Rasiwasia et al. 2010). Corr-LDA computes the score of an image given a text query similar to Eq. 5. In LDA-CCA, two standard LDA models are 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 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: α^V=1, α^E=1, β^V=0.1, β^E=0.1, γ = 0.1, and η = 0.5. We then tune the number of visual topics (K) and emotional topics (E) in a grid search for each dataset (see Table 2 for the detailed settings). We similarly optimize the Corr-LDA and LDA-CCA baselines by searching for their best parameter settings.

4.4 Results and Analysis

Results on the cross-modal image retrieval tasks on the five datasets are shown in Figure 4. Each plot depicts the retrieval errors averaged over all testing queries in a specific dataset. For all five datasets, a one-tailed paired t-test with threshold 0.001 revealed that VELDA's performance gain is statistically significant. For POTD, we have additionally overlaid Jia et al.'s MDRF results, as taken from their paper. The MDRF results are not strictly comparable – our dataset is larger by 537 documents (approximately 20% larger) and we do not have their extracted features – but we feel that they are indicative of MDRF's performance, and further help to show VELDA's competitive performance.

For all graphs, better performance is equated with lower error rate (curves closer to the bottom left corner). From Figure 4, 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 in the very early positions of the ranked list (*e.g.*, the top 0.8% for Weibo). For concrete comparison, we focus on recall on the top 10% level, reported separately in Table 4. 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, respectively. For the POTD dataset,



(a) Weibo (b) Twitter (c) Google-Zh (d) Google-En (e) Wikipedia POTD

Figure 4: Retrieval error rate by the percentage of the ranked list considered. Curves closer to the axes represent better performance.

Table 4: 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 one-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: 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%

VELDA outperforms MDRF by 8%. 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 images that are bright and tranquil.

We note the lower performance of all methods on Twitter. We attribute this to the brevity of Twitter: each Twitter image tweet has only 6.7 textual words on average (after text processing), far shorter than the other datasets (*e.g.*, 18.9 for Weibo). This passes little textual information to the model, making the image-text correlation learning difficult. Even in such sparse data scenarios, VELDA still betters 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. 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 to tune is the hyperparameter η, which determines the relevance distribution (λ) of textual words, while other parameters can be easily set with standard LDA heuristic rules. So we further investigate the impact of η. Theoretically speaking, large (small) η makes the relevance distribution λ more skewed (balanced). From Figure 5, we see that VELDA's performance remains relatively stable for varying values of η. This insensitivity to η shows that VELDA is robust and does not require careful tuning to perform well.

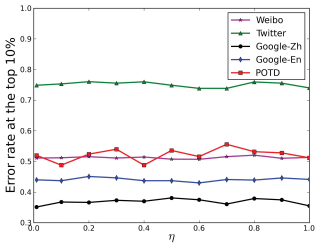


Figure 5: Parameter η versus error rate for top 10% retrieval.

[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 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 6: Three (translated) Weibo posts, along with VELDA's top 4 (from top left to bottom right) suggested illustrations.

Finally, in Figure 6, we show three typical Weibo posts (translated to English) and their top four recommended images by VELDA. As we can 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 varieties (*e.g.*, capturing three different nuts in the top four illustrations). For sentimental tweets, *e.g.*, the bottom example post, our recommended images match the emotions of the text well. This suggests a real possibility of applying VELDA to an automated microblog illustration task.

5 Conclusion

Image tweets match text with image to help convey a unified message. We examine the image–text correlation and its modeling for cross-modality image retrieval, in both microblog posts as well as other image–text datasets.

We discover that an image tweet's image and text can be related in different modes, not limited to visual relevance but including emotional relevance. A key contribution is our development of Visual-Emotional LDA (VELDA), a novel topic model that captures such two image–text correlations. Experiments on both English and Chinese image tweets show that VELDA significantly outperforms baseline methods. VELDA also demonstrates its robustness and generalization, being applicable not only to its intended domain of image tweets but also general image–text datasets.

In the future, we may develop a mobile application to assist users discover and use visually and emotionally relevant images to adorn their textual posts. In a separate vein of work, we hope to validate VELDA as a generic model that can apply to other instances of multimedia correspondence. Moreover, the current VELDA regards image's visual and emotion as two independent views, it will be interesting to explore how these two are correlated with each other.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. We also would like to thank our colleagues, Kazunari Sugiyama, Jun-Ping Ng, Jovian Lin and Jingwen Bian, for the discussions.

References

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching Words and Pictures. *Journal of Machine Learning Research* 3:1107–1135.

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

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

Chen, T.; Lu, D.; Kan, M.-Y.; and Cui, P. 2013. Understanding and Classifying Image Tweets. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, 781–784.

Colombo, C.; Del Bimbo, A.; and Pala, P. 1999. Semantics in visual information retrieval. *IEEE MultiMedia* 6(3):38–53.

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

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

Jia, J.; Wu, S.; Wang, X.; Hu, P.; Cai, L.; and Tang, J. 2012. 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.

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

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

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

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

Putthividhy, D.; Attias, H.; and Nagarajan, S. 2010. 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, 3408–3415.

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

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

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

van de Weijer, J.; Schmid, C.; and Verbeek, J. 2007. Learning color names from real-world images. In *Proceedings of the 20th IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '07, 1–8.

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

Wang, Y.; Bai, H.; Stanton, M.; Chen, W.-Y.; and Chang, E. Y. 2009. 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, 301–314.

Yakhnenko, O., and Honavar, V. 2009. 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, 283–293.

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

Zhao, X.; Zhu, F.; Qian, W.; and Zhou, A. 2012. 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)*.