Inferring User Interests on Social Media From Text and Images

Yagmur Gizem Cinar
Computer Science
KU Leuven
Leuven 3001 Belgium
Email: yagmurgizem.cinar@student.kuleuven.be

Susana Zoghbi
Computer Science
KU Leuven
Leuven 3001 Belgium
Email: susana.zoghbi@cs.kuleuven.be

Marie-Francine Moens
Computer Science
KU Leuven
Leuven 3001 Belgium
Email: sien.moens@cs.kuleuven.be

***Abstract*—Inferring user interests on social media from text and images is addressed as a multi-class classification problem. We proposed approaches to infer user interest on Social media where often multi-modal data (text, image etc.) exists. We use user-generated data from Pinterest.com as a natural expression of users' interests. We consider each pin (image-text pair) as a category label that represents a broad user interest, since users collect images that they like on the social media platform and often assign a category label. This task is useful beyond Pinterest because most user-generated data on the Web is not necessarily readily categorized into interest labels. In addition to predicting users' interests, our main contribution is exploiting a multi-modal space composed of images and text. This is a natural approach since humans express their interests with a combination of modalities. Exploiting multi-modal spaces in this context has received little attention in the literature. We performed eleven experiments using the state-of-the-art image and textual representations, such as convolutional neural networks, word embeddings, and bags of visual and textual words. Our experimental results show that in fact jointly processing image and text increases the overall interest classification accuracy, when compared to uni-modal representations (i.e., using only text or using only images).**

Keywords*—*inferring user interests, user modeling, term frequencies, bag of words (BoW), convolutional neural networks (CNN), word embeddings.

I. Introduction

The goal of this study is to infer broad users' interests from user-contributed images, user-generated text and categories on social media. This task is useful in several fields, such as providing a personalized online experience or recommending relevant items to users. For instance, we may recommend sports products from Amazon to a user interested in sports, or furniture to a user interested in home decoration. In addition, it can be used to analyse upcoming trends such as likes on fashion, books, healthy organic food, etc, which in turn is useful for market analysis. Every day, thousands of people contribute to social media by uploading photos, videos, status updates, likes, comments, etc. Every action may reveal different users' aspects. In our study we leverage this information to infer users interest.

Specifically, we use pins from Pinterest.com as expressions of broad categories that users are interested in. Pinterest is a social media platform for sharing photos. Users usually pin or repin photos that they find interesting and add a textual comment to the corresponding photo. Pins are quite interesting. They provide real examples often labelled by the users and indicates the users' interests.

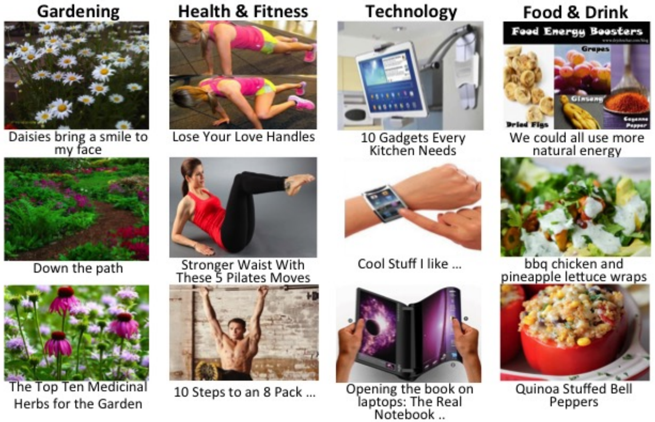
Gardening
Health & Fitness
Technology
Food & Drink
Food Energy Boosters
Grapes
Ginseng
Dried Figs
Cayenne Pepper
Daisies bring a smile to my face
Lose Your Love Handles
10 Gadgets Every Kitchen Needs
We could all use more natural energy
Down the path
Stronger Waist With These 5 Pilates Moves
Cool Stuff I like ...
bbq chicken and pineapple lettuce wraps
The Top Ten Medicinal Herbs for the Garden
10 Steps to an 8 Pack ...
Opening the book on laptops: The Real Notebook ..
Quinoa Stuffed Bell Peppers

Fig. 1: Examples of pins (image + text) on Pinterest and the corresponding category. We use the user-indicated category as an expression of the user's interest and try to predict this label. This is useful because not all pins are categorized.

Pinterest users often categorize the pins on their boards. However, not all pin boards are categorized. Figure 1 presents several pin examples and the corresponding user-indicated category. We use these categories as *interest labels* and our aim is to automatically predict them. We address this as a supervised multi-class classification problem.

This task is also useful beyond Pinterest. The Web is a multi-modal space, where users express their thoughts using different media (visual, textual, audio, etc). We can find billions of such examples on blog posts, Facebook updates, Flicker photos and comments, product reviews, etc. Our approach for interest prediction may be used on other sites, where users have not explicitly indicated the type of items that they are interested in. With a multi-class classification approach, we can learn the user interests for a given set of images with some captions. Hence, we can learn the user interest when there is no available category label such as Pinterest boards with no category label and other social media platforms such as Instagram, Facebook.

Given that users may express their interests both visually and through textual comments, we propose to exploit this multi-modal space to accomplish our task. The rationale is that only text or only images may not be sufficient to obtain a good representation. For instance, one post on Pinterest states that *"We could all use more natural energy"*. Natural energy may be interpreted as renewable energy. However, when combining this text with the top right image in Fig. 1, one can see that this energy refers to natural food products.

Thus, in addition to predicting users' interests on social media, our second goal is to compare the performance of uni-modal and multi-modal approaches. Our motivation is that joint image-text features would improve accuracy w.r.t. using a uni-modal space. To test this, we conduct eleven experiments with two different text representations: bag of words and word embeddings; and two image representations: bag of visual words, CNN as pre-trained network and fine tuned for Pinterest dataset (ft-CNN). Our experimental results show that indeed accuracy is improved by combining text and image features.

II. Related Work

Regarding input features, there exist many studies to infer user interests based exclusively on text data [1], [2], [3], [4] or exclusively on image data [5], [6], [7]. Unlike these previous studies, we focus on the rich multi-modal data from the social media platform, Pinterest.

There exist many different approaches to model user interests. For example, in [8], [6] clustering techniques are used, where each cluster is one type of interest. Similarly, k-nearest neighbors [9] and classification [10] methods might be used for representing interests. We use a multi-class classifier to classify user pins (image + text pairs) to different interest class labels, e.g. 'science nature', 'photography'. With multi-class classification approach, we can learn the distribution of interests of the users from their unlabeled shares (image + text) on different social media platforms.

We use two main frameworks to represent our data: bag of words and neural networks. Under the bag of words framework, text and images are represented by a histogram of terms from a dictionary [11]. Many studies have exploited the quantization of visual descriptors into visual words [12], and utilized the bag of words approach [13], [14], [15].

Under the neural network framework, text may be represented using word embeddings and images may be represented using Convolutional Neural Networks (CNN) [16]. A powerful word embedding approach was recently proposed by Mikolov et al. [17]. Their *word2vec* algorithm projects discrete words into a low-dimensional continuous vector space. These vectors have been shown to capture semantic and syntactic relations very well [17], [18]. Text represented as neural probabilistic vector of words based on n-gram models have also been studied in [19].

For image representations, CNN have achieved outstanding performance in the largest image classification and object detection tasks [16], [20]. They were first studied by Fukushima and LeCun et al. [21], [22] and are now used for other challenging tasks such as automatic image description generation [23].

III. Methodology

We approach the task of inferring a user's interest as a multi-class classification task on image and text pairs. In other words, given an image and corresponding caption, we want to assign a label from a set of interest labels. This set of labels corresponds to the Pinterest categories. In this section, we describe state-of-the-art representations for text and image data.

A. Text Representations

We study two kinds of text document representations: term frequencies and sentence vectors composed of word embeddings. The term frequency approach represents documents as a distribution of terms given the corpus vocabulary. It disregards the word order and grammar. This representation is often referred to as *bag of words* (BoW).

Word embeddings are given by the weights of a shallow neural network, which encode semantic relations, linguistic patterns and syntactic information of any given word based on its co-occurring words. The network learns to encode discrete words into a continuous semantic space, where similar words are placed close together. Word embeddings enable us to represent words in a much more compact way than the BoW approach. In our experiments, we have used pre-trained word vectors (also known as word2vec) on a Google News dataset. These pre-trained vectors are readily available and have become quite popular due to their outstanding performance in various semantic and syntactic tasks in the NLP community. They are 300-dimensional and were trained on 3 million words and phrases. For a full treatment of this representation, we refer the reader to [17]. To represent sentences/documents, we applied a simple summation of the word vectors that compose the text data. For a sentence $s = \{w_x, w_y, w_z\}$, we compute the sentence vector $\vec{s}$ as the vector sum, $\vec{s} = \vec{w}_x + \vec{w}_y + \vec{w}_z$.

B. Image Representations

In this study, three image representations are explored: Bag of Visual Words (BoVW) [12], [13], Convolutional Neural Networks (CNN) as pre-trained in the ImageNet classification challenge [20], and Convolutional Neural Networks fine-tuned to Pinterest categories (ft-CNN).

The BoVW approach starts off by computing the popular Scale-Invariant Feature Transform (SIFT) descriptors for each image. SIFT finds interest points and considers a grid of subregions around it. For each subregion it computes a gradient orientation histogram. The standard setting uses 4 by 4 subregions with 8-bin orientation histograms resulting in a 128-bin histogram. For an in-depth treatment the reader may refer to [24]. We then create visual words by quantizing the descriptors into a number of clusters. We use the k-means algorithm to cluster the descriptors around centroids. These centroids are sometimes called a visual codebook. Each descriptor in each image may then be assigned a visual word and each image may be thought to contain a set of visual words. Thus, the BoVW approach enables us to represent documents with visual term frequencies. We compute SIFT features densely across the image using the open source library VLFeat [25].

CNN is a type of feed-forward artificial neural network where the individual neurons are connected to respond to overlapping regions in the image [22]. They have been extremely successful in image recognition tasks in computer vision. The training process may be interpreted as learning a template given by a set of weights, which aim to maximize the probability of the correct class for each of the training instances. These networks may contain many layers. The weights corresponding to the fully connected layer of CNN are often used as image features. Thus, we used two approaches of CNN image features: first is fully connected layer weights of a pre-trained CNN in the ImageNet classification challenge, and second is the fully connected layer weights of CNN that fine-tuned to Pinterest categories. For full details of pre-trained network, we refer the reader to the study of Krizhevsky et al. [16]. We used the Caffe implementation of CNN [26]. For fine-tuned CNN to Pinterest categories, we mainly updated the (pre-trained network in the ImageNet challange) weights of the last layer and also changed the number of outputs to 32 (from 1000) which reflects the number of classes. The choices of hyper-parameters are similar to previous work [27].

The Bag of Words approach does not utilize the localization of the image features. Whereas, the CNN approach uses the localization of the features with the convolution steps.

C. Joint Text and Image Representation

To exploit the multi-modal nature of the data, we employ late fusion (decision fusion) of the image and text representations. For the i^{th} pin, the text feature (representation), fea_{t_i}, and image feature (representation), fea_{im_i}, are extracted separately as described in Section III-A and Section III-B, respectively and we combine their class predictions as a weighted sum as in Equation 1.

$$pred_{pin_i} = \lambda * pred_{im_i} + (1-\lambda) * pred_{t_i} \quad i \in \{1, ..., N\} \quad (1)$$

where λ value leverages between image and text class prediction decision. Higher λ value gives higher weight to classification prediction of image feature. Whereas, smaller λ gives higher weight to classification prediction of text features. $pred_{input\_data}$ is the likelihoods of the input data (image, text or image + text) for the 32 categories. We investigate three different λ values as [0.3, 0.5, 0.7] where λ = 0.5 gives equal weights to both predictions when image and text data are used.

D. Classification

A one versus rest (OVR) learning scheme was used to train a Support Vector Machine (SVM) for the multi-class classification problem. There are 32 predefined categories on Pinterest. Thus, 32 different classifiers were trained to address each class separately. Table I presents these categories. A Radial Basis Function Kernel was utilized for feature mapping and C and γ parameters were optimized by using grid search.

E. User Interest Representation

We represent users' interests as a frequency distribution of pins over Pinterest categories, which correspond to broad interests. Figure 2 presents some interest distributions under this representation. It clearly shows that each user has different category preferences with varying number of pins. For example, User 4 has about 60% of her pins dedicated to 'weddings', followed by about 15% dedicated to 'home decor', and the rest of her pins focus on 'women's fashion', 'food and drink', and 'design'. In this paper we focus on learning to correctly classify each pin. This can be translated into a distribution of interests by simply counting the number of pins on each category for each user.

F. Evaluation

We focus on assessing how well we can predict the category of each pin in an unseen test set. To this end, we compute classification accuracy, and $Recall@K$ for our pin classification task. Classification accuracy is defined as the ratio of correctly classified examples and the total number of instances. For instance, classification accuracy equals to $(tp+tn)/(tp+fp+fn+tn)$ for two class classification where tp is true positives, tn true negatives and fn is false negatives and tn equals to true negatives. Accuracy measure for multi-class classification gives overall performance of classification task. We used True Positive Rate to evaluate classification performance per category. True Positive Rate is the ratio of the correctly classified instances for the specific category to all examples of that category and equivalent to recall and specificity. True Positive Rate equals to $tp/tp + fn$ for two class classification.

$Recall@k$ is the fraction of correct labels that were retrieved at rank k. That is, for each test case, we compute the output of each classifier and sort them in descending order according to the probability of the class. $Recall@k$ provides a measure of how highly ranked the correct class is.

IV. Dataset

We used a crawler to find Pinterest users, their boards and pins. A user page contains a set of boards. Our crawler performed a depth-first search starting from a popular (many followers) user. Our initial dataset consists of over one million pins, corresponding to over 18,000 boards and 650 users. The number of pins in a board varies from a couple to several thousands. The average number of pins per user is 2,476, while the average number of pins per board is 55.6.

TABLE I: 32 predefined Pinterest categories.

animals
architecture
art
cars motorcycles
celebrities
design
diy crafts
education
film music books
food and drink
gardening
geek
hair beauty
health fitness
history
holidays events
home decor
humor
illustration posters
kids
mens fashion
outdoors
photography
products
quotes
science nature
sports
tattoo
technology
travel
weddings
womens fashion

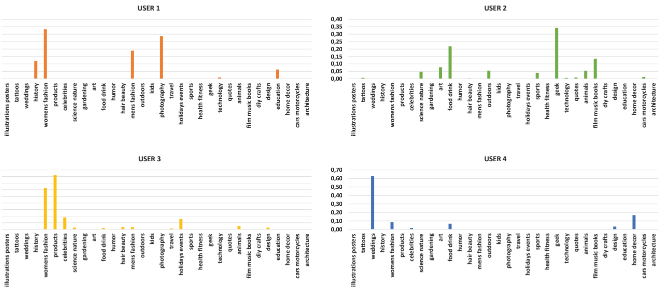

Fig. 2: Examples of pin frequency per category for different users.

For our experiments, we use the user boards that were labeled into one of the predefined Pinterest categories and we discard the rest (users may choose to not categorize their board). We are left with 216 users and 547,000 pins.[1] However, we discovered many (104,000) duplicated images, as people tend to repin items that already exist. To provide a fair evaluation, we removed all duplicate images from the dataset and ended up with 443,000 pins. From these, we randomly selected 1000 pins per category to train and test for the task of inferring user interests.

[1] Links to these pins with training-validation and test splits can be found by category at https://goo.gl/bfCwUw

In Table II, we can see the most frequent words in 'animals', 'gardening', 'diy crafts' and 'weddings' categories. For instance, in 'gardening' category the most frequent word is 'garden' and it is followed by 'love', 'plant' and 'flowers'. Eventhough, in social media platforms users often do not generate elaborate descriptions of the images, they use the words that give clues about the category of the image and class. In addition, their caption is mostly a clear indication of their interest perception of the given image. We exploit both visual cues and textual cues to infer user interests using a multi-modal approach.

However, our task is challenging due to within-class variations, inter-class similarity and misleading class labels. For instance, high variation is observed in the 'animals' category, since pins range from wild animals to pets and farm animals. This also causes a high variation of scenes and backgrounds, from jungle to house and streets. Furthermore, for some categories inter-class similarity is high. For instance, the categories 'hair beauty', 'women's fashion' and 'celebrities' are composed of very similar pins such as images of a woman with a fancy dress and stylish hair. Moreover, since the class labels are generated by the users, they are more likely to be noisy and can be highly context dependent. Some Pinterest data examples are presented [2].

TABLE II: The pairs of (the most frequent word, frequency of that word) of some Pinterest Categories.

animals	**gardening**	**diy crafts**	**weddings**
('dog', 815)	('garden', 1227)	('diy', 3227)	('wedding', 5233)
('baby', 751)	('love', 364)	('make', 2842)	('bridal', 1822)
('cute', 734)	('plant', 357)	('tutorial', 2373)	('dress', 1658)
('love', 710)	('grow', 356)	('love', 2031)	('love', 1173)
('cat', 568)	('flowers', 300)	('crochet', 1876)	('com', 1109)

V. Experimental Setup

To test our hypothesis, eleven experiments are conducted. Firstly, text features are extracted from N(=1000) randomly selected pins for each category.

We randomly split the dataset into 90% training-validation set and the remaining 10% is set aside as a test set. We applied 5-fold cross validation on the trainining-validation set to train a multi-class classifier. The same training-validation set is used to fine-tune CNN to Pinterest categories with larger validation set as 80% of the training-validation set as validation and remaining 20% as test set. Performance of multi-class classifier is tested on unseen test set. The same training-validation and test splits are used in all eleven experiments.

Two different text representations and three different image representations are used. The first text representation is the bag of textual words, where text documents described with their document term frequencies. The second one is sentence vectors where one sentence is the summation of word vectors that compose the sentence. Similarly, three different image representations are used: BoVW, CNN, and ft-CNN, respectively. Scale invariant feature transform is applied on images to extract local features. A k-means clustering algorithm is used to compute visual vocabulary. 2000 visual words are created and each image is described as histogram of words.

A SVM classifier is trained separately on each representation, such as sentence vectors or BoVW features. 32 SVM classifiers are trained for each category, conforming with the OVR scheme. OVR SVM class prediction is used to evaluate the predicted user interests. The eleven sets of experiments are summarized in Table III.

TABLE III: Input data and features for our 11 experiments.

	input data	features
1	images only	BoVW
2	images only	CNN
3	images only	ft-CNN
4	text only	BoW
5	text only	word2vec
6	images + text	BoVW + BoW
7	images + text	BoVW + word2vec
8	images + text	CNN + BoW
9	images + text	CNN + word2vec
10	images + text	ft-CNN + BoW
11	images + text	ft-CNN + word2vec

VI. Results and Discussion

We assess user-interest prediction performance of uni-modal features (only images or only text) in comparison to multi-modal predictions. Table IV illustrates uni-modal classification accuracy scores with corresponding feature dimensions. We observe that when we use text input only, we obtain classification accuracy of 37.16% with BoW features and 33.47% with word2vec features. Whereas, when we use images only, we achieve classification accuracy of 17.47% with BoVW features, 33.59% with pre-trained CNN, and 57.53% with fine tuned CNN for Pinterest dataset.

TABLE IV: Multi-class overall classification accuracy of uni-modal space for different input features: BoVW, CNN, ft-CNN, BoW and word2vec, and corresponding feature dimensions.

	images only			text only	
	BoVW	CNN	ft-CNN	BoW	w2v
feature dimension	2000	4096	4096	29779	300
accuracy	17.47%	33.59%	57.53%	37.16%	33.47%

We can further improve these results by linearly combining the class prediction of each uni-modal feature, as described in Eq. 1, where λ controls the contribution of each representation. We explore how prediction accuracy is affected for λ = [0.3, 0.5, 0.7].

TABLE V: Late Fusion effect of λ values on different input features

		lambda		
		0.3	0.5	0.7
BoVW	BoW	**38.44%**	34.66%	27.44%
BoVW	word2vec	**37.06%**	36.00%	29.66%
CNN	BoW	42.94%	**44.88%**	43.41%
CNN	word2vec	38.50%	42.06%	**43.09%**
ft-CNN	BoW	51.00%	58.22%	**60.94%**
ft-CNN	word2vec	45.91%	54.47%	**60.00%**

Table VI depicts multi-modal classification accuracies for late fusion of different feature pairs with corresponding feature dimensions. It shows that performance increases while leveraging the information to a multi-modal (images + text) space. BoVW + BoW improves the classification accuracy score to 38.44%, compared to uni-modal performance of 17.47% and 37.16% accuracy of uni-modal BoVW and BoW, respectively. Similarly, late fusion of the neural network representation CNN + word2vec, using both images and text yields to a classification accuracy of 43.09%. Whereas using only images (CNN) and only text (word2vec) leads to 33.59% and 33.47% classification accuracy, respectively. Furthermore,

[2] http://goo.gl/YqA5hv

classification accuracy of late fusion of CNN + BoW increases to 44.88% compared to uni-modal classification accuracies 33.59% (CNN) and 37.16% (BoW). We observe a similar behaviour with late fusion of BoVW + word2vec, using both images and text yields to a 37.06%, whereas individual classification accuracies of images and text are 17.47% and 33.47%, respectively. Moreover, ft-CNN + BoW and ft-CNN + word2vec enhance classification accurcies of uni-modals to 60.94% and 60.00% respectively.

TABLE VI: Multi-class overall accuracy of multi-modal space for different input features: BoVW, CNN, ft-CNN, BoW, word2vec and the dimension of each representation.

	images					
	BoVW	CNN	ft-CNN	BoVW -d	CNN -d	ft-CNN -d
BoW	38.44%	44.88%	60.94%	31779	33875	33875
word2vec	37.06%	43.09%	60.00%	2300	4396	4396

We performed Repeated Measures ANOVA to evaluate the significance of the improvement. Table VII presents the p-values when we compare multi-modal representations vs. uni-modal ones by employing pairwise t-tests with adjusted p-values. We observe significant differences between the multi-modal BoVW + BoW vs. the individual uni-modal representation BoVW ($p = 3.2 \times 10^{-11}$), multi-modal BoVW + word2vec vs. the uni-modal images only BoVW ($p = 1.6 \times 10^{-11}$) and word2vec ($p = 0.0004$). In addition, there is significant improvement with multi-modal late fusion of CNN + BoW vs. uni-modal images only CNN ($p = 9.7 \times 10^{-6}$) and BoW ($p = 2.2 \times 10^{-6}$), and multi-modal late fusion of CNN + word2vec vs. uni-modal representations: images only CNN ($p = 2.8 \times 10^{-8}$) and text only word2vec ($p = 2.1 \times 10^{-6}$). The multimodal BoVW + BoW vs. the uni-modal text-only (word2vec) and multi-modal ft-CNN + word2vec vs. uni-modal ft-CNN representation did not yield significant differences, but the p-value is low ($p = 0.0927$). These results strengthen our intuition that jointly processing images and text for inferring users' interests is beneficial. Finally, we obtain significant improvement with multi-modal late fusion of ft-CNN + BoW vs. uni-modal images only ft-CNN ($p = 0.0083$) and text only ($p = 8.8 \times 10^{-11}$), and multi-modal late fusion of ft-CNN + word2vec vs. text only word2vec ($p = 5.2 \times 10^{-15}$).

There are further observations from these results: 1) When using only images, convolutional neural networks outperforms the bag of visual words representation, by almost a factor of two and fine tuned CNN for pinterest dataset outperforms other image features. This shows the power of this representation for image processing, and in particular for the task of modeling user interests from images. 2) When using only text, the bag of word representation yields better results than the word embeddings. It is worth noticing that the dimensionality of word embeddings is much lower (300-d) than that of bag of words (29K-d). While the former obtains lower scores, it is still remarkable that it is able to compress meaningful information in a much smaller dimensional space. This lower dimensional space yields to lower computational complexity. Low dimensional representation is essential in real life applications to address large number of users. 3) A similar remark is true when comparing the joint image-text representation: while BoW achieves a higher score, it does so with a much larger dimensional space (31k-d), while NNs has just over four thousand dimensions. The dimensionality of each feature is presented on Table IV and VI.

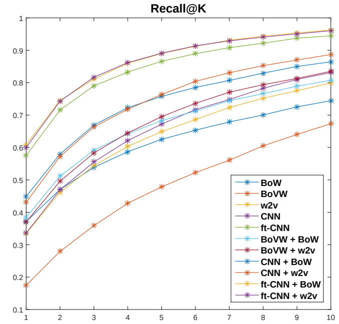

Fig. 3: Multi-class Classification $Recall@K$ on Different Input Features.

The $Recall@K$ results are given in Fig. 3. We can see that the multi-modal (image + text) approach provides higher average recall at each K than the uni-modal (image only/text only) constituents. This is valid for all multi-modal image-text feature pairs, 'BOVW + BoW', 'BoVW + word2vec', 'CNN + BoW', 'CNN + word2vec', 'ft-CNN + BoW', and 'ft-CNN + word2vec', respectively. For instance, 'BOVW + BoW' $Recall@K$ curve is always higher than 'BoVW' and 'BoW' $Recall@K$ curves.

Furthermore, we can break down the results from the highest-score representation (ft-CNN + word2vec) into individual accuracy scores for each class. The true positive rate of the results are shown on Fig. 4. We see that the 'cars motorcycles', 'tattoo', and 'food drink' classes are predicted with higher accuracy. This can be explained by less inter-class similarity. For instance, in the 'cars motorcycles' category most of the images include a motor vehicle. Whereas, it is less likely to see 'tattoo' category images in other categories. Furthermore, the low performance of the categories such as 'design' and 'photography' can be explained by high inter-class similarity with categories such as 'art', 'diy crafts', 'home decor', 'architecture' and 'travel'.

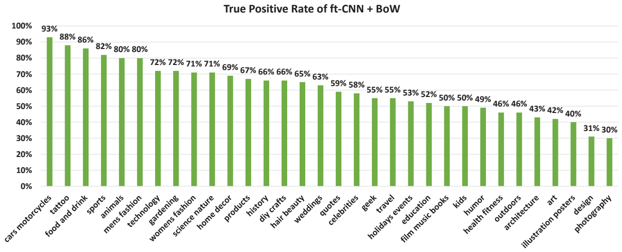

Fig. 4: Multi-class Classification True Positive Rates for Different Categories.

VII. Conclusion and Future Work

We have studied and compared different approaches for inferring users' interest on social media from text and images. This is a very useful task because the Web is a multi-modal space where relevant information can be extracted from visual and textual content. It is also a challenging task because of large variability between users and and within classes.

In this study, we have shown that we can learn a multi-class classifier to infer broad interest by using user contributed

TABLE VII: Repeated Measures ANOVA p-values to compare multi-modal vs uni-modal representations.

		multi-modal					
		BoVW + BoW	BoVW + word2vec	CNN+ BoW	CNN + word2vec	ft-CNN + BoW	ft-CNN + word2vec
uni-modal	images only	3.2×10^{-11}	1.6×10^{-11}	9.7×10^{-6}	2.8×10^{-8}	0.0083	0.0927
	texts only	0.25	0.0004	2.2×10^{-6}	2.1×10^{-6}	8.8×10^{-11}	5.2×10^{-15}

content and user generated text data. We used Pinterest which enable us to model the broad user interests by using multi-modal media (visual, textual, audio, etc) with available broad interest labels. The multi-class classifier can be used to infer user interests on other social media platforms such as Instagram, Facebook. In addition, it can also be used for pins without category label to increase the personalized user experience on Pinterest.

Our results show that using a multi-modal approach outperforms uni-modal features for inferring a user's interest. In the future, we will explore fine-tunning CNN Pinterest categories to achieve higher results. Moreover, other sentence representations can be explored such as Latent Dirichlet Allocation (LDA) or Probabilistic Latent Semantic Analysis (pLSA) which also represent semantics and some patterns as word embeddings. In such setting the distribution of topics in a sentence obtained with the LDA or pLSA can be used as features. In addition, word embeddings can be trained with convolutional neural networks to train sentence vectors from word vectors [28].

References

[1] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: A bibliography," *SIGIR Forum*, vol. 37, no. 2, pp. 18–28, Sep. 2003.

[2] X. Shen, B. Tan, and C. Zhai, "Implicit user modeling for personalized search," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ser. CIKM '05. New York, NY, USA: ACM, 2005, pp. 824–831.

[3] W. Shen, J. Wang, P. Luo, and M. Wang, "Linking named entities in tweets with knowledge base via user interest modeling," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 68–76.

[4] D. Yin, S. Guo, B. Chidlovskii, B. D. Davison, C. Archambeau, and G. Bouchard, "Connecting comments and tags: Improved modeling of social tagging systems," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 547–556.

[5] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu, and H.-Y. M. Liao, "Personalized travel recommendation by mining people attributes from community-contributed photos," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 83–92.

[6] P. Xie, Y. Pei, Y. Xie, and E. P. Xing, "Mining user interests from personal photos," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 2015, pp. 1896–1902.

[7] Q. You, S. Bhatia, and J. Luo, "A Picture Tells a Thousand Words – About You! User Interest Profiling from User Generated Visual Content," *ArXiv e-prints*, Apr. 2015.

[8] M. Szomszor, H. Alani, I. Cantador, K. OHara, and N. Shadbolt, "Semantic modelling of user interests based on cross-folksonomy analysis," in *The Semantic Web - ISWC 2008*, ser. Lecture Notes in Computer Science, A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, and K. Thirunarayan, Eds. Springer Berlin Heidelberg, 2008, vol. 5318, pp. 632–648.

[9] I. Schwab, A. Kobsa, and I. Koychev, "Learning user interests through positive examples using content analysis and collaborative filtering," in *30 2001. Internal Memo, GMD*, 2001.

[10] R. Kawase and E. Herder, "Classification of user interest patterns using a virtual folksonomy," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '11. New York, NY, USA: ACM, 2011, pp. 105–108.

[11] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. McGraw Hill Book Co, New York, 1983.

[12] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 1470–1477 vol.2.

[13] G. Csurka and C. Dance, "Visual categorization with bags of keypoints," *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22, pp. 1–2, 2004.

[14] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1458–1465 Vol. 2.

[15] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval - MIR '07*. New York, New York, USA: ACM Press, Sep. 2007, p. 197.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.

[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003.

[20] 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, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, 2015.

[21] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[23] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *CoRR*, vol. abs/1412.2306, 2014.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] A. Vedaldi and B. Fulkerson, "VLFeat - an open and portable library of computer vision algorithms," in *ACM International Conference on Multimedia*, 2010.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[27] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," *arXiv preprint arXiv:1311.3715*, 2013.

[28] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *CoRR*, vol. abs/1404.2188, 2014.