ISSN 1870-4069

User Identification in Pinterest Through the Refinement of a Cascade Fusion of Text and Images

Juan Carlos Gomez, Mario-Alberto Ibarra-Manzano, and Dora-Luz Almanza-Ojeda

Universidad de Guanajuato, Departamento de Ingeniería Electrónica, DICIS,
Salamanca, Mexico

{jc.gomez,ibarram,dora.almanza}@ugto.mx

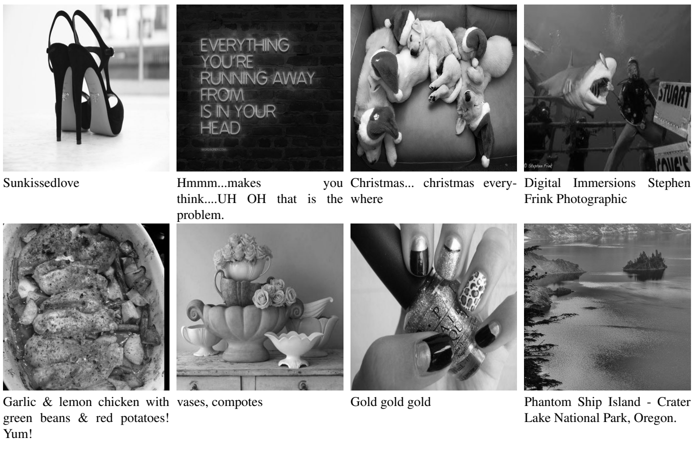
Abstract. User identification in social media is of crucial interest for companies and organizations for purposes of marketing, e-commerce, security and demographics. In this paper, we aim to identify users from Pinterest, a platform where users post *pins*, a combination of an image and a short text. This type of multi-modal content is very common nowadays, since it is a natural way in which users express their interests, emotions and opinions. Thus, the goal is to identify the user that would post a particular pin. For solving the problem, we propose a two-phase classification model. In a first phase, we train independent classifiers from image data, using a deep learning representation, and from text data, using a bag-of-words representation. During testing we apply a cascade fusion of the classifiers. In a second phase, we refine the output of the cascade for each test pin by selecting the top most likely users for the test pin and re-weighting their corresponding output in the cascade by their similarity with the test pin. Our experiments show that the problem is very hard because several reasons with the data distribution, but they also show promising results.

Keywords. Social media, user identification, pinterest, deep learning, information fusion.

1 Introduction

In the Big Data era, a large amount of information is created and transmitted around the world through the Internet [3]. Much of the traffic occurs on social media and similar platforms, where users create and share multimedia content such as news, reports, videos, emotions, music, opinions, etc. The content generated by users in social media has some particularities [26,34]: is plentiful, is constantly generated, is dynamic (shared and distributed by the users in real time), is representative of users or groups of users, and is multi-modal [27], meaning it is composed by a mixture of text, images, videos, audio and links that connect users and websites (friends, followers, shares, reactions, related websites, etc.). All the information that users generate in social media is a tool that can help to draw specific user profiles [21,7], and represents a digital footprint that identify how the persons use the social media, indicating their tastes, likes, preferences, personalities, sentiments, types of friendships or connections, etc.

Currently, there is a great interest from companies and organizations to analyze user generated content in social media, with the purpose of obtaining useful indicators for



Sunkissedlove

Hmmm...makes you think....UH OH that is the problem.

Christmas... christmas everywhere

Digital Immersions Stephen Frink Photographic

Garlic & lemon chicken with green beans & red potatoes! Yum!

vases, compotes

Gold gold gold

Phantom Ship Island - Crater Lake National Park, Oregon.

Fig. 1. Examples of pins from Pinterest. Each pin is formed by an image and a short comment about the image.

making decision on several areas [6,30,9], such as security, politics, social, educational, commercial and marketing. In this direction, user identification serves to several purposes. For companies and organizations, it helps to group persons by similar interests and behavioral patterns, recognize lead users, influencers, potential customers, political sympathizers and even to detect trolls, intruders, terrorists or persons that are a threat for public security [5]; considering that in social media users can assume multiple or false identities. Additionally, users can benefit by receiving personalized content that is adequate for their needs of entertainment, shopping, security, health and education.

Social media posts present several characteristics that are challenging for user identification. First, they are multi-modal, since this is a natural way for persons to express an opinion, emotion or to share content. This type of content suffers of a semantic gap between modalities, since what is expressed in text could not be representative of what is shown in an image. Second, users in general do not follow grammar and spelling rules, which makes difficult to use high-level language attributes such as syntax and semantics [21]. Third, text varies largely in length among posts. Fourth, text and images are highly heterogenous, including a large diversity of topics. Fifth, there are many users, and their content could overlap.

In this work, we present a two-phase model for user identification in Pinterest, which is one of the most popular social networks in the world with over 150 million of monthly active users. In Pinterest, users post *pins*, a combination of an image and a short text about the image. Figure 1 presents some example of pins, showing the diversity of content that we can find in this social site.

We define our task as identifying the specific user that would have post a pin (combination of image and text). We pose the problem as a single-label multi-class classification task, where the classes correspond to specific users and a pin could belong to only one user. For solving the problem, we propose a two-phase classification model. In a first phase, we train independent classifiers from image data, using a deep learning representation, and from text data, using a low-level bag-of-words representation. During testing, we apply a cascade fusion of the classifiers, where the probabilities outputs of both classifiers are multiplied sequentially. The cascade fusion classifier is aimed to combine the two data modalities for reducing the semantic gap and exploiting better the whole content. In a second phase, we refine the output of the cascade fusion for each test pin. The refinement is done per test pin, selecting the top most likely users for such pin, and then adjusting their corresponding probability scores by measuring the maximum similarity of all their training pins with the test pin. Our experiments confirm that the task is highly complex, due to the reasons explained above, but also show promising results.

The rest of the paper is organized as follows. Section 2 briefly reviews the related works. Section 3 presents our methodology, including the data representations and the two-phase classification model. Section 4 describes the setup used for experimentation, including the dataset and the evaluation metrics. Section 5 shows the results. Finally, Sect. 6 concludes the paper with an overall discussion and future research directions.

2 Related Work

User identification has been conducted using different representations for the user generated data, but most of the works are based exclusively on text data [21,28,31], image data [13,32], link data [35,16] and metadata [29]. In this work, we aim to combine multi-modal data from text and images trying to exploit most of the content for a better user identification.

There exist several approaches for user identification [1]. In [7,2,20] clustering techniques such as K-means, DBSCAN and normalized mutual information are used as a first approach for grouping users with similar patterns, and then identified particularities of each one. Similarly, classification methods such as K-Nearest Neighbors [21], Gradient Boosted Decision Trees [23] and Support Vector Machines [17] have been used for identifying user based on their profiles. In addition, some works follow an information retrieval approach and estimate distance metrics for ranking possible users for a given content [22,33]. In this work, we employ a single-label multi-class model approach, where each class correspond to a specific user, and the content could be assigned exclusively to one user.

Image and text data can be represented in different manners. In text, while high-level [24] and deep learning [18] text features can reflect users' writing styles and syntactic and semantic relations [19], lower-level text features such as words frequencies and n-grams could produce similar results with a lower cost. For example, in [4] the authors found that a bag-of-words representation outperforms the deep learning representation of word2vec [18] when inferring users' interests in Pinterest. In images, the state of the art is to use deep learning features, mainly using Convolutional Neural Networks (CNN), that have achieved good performance in many image classification and object detection tasks [15,25], as well as in image description generation [14] and automatic comments generation [12]. In this work, we employ a bag-of-words representation of text and a deep learning representation of images, both for its good results working with similar data.

Our work is particularly close related to [4], where the authors present a model that linearly combines the output of two independent classifiers trained over the modalities of text and images for inferring user interests in Pinterest. We use their model as a baseline for comparison with ours, considering that we are trying to identify specific users, rather than their interests. User identification is a more challenging task because of the number of possible classes involved. Additionally, we use as baselines models that are built using exclusively text or image data.

3 Methodology

In Pinterest, users post *pins* which are organized in boards. A pin is an ¡image,text¿ pair, and a board groups several pins that represent particular users' interests. Here we discard the board information, and put all the pins from a user in a single collection. Be u a user and $\mathbf{P}_u = (\mathbf{g}_i^u, \mathbf{x}_i^u)$ his collection of pins, where $\mathbf{g}$ refers to the image and $\mathbf{x}$ to the text. There are $i = 1, \ldots, n_u$ pins per user, m total users and $n = \sum_i = 1^m(n_{u_i})$ total pins. The task consists in identifying the user $\widehat{u}$ that would have generated a pin $p_t = (\mathbf{g}_t, \mathbf{x}_t)$

We stack all the pins of all the users in a single dataset, and we split it in a training, a validation and a test set, keeping the same proportions for each user. With text, we first cleaned each pin by removing special symbols (e.g. hash tags, asterisks, etc.), stop words, URLs, one-letter words and long words (> 30 characters). Using exclusively the training set we extract a dictionary removing words appearing in only one pin. Afterwards, we use a tf-idf weighting schema to build document-term matrices $\mathbf{X}_{tr}$, $\mathbf{X}_v$ and $\mathbf{X}_t$ for each part of the data. We preprocessed each image using a Convolutional Neural Network to obtain the corresponding row vector $\mathbf{g}^u$ of features for user u. We used for that the DeCAF library [8], considering the activation values of the 4,096 neurons in the 7-th layer as the image features. The DeCAF model used for the transformation of images, was pre-trained with the Imagenet dataset [15]. We join all the image vectors corresponding to the training, validation and test sets in matrices $\mathbf{G}_tr$, $\mathbf{G}_v$ and $\mathbf{G}_t$ that are row paired with matrices $\mathbf{X}$ and sorted by user.

3.1 Classification Model

Our classification model consists of two phases. In the first phase, we train two independent classifiers using separately the pins' text data $\mathcal{F}_{\mathbf{x}}$ and image data $\mathcal{F}_{\mathbf{g}}$ considering we have m different users representing the classes. In this work, we employ logistic regression as the individual classifier, since this model can be trained naturally for a multi-class problem, and directly outputs a probability for each user. The models are first built using the training sets $\mathbf{X}_{tr}$,$\mathbf{G}_{tr}$ and use the validation sets $\mathbf{X}_v$,$\mathbf{G}_v$ for independently optimize the regularization parameter C, considering the values [0.1,1,10,100]. After optimization, we merge the training and validation sets in two single sets and train independent models with the whole data using the optimal C.

During testing, for each test pin $p_t = (\mathbf{g}_t, \mathbf{x}_t)$ we classify independently its text part as $\mathcal{F}_{\mathbf{x}}(\mathbf{x}_t)$ and its image part as $\mathcal{F}_{\mathbf{g}}(\mathbf{g}_t)$, producing the corresponding two vectors of probabilities of belonging to each user: $\mathbf{r}_x = [r_{u_1}^x, r_{u_2}^x, \ldots, r_{u_m}^x]$ and $\mathbf{r}_g = [r_{u_1}^g, r_{u_2}^g, \ldots, r_{u_m}^g]$. Afterwards, we combine the two outputs in a cascade fusion as:

$$\mathbf{r} = \mathbf{r}_x \cdot \mathbf{r}_g. \qquad (1)$$

This equation outputs a fusion probability vector $\mathbf{r} = [r_{u_1}, r_{u_2}, \ldots, r_{u_m}]$ that merges together the two data modalities. We can select from this vector the user $\widehat{u}$ for the test pin p_t as the one with the highest probability.

In a second phase, for the test pin p_t we extract the top 10 most likely users $top = [\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_10]$ and their fusion probabilities $\mathbf{r}_{top} = [r_{\tilde{u}_1}, r_{\tilde{u}_2}, \ldots, r_{\tilde{u}_10}]$. Using the training pins of each top user $\mathbf{P}_{\tilde{u}_i}$ we compute the cosine similarity of all their pins with the test pint p_t and find the maximum similarity $l_{\tilde{u}_i}$, and re-weight its corresponding probability in vector $\mathbf{r}_{top}$ as $r_{\tilde{u}_i} = r_{\tilde{u}_i} * l_{\tilde{u}_i}$. Finally, from vector $\mathbf{r}_{top}$ we select user $\widehat{u}$ for the test pin p_t as the one with the highest value.

4 Setup

In this work, we use a dataset of 70,200 pins belonging to 117 users, that were randomly collected by directly crawling the Pinterest website[1]. Thus, there are 117 different classes. We selected at random three boards per user, saving 200 pins per board. We discarded the board information and merged all the pins for a total of 600 pins per user. For the experiments, we split the dataset selecting at random 400 pins per user for training, 100 pins per user for validation and 100 pins per user for testing.

All the pins contain and image and a text. The comments are in English and are of a variable length from one (12.33% of the pins) to a maximum of 552 words. The dictionary to build the document-temr matrix was extracted from the training set during validation and from training and validation sets when building the final model. The final dictionary was composed of 17,145 words. In Table 1, we show the five top common words, with their pin frequency, used by five random users. We observe there the diversity of topics the users talk about. In Figure 2, we show statistics about the minimum, maximum, median and average number of words in pins per user (using a logarithmic scale). The distribution of words is generally regular, with some users having larger comments, considering that the minimum and the median are similar. The statistics for the number of words in pins in the whole dataset are a minimum of 1, a maximum of 552, a median of 4 and average of 8.5.

Table 1. Most frequent words for a sample of 5 users; indicating the number of pins for that user on which such word appears.

User 1	User 2	User 3	User 4	User 5
logo(77)	make(45)	love(28)	crochet(107)	diamond(155)
design(64)	diy(34)	vintage(20)	pattern(56)	ring(144)
via(51)	cream (32)	one(14)	free(48)	necklace(127)
infographic(46)	chicken(30)	black(12)	com(33)	gold(82)
designspiration(46)	cake(30)	elizabeth(10)	art(30)	sapphire(43)

[1] Dataset is available at: `https://data.mendeley.com/datasets/fs4k2zc5j5/2`

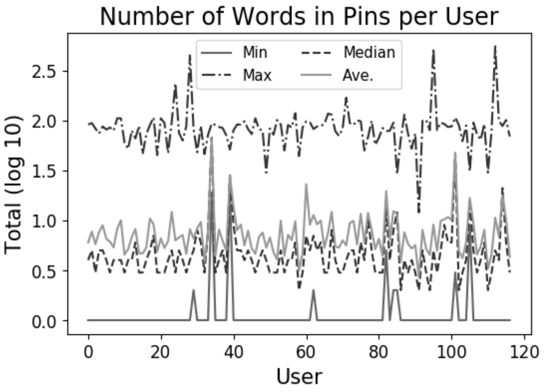


Fig. 2. Words statistics in pins over users.

The images in the dataset also vary in size, but they are all in JPG format. Only 4.2% of images are shared by 2 or more users, which shows the great diversity of content in the pins, including products (e.g. clothes and jewelry), interests (e.g. food and decoration), photographs (e.g. animals and landscapes) and more abstract content (e.g. paints and designs).

Text and images even if coupled, represent different modalities of data, and between them exist a semantic gap, since a comment about an image could not be expressing the same meaning as the visual depiction. That is, a text comment could be applied to several images, and an image could be described/commented in many ways. In Figure 3 we show the whole of text and image data projected independently over the two principal components (PC) after transforming the data using principal component analysis (PCA). In the figure, we observe how the text and image data are distributed differently along the PCs. For text data, the general explained variance using 10 PCs is 0.043, while for image data is 0.321. Images in our dataset are more compact in their features than text, and could be better compressed using PCA. That means that their features are more homogenous, even if its content is properly highly heterogenous. This could serve to reduce the feature space and simplify some computations, but also it does not help for separating data of different users. On the other hand, text data is more heterogenous in its features, like its content, and could not be well compressed with PCA, but it serves to better distinguish the data from different users up to a certain level.

4.1 Baselines and Evaluation

As first baselines, we use the two logistic regression models that are built using separately the pins' text data $\mathcal{F}_{\mathbf{x}}$ and image data $\mathcal{F}_{\mathbf{g}}$. Such models are the base for the cascade fusion model and are optimized for the regularization parameter C, using the validation set and considering values of $C = [0.1, 1,]$. The other set of baseline models are taken from [4]. In that work, the authors built a model that when testing

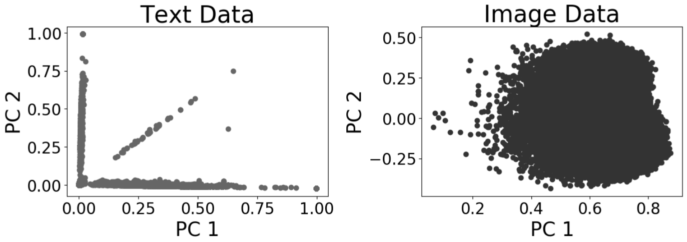


Fig. 3. Distributions of text and image data over two principal components.

for a pin_t, it takes the output of classifiers $\mathcal{F}_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{g}}$, and using a late fusion schema, weight the contribution of each one depending on a parameter λ, as follow $pred_{p_t} = \lambda \mathcal{F}_g(\mathbf{g}_t) + (1 - \lambda)\mathcal{F}_x(\mathbf{x}_t)$. We consider three values for $\lambda = [0.3, 0.5, 0.7]$, as in the mentioned work.

We compare the performance of all the methods using the standard classification measures: accuracy and F1. These are defined as $accuracy = \frac{(tp+tn)}{(tp+fp+fn+tn)}$ and $F1 = 2\frac{precision \cdot recall}{precision + recall}$, where tp are the true positives, tn the true negatives, fp the false positives and fn the false negatives. Accuracy measures the proportion of corrected identified users, while the F1 measure represents the harmonic mean of precision and recall, that are in turn defined as: $precision = \frac{tp}{tp+fp}$, $recall = \frac{tp}{tp+fn}$. We compute the macro averages for F1, since in our case the classification is single-label, and the micro averages of F1 are equal to the accuracy [11].

We ran all our experiments using a Windows PC with a 2.5 GHz Core i5 processor and 8 GB in RAM. We implemented all our methods in Python[2], using the scikit-learn[3] and NumPy[4] libraries.

5 Results

Table 2 shows the summary of results for all the models. We can see than in general, the performance of the models is low, with less than 40% for accuracy and F1, making clear that the problem is hard to solve. We observe that using only images for user identification produces the lowest results. As mentioned before, image data is more homogenous on its features, but not on its content, and could have problems to separate the data of different users. We believe this is associated with the transformation done with the DeCAF library, which is trained using a generic dataset of images (Imagenet). We guess that a fine tuning of the DeCAF model with images from Pinterest could help to improve this performance. When using only text, the results are better, as expected from the same argument as before, text reflects better the heterogeneity of the data and can discriminate better the data of different users. Our cascade fusion (CF) model performs better than all the baselines, with near 4% of improvement for accuracy and

[2] Code is available at: `https://github.com/jcgcarranza/2017rcs_code`

[3] `http://scikit-learn.org`

[4] `http://www.numpy.org/`

F1 regarding text, and 1% for F1 regarding the $\lambda = 0.3$ model. The cascade fusion model reaches a recall@10 of 75%, that would be the maximum accuracy performance expected if we ordered correctly the 10 most likely users per test pin. When we use refinement over the cascade fusion model (CF+Ref) the results improve in about 1% more for accuracy and F1. It is thus clear, that the cascade fusion takes better advantage of the information coming from both modalities.

Table 2. Results for the different models with the different metrics.

Model	**Accuracy**	**F1**
Text only	33.75	33.21
Images only	21.62	19.72
[4] $\lambda = 0.3$	37.18	35.01
[4] $\lambda = 0.5$	34.68	32.11
[4] $\lambda = 0.7$	30.73	28.10
CF	37.34	36.16
CF+Ref.	**38.30**	**37.46**

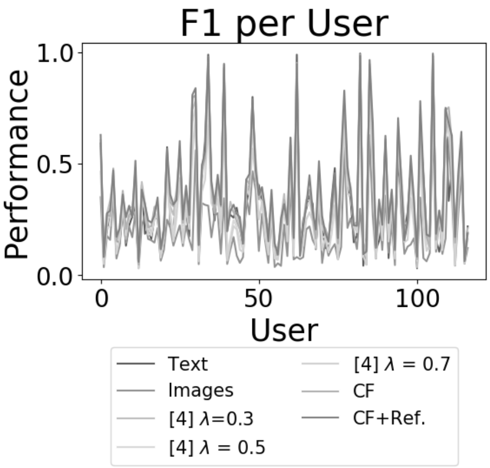


Fig. 4. Performance in F1 for each user.

To better understand the low performance in the task, in Figure 4 we show the performance of the different models per user for F1. All the models present a similar behavior along the different users, with the cascade fusion model and the refined model showing plots a little above than the others. We observe in this plot that some users are more easily identified than others. There are seven users for which the performance is less than 10%, and there are 6 for which the performance is more than 90%. In Figure 5 we plot a histogram of the F1 performance for the CF+Ref model. We can see that for most of the cases user identification is difficult, with general performances below the 50%.

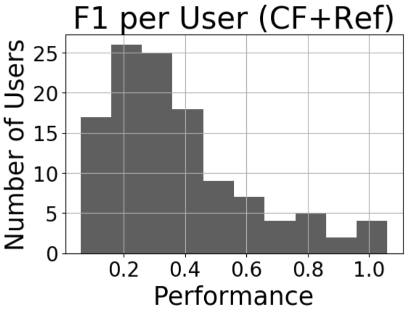


Fig. 5. Histogram of the performance in F1 for the CF+Ref model.

Table 3. Users with the lowest and the highest F1 performance for the CF+Ref model, paired with the median number of words in their pins.

Lowest Performance			**Highest Performance**		
User	**Median**	**F1**	**User**	**Median**	**F1**
13	4	0.059	106	16	0.995
116	6	0.062	83	13.5	0.985
2	5	0.081	34	68	0.980
101	6	0.083	86	12	0.966
64	5	0.086	40	28	0.949

In Table 3 we show the top five users for which the performance is the lowest and the top five for which it is the highest, showing the F1 metric value and the median length of their pins. Users for which the performance is low, have pins with medians between 4 and 6, while users for which the performance is high have pins with medians between 13.5 and 68. Indeed, considering all the users there is a weak Pearson correlation (0.46) between the performance and the median length of the comments in pins. This means that in general, users that post longer comments provide with more information to better identify them. The general median for all the users is 4, which partially would explain the low performance for most of the users, meaning that users would tend to post very short comments in Pinterest. A second part to explain the low performance comes from the images. As mentioned before, the deep learning features are good for representing the data, but not for discriminate the data of different users. Moreover, in this case for images it is more complicated to analyze the features in detail, since the used image features are convolutions of other more general features, and the semantic meaning is lost in the process. Another part playing in the low performance is the number of users, there are 117 in total, and from Figure 3 we observe that even if the content shows heterogeneity, the distributions of text and image data could be similar among different users.

6 Conclusions and Research Directions

In this work, we have presented a two-phase classification model for user identification in Pinterest, a very popular social media site where users post pins (a mixture of an image and a companion text comment). Our task consisted on identifying the specific user that would have generated a pin. We treated the problem as a single-label multi-class classification task, where the classes correspond to specific users and a pin could belong to only one user. Our classification model in a first phase it trains independent classifiers from image data, using a deep learning representation, obtained from a convolutional neural network mode, and from text data, using a low-level bag-of-words representation weighted using a tf-idf schema. During testing, we apply a cascade fusion of the classifiers, where the probabilities outputs of both classifiers are multiplied sequentially. The cascade fusion model tries to combine the two data modalities for reducing the semantic gap and exploiting better the whole content. In a second phase, we refine the output of the cascade fusion for each test pin. The refinement is done by selecting the top most likely users for the test pin, and then adjusting their corresponding probability scores by measuring the maximum similarity of all their training pins with the test pin.

We tried our model with a dataset of 70,200 pins from 117 users, and compare it with some baselines from the literature. The results showed that the task is very hard, since all the models reached less than 40% of performance for accuracy and macro F1. When analyzing individual performance per user, we observe that only a few of them (6) are identified with more than 90% of accuracy, while most of them are identified with less than 50% of accuracy. In the lowest extreme, there are 7 users for which the performance is less than 10%. In general, we observe a weak Pearson correlation between the length of the user pins (as measured by the median length) and the associated performance. Additionally, image features obtained by a convolutional neural network are good for representing the data, but not for discriminate the data of different users. Text features reflex better the heterogeneity of the data, but also present similar distributions among users. Finally, the number of possible classes is also a challenge, since there is a bigger chance than the distribution of images and text overlap between users.

Future research directions include the use of a DeCAF model fine-tuned with data collected from Pinterest, for computing the image features. Additionally, we guess that it is necessary to transform both modalities of data to other feature spaces, where the data would be better separate [10]. This could be done for each modality independently or both at the same time. Finally, other late fusion models could be explored to combine the outputs of independent classifiers, especially considering that the recall@10 of the cascade fusion model is 75%, meaning that between the 10 most likely users selected per test pin, there is a 75% of chance of finding the correct one. Thus, a better model to re-order the selected users could help to improve the performance.

Acknowledgments. This research was supported in part by PRODEP-SEP project folio UGTO-PTC-595, oficio 511-6/17-8074, by CONACYT-INEGI project Fondo Sectorial Conacyt-Inegi No. 290910, and by Universidad de Guanajuato.

References

1. Barforoush, A.A., Shirazi, H., Emami, H.: A new classification framework to evaluate the entity profiling on the web: Past, present and future. ACM Computing Surveys (CSUR) 50(3), 39 (2017)
2. Cerquitelli, T., Servetti, A., Masala, E.: Discovering users with similar internet access performance through cluster analysis. Expert Systems with Applications 64, 536–548 (2016)
3. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile Networks and Applications 19(2), 171–209 (2014)
4. Cinar, Y.G., Zoghbi, S., Moens, M.F.: Inferring user interests on social media from text and images. In: Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. pp. 1342–1347. IEEE (2015)
5. Cohen, K., Johansson, F., Kaati, L., Mork, J.C.: Detecting linguistic markers for radical violence in social media. Terrorism and Political Violence 26(1), 246–256 (2014)
6. Criado, J.I., Sandoval-Almazan, R., Gil-Garcia, J.R.: Government innovation through social media. Government Information Quarterly 30(4), 319–326 (2013)
7. van Dam, J.W., van de Velden, M.: Online profiling and clustering of facebook users. Decision Support Systems 70, 60–72 (2015)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
9. Fan, W., Gordon, M.D.: The power of social media analytics. Communications of the ACM 57(6), 74–81 (2014)
10. Gomez, J.C., Boiy, E., Moens, M.F.: Highly discriminative statistical features for email classification. Knowledge and information systems 31(1), 23–53 (2012)
11. Gomez, J.C., Moens, M.F.: Minimizer of the reconstruction error for multi-class document categorization. Expert Systems with Applications 41(3), 861–868 (2014)
12. Gomez, J.C., Tommasi, T., Zoghbi, S., Moens, M.F.: What would they say? predicting user's comments in pinterest. IEEE Latin America Transactions 14(4), 2013–2019 (2016)
13. Hum, N.J., Chamberlin, P.E., Hambright, B.L., Portwood, A.C., Schat, A.C., Bevan, J.L.: A picture is worth a thousand words: A content analysis of facebook profile photographs. Computers in Human Behavior 27(5), 1828–1833 (2011)
14. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3128–3137 (2015)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
16. Liu, S., Wang, S., Zhu, F., Zhang, J., Krishnan, R.: Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data. pp. 51–62. ACM (2014)
17. Malhotra, A., Totti, L., Meira Jr, W., Kumaraguru, P., Almeida, V.: Studying user footprints in different online social networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. pp. 1065–1070. IEEE (2012)

18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
20. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 251–260. ACM (2010)
21. Peng, J., Choo, K.K.R., Ashman, H.: Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. Journal of Network and Computer Applications 70, 171–182 (2016)
22. Peng, J., Detchon, S., Choo, K.K.R., Ashman, H.: Astroturfing detection in social media: a binary n-gram–based approach. Concurrency and Computation: Practice and Experience 29(17) (2017)
23. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. Icwsm 11(1), 281–288 (2011)
24. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Detecting and tracking the spread of astroturf memes in microblog streams. arXiv preprint arXiv:1011.3768 (2010)
25. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3), 211–252 (2015)
26. Russell, M.A.: Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. O'Reilly Media, Inc. (2013)
27. Scott, J.: Social network analysis. Sage (2017)
28. Shen, X., Tan, B., Zhai, C.: Implicit user modeling for personalized search. In: Proceedings of the 14th ACM international conference on Information and knowledge management. pp. 824–831. ACM (2005)
29. Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PloS one 10(3), e0115545 (2015)
30. Tan, W., Blake, M.B., Saleh, I., Dustdar, S.: Social-network-sourced big data analytics. IEEE Internet Computing 17(5), 62–69 (2013)
31. Tuarob, S., Tucker, C.S.: Automated discovery of lead users and latent product features by mining large scale social media networks. Journal of Mechanical Design 137(7), 071402 (2015)
32. Wu, Y.C.J., Chang, W.H., Yuan, C.H.: Do facebook profile pictures reflect user's personality? Computers in Human Behavior 51, 880–889 (2015)
33. Yang, Y.C.: Web user behavioral profiling for user identification. Decision Support Systems 49(3), 261–271 (2010)
34. Zafarani, R., Abbasi, M.A., Liu, H.: Social media mining: an introduction. Cambridge University Press (2014)
35. Zafarani, R., Liu, H.: Connecting users across social media sites: a behavioral-modeling approach. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 41–49. ACM (2013)