chist-era

User Context Mining Module

Alexandru L. Ginsca*, Mihai Lupu**, Adrian Popescu*

*CEA, LIST, LVIC France

** Vienna University of Technology, ISIS, IMP

Contacts: lupu@ifs.tuwien.ac.at, adrian.popescu@cea.fr

MUCKE Project, Deliverable 3.2

01/12/2014

Contents

Abstract

In the MUCKE framework, user credibility estimates are used to filter or rerank a list of retrieved items according to the users who produced them. While hints about a user's credibility can be directly derived from the content of his contributions, we explore in this work possible credibility indicators that can be extracted from the context provided by the Flickr platform. We propose context features stemming from various data sources, such as Flickr groups, photo favorites or a user's contacts network.

Our first contribution is the creation of a new dataset specifically built to help us evaluate potential indicators for credibility but also to serve as a training dataset on which we can compare multiple learning models and features. We present the motivation behind the need for such a dataset, our methodology used for the creation of the dataset and detail important statistics on the number of users, images and rater agreement scores. We then propose a large set of features, for which we describe the data acquisition process and test their usefulness as individual credibility estimators. Finally, we define a credibility prediction problem, in which we learn regression models that provide better credibility estimators than the individual features. We also test several feature ranking and selection methods and, for the best configuration, we notice a 30% improvement over the best individual feature.

1 INTRODUCTION

Features extracted from a user's activity in the community have been successfully used to classify users in several social media platforms. In [22], the authors use, among other indicators, statistics about the user's immediate network (e.g., number of followers/friends) and communication behavior (e.g., retweet frequency) to classify latent user attributes, including gender, age or regional origin of Twitter users. A combination of features extracted both from the user's profile and interactions in the community and from user generated content have been proposed for expert identification in community question answering websites. Liu et al.[14] use a vector space model to represent the question and user profiles as term vectors. The proposed expert-finding method compares the similarity of questions and user profiles and takes into consideration the differences of expertise level, posting time of query, and the number of replies to questions. In [13], the authors propose an approach that considers user subject relevance, user reputation and authority of a category in finding experts. There, a user's subject relevance is defined as the relevance of a user's domain knowledge to the target question and a user's reputation is derived from the user's historical question-answering records, while user authority is derived from link analysis. In [19] and [20], the authors focus on temporal cues that contribute to expert identification and discuss their influence in community dynamics. One important finding reported is that the temporal cues based method outperforms user statistics based ones.

While studies involving the analysis of users in a community are well represented when dealing with social media websites, such as Twitter or Facebook or blogs, there are few works that directly target users in image sharing platforms. Since its early years, Flickr has been primarily used as a playground for identifying the motivations and scope of users for tagging their images. Although we can not use directly the findings reported in these types of studies, they offer both important insights on how we can link previously defined user categories to credibility and a theoretical motivation for exploring different data sources when proposing context features. In a seminal work, Ames and Naaman [1] propose a taxonomy of motivations for annotation in this system along two dimensions (sociality and function), and explore the various factors that people consider when tagging their photos. They base their work on user interviews and other qualitative methods. The first dimension, *sociality*, relates to whether the tag's intended usage is by the individual who took and uploaded the photo or by others, including friends/family and strangers. The second dimension, *function* refers to a tag's intended uses. They found that users tagged their pictures either to facilitate later organization and retrieval or to communicate some additional context to viewers of the image (whether themselves or others). The *function* dimension focuses on the motivation for adding tags. In MUCKE, we look for user credibility indicators that can improve an image retrieval framework. This is why, when considering the *sociality* dimension, we are interested in discriminating between users that upload their images to Flickr with the goal of showing his contributions to the community and those that do it to have a backup of their photo collection or only show them to a limited set of people. Contributions from the first category of users are potentially more useful to be used for

general purpose retrieval than those coming from users falling under the second category. Similar, when looking at the *function* dimension, we are interested in identifying users that have as their main purpose of tagging the improvement of retrieval over their photo collection. In contrast to this type of users, we would like to be able to filter out the users that tag mostly to indicate the context in which the photo was taken (e.g. the year in which the photo was taken or the names of people attending a certain event) which would be relevant only to a small number of people. In [11], the authors also investigate the motivations behind user tagging and propose several incentives that can be used in an annotation framework. They argue that annotators gain widespread recognition and credibility by doing good work that can be used by many people, even if they are not receiving direct compensation for their work.

In our work, we view credibility as a mixture of user specific attributes, such as expertise and the quality of his contributions. In this sense, we share the observation made by Ye and Nov [24] who state that researchers need to take a user-centric approach to understanding the dynamics of content contribution in social computing environments. They study the connection between different aspects of a user's motivation and both the quantity and quality of his contributions. Their results indicate, among others, that users with more social ties, especially ties with people they have not met in the physical world, tend to contribute better content to the community.

2 A NEW USER TAGGING CREDIBILITY DATASET

The only available dataset that provides manual credibility estimations for Flickr users is the one introduced in the MediaEval Retrieving Diverse Social Images Benchmarking Initiative [7]. A specially designed dataset that addresses the estimation of user tagging credibility in the context of landmark is provided. It offers Flickr photo information (the date the photo was taken, tags , user's id and photo title etc.) for 685 different users. Each user is assigned a manual credibility score which is determined as the average relevance score of all the user's photos. To obtain these scores, 50,157 manual annotations were used, averaging 73 photos per user. We propose here a novel dataset, designed with the scope of analyzing user credibility for a diversified set of topics.

2.1 USER CREDIBILITY DATASET MOTIVATION

We describe here the process of creating a dataset tailored for the investigation of features that are potentially useful to hint a user's tagging credibility. We identify the following needed requirements for a dataset of this nature:

- It should contain contributions from a consistent number of different users. This allows the exploitation of the dataset both as a relevant collection on which correlations between synthetic features and manual credibility scores can be estimated but also leaves room for a learning scenario in which the credibility score can be predicted by a trained model. It should offer

enough training instances so that commonly used machine learning models are able to learn a pattern, if one would exist.

- Each user should have a significant number of contributions evaluated so that we could derive a reliable manual credibility score. This score will be obtained by averaging the relevance scores of individual contributions.

- Contributions sampled for each user should be images depicting a diverse set of topics. On the first hand, this choice is imposed by the nature of how we define the credibility score in an image tagging context. Our goal is to study a user's global credibility score. Having more than one topic represented for each user also promotes the re-usability of this dataset and enables studies on domain specific user credibility.

In practice, all of the desired features mentioned above are subject to limitation coming from the availability of data but mostly from the cost of annotation. As a result, when setting the targeted values for each of the three features, a trade-off between any of them has to me made. After a series of internal studies, we settled for the following approximate values: Around 1000 users, 50 images for each user and at least 5 topics represented in the contributions of each user. Next, we present the dataset annotation protocol and the dataset statistics.

2.2 DATASET CREATION

For the annotation effort, we follow a similar methodology as that proposed for the construction of the datasets used in the ImageCLEF Wikipedia retrieval evaluation campaigns [23]. For each topic, we present the annotator with a couple of relevant images and a narrative which has the purpose of clarifying what is relevant and what is not for each topic. For example, in the case of the *sunset* topic, we provide the following narrative: *Assume that you want to illustrate different aspects of sunset with images. Please select all images which are relevant for sunset from the list below. Diversified views or aspects of sunset are relevant.*. Then, for each topic we present a maximum of 300 images per page. The annotator's task is to select only the images he/she finds relevant for the given topic.

The relevance assessments of the images in the dataset were provided by a total of 6 trusted annotators (faculty members). Before starting the annotation process, the raters were first involved in a feedback loop. This entailed them expressing the ambiguities they identified in some topics and, from our side, modifying the narratives, where necessary.

We first fix a number of diverse but simple topics that have a clear visual representation. Our proposal diverges from recent image retrieval datasets that are either domain specific [8] or built for ad-hoc retrieval of complex topics [23] and is closer in terms of topic coverage to the original MIR Flickr collection[6]. Our target is obtaining a reliable score for user credibility. This entails having confident assessments of his images that depict easily recognizable topics.

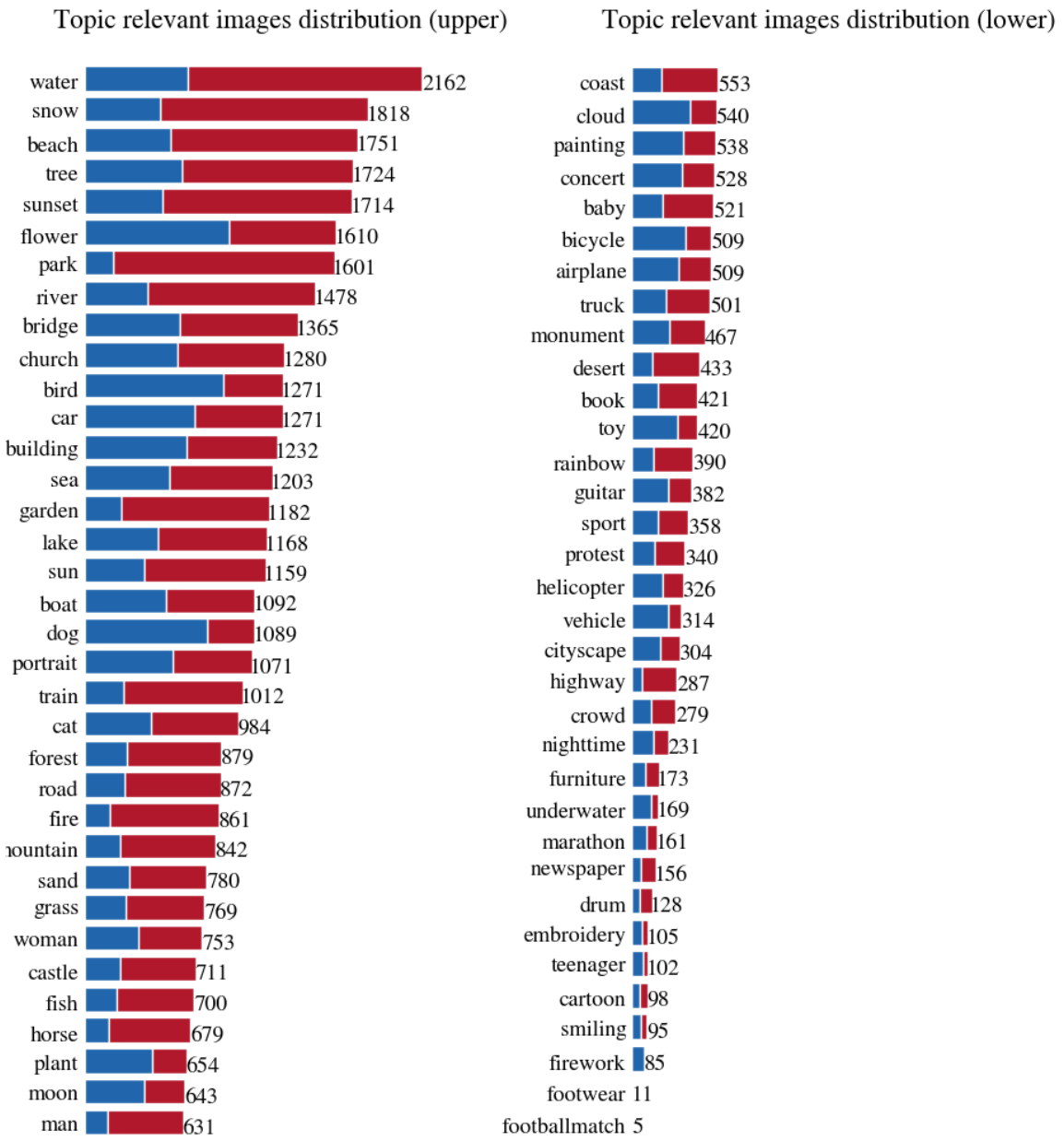

FIGURE 1: DISTRIBUTION OF RELEVANT AND NON RELEVANT IMAGES FOR EACH TOPIC

We use the Flickr API [1], to download both user and image metadata. We start with the *flickr.photos.search* function to download photo metadata for more than 90 topics. Then, we collect statistics on the users that have contributions among the retrieved set of images for all the topics and we retain the users with most images across topics. We keep the top 3000 users as candidates for the credibility dataset. For each of these users, we call the *flickr.people.getPhotos* function to gather metadata for the users' photos. We downloaded metadata for a maximum 10.000 images per user. Finally, we keep only the users that have at least 50 images covering at least 10 topics.

[1] http://www.flickr.com/services/api/

2.3 DATASET STATISTICS

Using the protocol described above, we obtain a dataset containing a total of **1009** users and 50,450 images evaluated for relevance covering 69 topics. Each user has exactly 50 images in the dataset. In Figure 1, we present the names of the topics we retained in the dataset, the number of images that were evaluated for each topic. We also show the distribution of positive and negative images for each topic. The blue bar represents the percentage of images found relevant by the raters and the red bar gives the percentage of non relevant images. We observe that some topics are very well represented (e.g. *water, snow, beach*), while other have fewer than 100 images (e.g. *firework, footwear, footballmatch*). We can also see from this figure that most of the images are rated as being non relevant to the queries. A few notable exceptions, where the relevant images are predominant are the *dog, plant, vehicle or firework* topics.

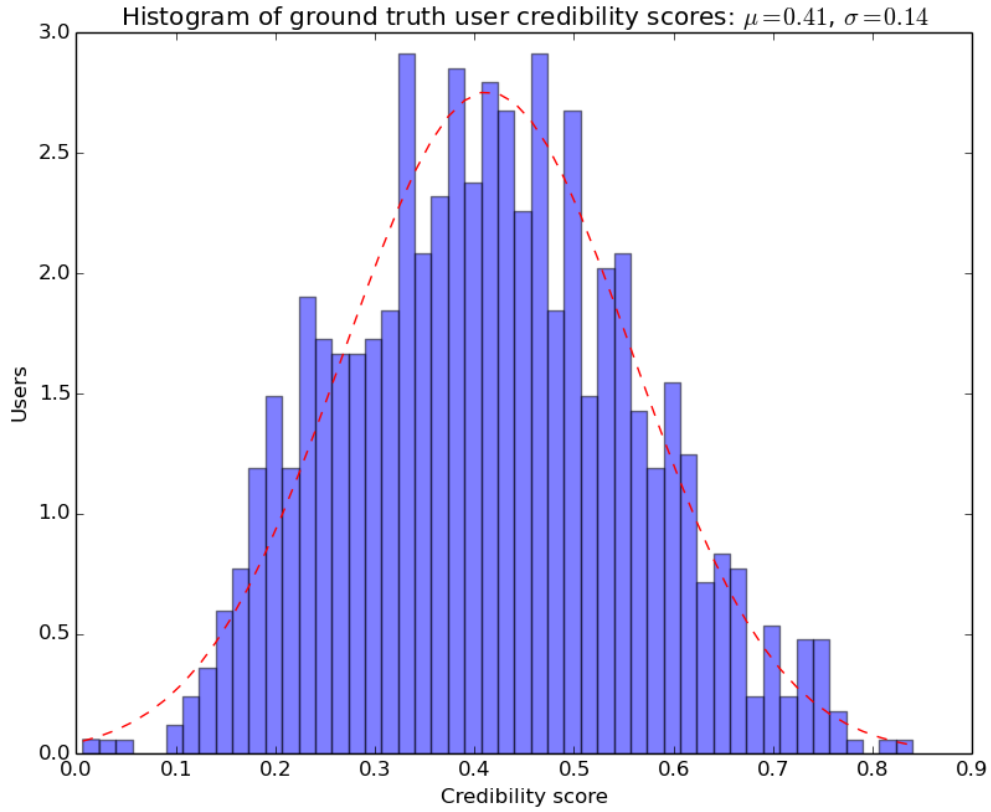

FIGURE 2: HISTOGRAM OF MANUAL CREDIBILITY SCORES

As for the dataset introduced in [7], we build the manual user credibility scores by taking the percentage of images found relevant among the 50 images that were evaluated for each user. In Figure 2, we show the distribution of the manual credibility scores. We observe that the scores follow an approximate normal distribution. The fact that the majority of images are labeled as non relevant can also be observed in the distribution of credibility scores. The mean of the credibility scores is 0.41.

We observe the agreement between raters by measuring Randolph's free marginal multirater kappa score [21]. We use this method to evaluate agreement, as opposed to Fleiss' multirater kappa because we do not know a priori the quantities of cases that should be distributed into each category (relevant vs. non relevant images). We observe an agreement score of 0.581 when combining annotation for all the topics. This can be interpreted as moderate to high agreement. This score shows that although we took precautions to ensure a simple and clear annotation process, providing relevance ratings for a diverse set of topics remains a difficult task.

Topics with high agreement			**Topics with low agreement**		
Name	***Kappa***	***#Images***	***Name***	***Kappa***	***#Images***
fire	0.86	861	truck	0.337	501
man	0.854	631	teenager	0.359	102
cat	0.838	984	lake	0.375	1168
marathon	0.776	161	embroidery	0.377	105
rainbow	0.770	390	sea	0.379	1203
helicopter	0.762	326	building	0.393	1232
horse	0.756	679	boat	0.406	1092
vehicle	0.749	314	nighttime	0.411	231
castle	0.735	711	church	0.413	1280
baby	0.733	521	grass	0.422	769

TABLE 1: RANDOLPH'S FREE MARGINAL MULTIRATER KAPPA SCORE FOR INDIVIDUAL TOPICS.

In Table 1, we show the first 10 (left column) and last 10 (right column) topics ranked by Randolph's free marginal multirater kappa score for the relevance annotation of the images found in the topic. As expected, we notice high scores for some of the least ambiguous topics (e.g. *fire, man, cat*). Among the topics with low agreement scores, we find those that may present with some level of incertitude, such as *teenager or like* but also, surprisingly, topics that seem to have a clear visual representation, such as *boat or truck*.

3 CONTEXT FEATURES FOR CREDIBILITY

Alongside its main goal of photo storage and sharing, the Flickr platform provides it's users a considerable amount of means to organize their photo collection but also to interact among themselves. Besides tagging, users can group their photos in photosets and can add their photos to groups, which unite contributions for different users with common interests. Following a social network recipe, users can have contacts and are able to provide feedback on the photos of other members of the community through the means of favorites and comments. We are interested to exploit as

much of this data as possible to propose user features that may be good indicators for credibility. In the study we present here, we group the features in feature families, according to the nature of data from which they are extracted. Due to limitations imposed by the number of calls per day to Flickr APIs, we settle for the following feature families: *photo metadata, groups, photosets, given photo favorites and contacts*. All the experiments and analysis presented in this section can be easily extended to include features coming from any other data source, when they becomes available. Note that the features we extract can be explicit and may come from a user's direct action, such as adding a new contact or indirect or implicit, which we derive from a user's actions, such as the temporal aspect of his upload behavior.

3.1 DATA ACQUISITION

Flickr exposes a large number of APIs through which data about a user's contributions or his interactions with other members can be easily gathered. Due to API constraints, we limit the number of samples we download. This limit is specific to each feature family and is chosen so that we can obtain a significant sample of data for each user. We will now provide details about how we downloaded the collection from which we extract our proposed context features and give statistics on the number of metadata items for each feature family:

- *Photo metadata:* We use the *flickr.people.getPublicPhotos* function of the Flickr API to download metadata associated with a user's photos. For each photo we get data such as the title, tags, or when was the photo taken and when was it uploaded. We first make a request to retrieve the first page, from which we extract the tonal number of photos the user uploaded in Flickr. Then, we request 500 items per page for each API call and, when available, we download up to 20 pages for a user. Finally, for the 1009 users in our evaluation dataset, we end up with 10,540 metadata files.

- *Groups:* We use the *flickr.people.getPublicGroups* function to get the list of public groups a user is a member of. Note that we do not have access to a user's private groups. For each group, we retrieve its name, the number of users that are part of the group and the total number of photos that have been included in the group. For each user, we get a single file with the data about the public groups his part of, gathering a total of 1009 group metadata files.

- *Photosets:* Photosets metadata is retrieved by calling the *flickr.photosets.getList*. Similar to photo metadata, we first get the first page, with the number of total photosets and download a maximum of 10 pages, where each page contains data about at most 500 photosets. This leads to a total of 2,733 downloaded photosets metadata files.

- *Given Photo Favorites:* We retrieve the list of photos a user has marked as favorite by calling the *flickr.favorites.getPublicList* API function. We use the same procedure for retrieving the

data as that used for photosets and we get 6,337 files, each containing details about up to 500 photos the user has marked as favorite.

- *Contacts:* In Flickr, if a user has added another user as a contact, it does not imply that there will be reciprocity. The contact relationship is not symmetric. If user A designates user B as a contact, user A can see the photo stream of user B, but not vice versa. This makes the contact relationship closer to the follower structure in Twitter. The Flickr API provides access only to the contacts of a user but we can not retrieve a list of users that have the target user among their contacts. In order to be able to user network analysis methods, we crawl a subsample of the Flickr contacts network by recursively calling the Flickr API function *flickr.contacts.getPublicList*. We use a similar methodology for sampling the network as the one presented in [15] We start from the list of users in our dataset and we we the contacts up to a depth of 2 (i.e. contacts of a contact of the original user). In order to have a sample of contacts for all of our evaluation users in a reasonable amount of time, we impose a limit to API calls for second degree contacts. If a contact of the original user has more than 500 contacts, we retain only a sample of 500 for which, we download their contacts information. Using this approach, we obtain a contacts network comprised of 5,811,652 unique users and 91,205,141 links. To put these numbers in perspective, Cha et al. [4] estimate that a network including 2.5 million Flickr users and 33 million links, represents 25% of the entire Flickr network. This statement is made for the Flickr network as of the end of 2007. Newer data suggest that in 2014 there were around 92 million active users in Flickr[2].

Next, we describe for each feature family the list of features we extracted and the motivation behind them. This list is not exhaustive and, if proposed, other features can be easily added from the downloaded collection of data files presented above.

3.2 FEATURE EXTRACTION

Tags, alongside the actual image are the main content produced by a Flickr user. Here, we focus only on features that can be extracted from the context. In the case of a Flickr user, we consider the context built around his activity to encompass any action he performed or any action that concern him done by other users, except the act of tagging.

3.2.1 METADATA FEATURES

The metadata that accompany photos that a user uploads to Flickr represent the main source of information about a user's direct contributions in Flickr. Through its API, Flicker provides for each public photo, the associated tags, the title put by the uploader, the date it was uploaded on Flickr or the date when the photo was taken (if available). We exploit all of these sections and extract the following features:

[2]http://www.thesocialmediahat.com/active-users

Title related features. Users may choose a different title for one or small number of his photos or may use the same title for a large set of photos (e.g. all of the photos taken in the same trip). We hypothesize that a user who takes his time to provide a detailed topic for as many photos as possible, is more likely to provide contributions that are meant to be shared with the community. In the opposite case, when a user few titles for most of photos (i.e. usage of bulk titles), we may be facing with a user that only wants to store his photo collection mostly for personal usage. Besides bulk tags, we are also investigate the diversity of the vocabulary used in titles. A large diversity of title words may indicate a user who has either interest in a large number of topics or takes his photos in many different scenarios. Finally, we look at capitalized words found in titles. A high percentage of capitalized words may indicate for a user's photos a focus on locations or people. We extract the following tag related features:

- *title_bulk_percentage:* the percentage of titles that appear at least 3 times from the set of all the titles given by a user.
- *title_vocabulary_size:* the number of unique words used in the titles given by a user.
- *title_capitalized_words_percentage:* the percentage of capitalized words found in a user's title vocabulary.

Temporal features. A photo upload behavior uniformly distributed over time may be an indicator for a user's strong involvement in Flickr. Temporal data could contribute to separate casual users, that upload images from time to time from those that are more passionate about photography or even professionals. We propose the following time related features:

- *different_upload_days:* the number of unique days in which a user has uploaded at least one photo.
- *average_upload_time_delay_minutes:* the average time elapsed between two consecutive uploads, measured in minutes
- *average_upload_time_delay_days:* the average time elapsed between two consecutive uploads, measured in days. We look only at the number of days passed between the last upload of one day and the first upload of the next day in which an upload was made.
- *different_photo_taken_days:* the number of unique days in which a user has taken photo.
- *average_photo_taken_time_delay_minutes:* the average time elapsed between two consecutive photo taken timestamps, measured in minutes.
- *average_photo_taken_time_delay_days:* the average time elapsed between two consecutive photo taken timestamps, measured in days.
- *average_date_taken_upload_delay_hours:* the average time elapsed between the time a photo was taken and the time it was uploaded, measured in hours.

Photo related features. Ye and Nov [24] find that in Flickr, the quantity of a user's contributions is negatively associated with the quality of contributions. Besides this straightforward statistic, we also look at how many times user's photos have been seen by other members of the community. A user whose contributions receive increased attention from the community may be viewed as an expert photographer. We propose 3 features extracted from photo uploads and views statistics:

- *total_photos:* the number of photos a user has uploaded to Flickr.
- *avg_photo_views:* the number of times a user's photo has been viewed by other users on average.
- *photos_with_at_least_100_view_percent:* the percentage of photos that have been viewed at least 100 times. We propose this feature so that we would have an indicator for users that have a more uniform distribution of photo views. This counteracts the case in which there is a strongly skewed distribution of views which would lead to a high average from only few contributions.

3.2.2 GROUPS

In Flickr, users have the option to create groups that allow people who have similar interests to get together and share their photos. Flickr groups may form around users sharing a common interest (brands of cars, animals etc.), or they may gather images taken with a specific brand of camera or camera setting (black and white, light setting). It is also possible for a group to be created so that users coming from the same geographical location to share their contributions. In a pioneering work, Negoescu et al. [16] looked at the involvement of users in groups and found, among others, that user group loyalty is generally low and most users share the same photos in different groups. We are more interested on what we could infer about an user from the groups he is part of. For instance, a user that is part in many groups, may be more motivated to share high quality content than a user that prefers to keep his photos only in his collection. We look at the number of groups a user belongs to but also at the nature of those groups. To summarize, we extract the following features from group data:

- *groups_count:* number of groups a user is part of.
- *avg_groups_members:* the average number of members of the groups the user belongs to. A low membership may indicate more specialized groups or groups that limit the number of members. We assume that a user that belongs to many of this types of groups may suggest a higher level of expertise.
- *avg_groups_photos:* the average numbers of photos found in the groups the user belongs to. As the number of users, we consider this feature to be a possible indicator for a group's level of specificity.

3.2.3 PHOTOSETS

Flickr users can organize their images in photosets either at upload time or later, by selected a list of images to be grouped. When kept private, photosets serve the purpose to improve the organization of a user's personal collection. Public photosets give hints on a user's interest to group his photos for the indirect benefit of other members of the community. A user can receive feedback for his photosets through comments given by other users. We propose the following features from photosets data:

- *total_photosets:* the total number of photosets created by a user.
- *photosets_avg_views:* the average number of times a photoset was viewed by other members of the community.
- *photosets_avg_comments:* the average number of comments given for a photoset by other members of the community.

3.2.4 GIVEN PHOTO FAVORITES

A user can show his appreciation for photos of other members my marking them as favorites. We see this as another indicator for the user's involvement in the Flickr community. Here, we look only at the number of photos a user has favorited and metadata associated to those photos. Although the number a favorites a user receives for his contributions may serve as a feature for credibility, Flickr does not include this information in the photos metadata and a separate API call is required for each photo individually. This renders it impractical for the immediate scope of this work. We propose the following features:

Metadata features. We include here photo, user and title words counts.

- *total_favorited_photos:* the number of photos a user marked as his favorites.
- *%_unique_users_favorited:* the number of unique user for which the target user has favorited at least one photo divided by the total number of given favorites. Through this feature, we want to differentiate users that have a narrow circle of ties in the community for which they give favorites from those that give favorites to a more diverse set of users.
- *%_unique_words_int_favorited_titles:* the percentage of unique words found in the titles of the photos the user marked as favorites. This feature can be seen as a signal for the diversity of topics a user is interested in.

Temporal features. Sharing the same motivation as that behind proposing temporal features for a user's uploads, we consider the distribution over time of a user's favorites as a clue for his engagement in the community.

- *different_favorited_days:* the number of different days in which a user has marked as favorite at least one photo.

- *average_favorited_time_delay_days:* the average time elapsed between two consecutive given favorites, measured in days.

- *average_favorited_time_delay_minutes:* the average time elapsed between two consecutive given favorites, measured in minutes.

3.2.5 CONTACTS

Starting from the sample of the Flickr network we introduced in Section 3.2, besides number of contacts a user has and the number of other members that have the user among their contacts, we also investigate the use of well established link analysis algorithms, such as PageRank and HITS for estimating the credibility of Flickr users.

In the original PageRank algorithm [18], for improving the ranking of search query results, a single PageRank vector is computed, using the link structure of the Web, to capture the relative importance of Web pages. Besides Web pages, PageRank has been used to analyze users in networks where there is a unidirectional relationships between users, such as the *follower* relationship in Twitter [12]. Considering that link between two Flickr contacts is also unidirectional, we extract the PageRank score of the users in our evaluation dataset.

The HITS algorithm was developed by Kleinberg [10] and proposes the premise that web pages serve two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic. HITS algorithm mines the link structure of the Web and discovers the thematically related Web communities that consist of *authorities* and *hubs*. As described in [17], authorities are the central Web pages in the context of particular query topics. For a wide range of topics, the strongest authorities consciously do not link to one another. Thus, they can only be connected by an intermediate layer of relatively anonymous hub pages, which link in a correlated way to a thematically related set of authorities. Similar to PageRank, HITS has been used beyond the scope of Web pages and can serve as a method of identifying experts in online question answering communities [2] or influencers in Twitter [9]. Here, we use HITS in a similar fashion, with the goal of finding both influential users and hubs in their Flickr contact network. For a user in our dataset, we extract his HITS metrics but also statistics on the HITS scores of his immediate contacts.

In Figure 3, we give an example of a user's subgraph. We retain a first set of users containg the the user *12285897@N00* from our credibility dataset and all of his contacts. Then, we select the users that have at least one user from that set among their contacts. For visualization purposes, we keep only a random subsample of the nodes, making sure to include the original user. Nodes are colored in respect to their HITS authority score. The darkest the color, the higher the authority score of that user. Similar, the size of the labels representing users' Flickr ids are proportional to the authority score. Having the most outgoing links, the user *12285897@N00* has the highest hub score in this subgraph. We noticed that this observation does not hold for all the users in our dataset.

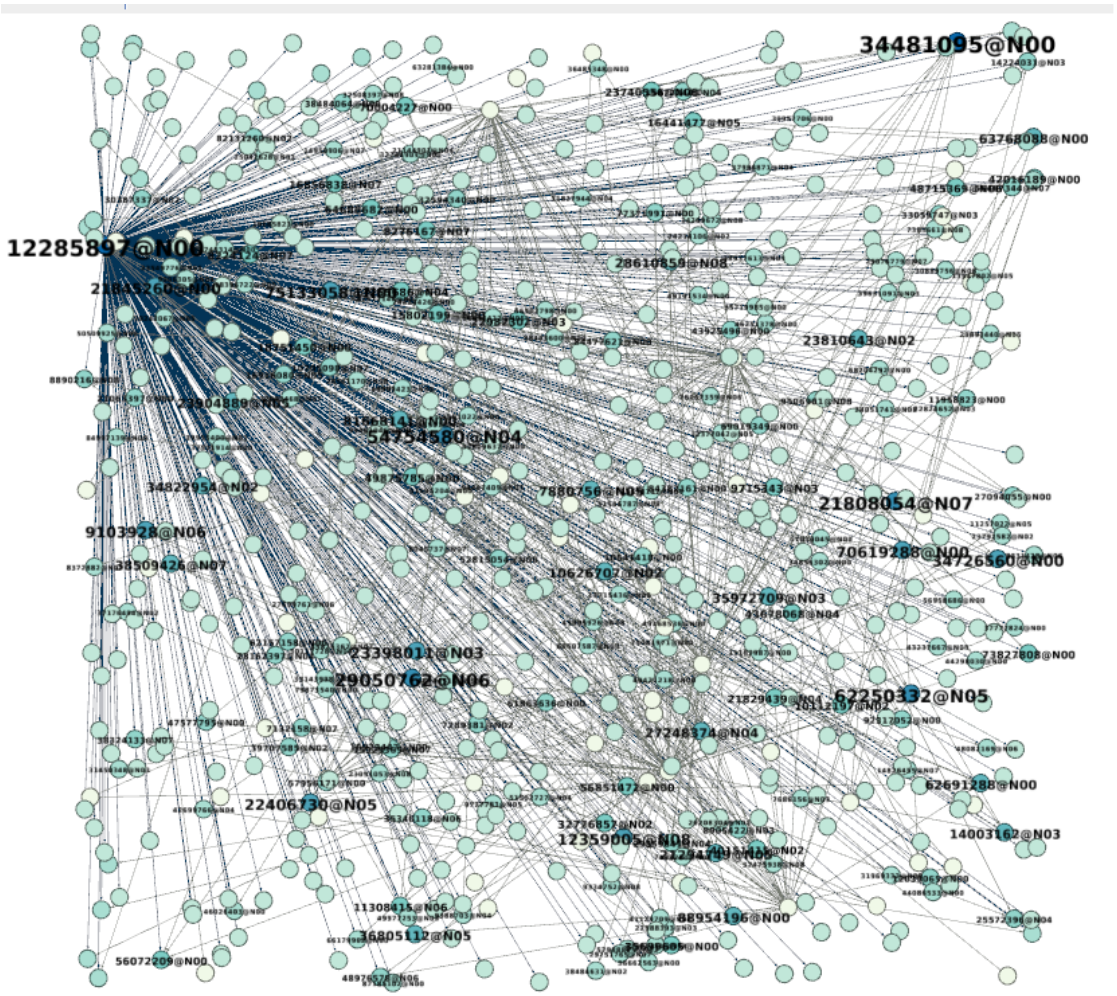

FIGURE 3: EXAMPLE OF A USER'S CONTACTS SUBGRAPH. NODE COLORS AND LABEL SIZES ARE PROPORTIONAL TO THE HITS AUTHORITY SCORE.

By analyzing subgraphs that were extracted by starting from other users, we discovered multiple nodes having higher hub scores than the original user. When looking at the authority scores, it is even more obvious that the bias towards the users in our dataset that we introduced collecting the contacts data does not have a strong impact on the link analysis methods. In the upper right corner of Figure 3, we can see a user with a higher authority score than the one used to generate the network subsample. These observations hint that we can obtain reliable network metrics even if we favor a set of users when collecting contacts data. Next, we briefly present the set of features we extract from contacts network data.

User network metrics features. This is a straightforward set of features, containing the network metrics computed directly for a target user.

- *user_in_degree:* the number of users that have the target user among their contacts.
- *user_out_degree:* the number of contacts of the user.
- *user_authority:* the HITS authority score of the user.

- *user_hub:* the HITS hub score of the user.
- *user_pagerank:* the PageRank score of the user.

Contacts network metrics features. Although when computing link analysis metrics for a users, his contacts are implicitly taken into consideration, we propose a subset of features that directly target a user's contacts. We chose this approach so that we can have a more detailed analysis of network features for credibility estimation, considering that we fully download data only for the set of immediate contacts of a user.

- *avg_contacts_in_degree:* the average *in_degree* of the user's contacts.
- *avg_contacts_out_degree:* the average *out_degree* of the user's contacts.
- *avg_contacts_authority:* the average HITS authority score of the user's contacts.
- *avg_contacts_hub:* the average HITS hub score of the user's contacts.
- *avg_contacts_pagerank:* the average PageRank score of the user's contacts.

3.3 FEATURE ANALYSIS

We evaluate the the features described in the previous sections by looking at how well they correlate with the manual credibility scores introduced in Section 2. We use Spearman's rank correlation for this purpose. The choice of Spearman correlation over Pearson is justified by our final goal of comparing an user ranking given my the manual credibility score to one dictated by a user feature and not necessarily to test if there is a linear relationship between the credibility scores and the features.

By looking at Table 2, we can draw the general conclusion that all of the proposed features are poorly correlated with the manual credibility scores. However, when comparing features, we observe that some of our suppositions made in the previous section are confirmed. Surprisingly, the strongest indicators for credibility are two of the photosets features (*photosets_avg_comments* and *photosets_avg_views*). Both features reveal the attention a user's contributions receive from other members of the community indicating that there is a weak positive correlation between the popularity of a user's photosets and quality of a users contributions. Notice that while the number of views of a user's photosets is the second most relevant feature, while the number of photo views has close to zero correlation with credibility. This may be explained by the fact that it is unlikely for a user with a large number of photos to have many views for the majority of them. On the contrary, a user has much fewer datasets. Datasets are also made with more consideration from the user and reflect his intention to provide curated content.

Feature name	Spearman	Feature name	Spearman
photosets_avg_comments	0.266	avg_date_taken_upload_delay_hours	0.063
photosets_avg_views	0.202	avg_contacts_out_degree	0.053
different_upload_days	0.166	avg_photo_views	0.053
different_favorited_days	0.161	user_hub	0.049
avg_upload_time_delay_minutes	0.149	photos_with_at_least_100_view_percent	0.033
total_photosets	0.116	title_vocabulary_size	0.017
avg_contacts_hub	0.114	user_out_degree	0.005
avg_contacts_in_degree	0.105	user_pagerank	0.005
avg_contacts_authority	0.105	avg_photo_taken_time_delay_days	-0.02
user_authority	0.102	%_unique_words_int_favorited_titles	-0.054
user_in_degree	0.102	avg_favorited_time_delay_minutes	-0.059
avg_photo_taken_time_delay_minutes	0.093	%_unique_users_favorited	-0.059
avg_contacts_pagerank	0.092	total_photos	-0.084
groups_count	0.092	avg_upload_time_delay_days	-0.093
total_favorited_photos	0.091	title_capitalized_words_percentage	-0.095
different_photo_taken_days	0.078	avg_favorited_time_delay_days	-0.099
avg_groups_members	0.076	avg_title_word_counts	-0.102
avg_groups_photos	0.069	title_bulk_percentage	-0.114

TABLE 2: SPEARMAN CORRELATION BETWEEN THE PROPOSED FEATURES AND THE GROUND TRUTH CREDIBILITY SCORES.

While we can not give a definitive statement, our assumption made in Section 3.2 about the bulk percentage among photo titles being a good indicator for low credibility is partially supported. Although the correlation between *title_bulk_percentage* and the manual credibility scores has a low absolute value, it still presents the highest inverse correlation among all of the proposed features. In the same register as the results reported by Ye and Nov [24], who find a negative correlation between the quantity and quality of a user's contributions, we a observe a negative correlation between the *total_photos* feature and the manual credibility score. Although negative, the correlation score is very small, falling close to indicating no correlation. Surprisingly, none of the proposed contact features seem to be strongly related to credibility. Except the number of contacts, the rest of the features are extracted from a sample of the Flickr network. In spite of the fact that we tried to minimize the impact of this shortcoming, without access to the full Flickr contacts network, we cannot give a final conclusion on the usefulness of contacts features. Nevertheless, most of these features are close together in the upper half of the ranked feature list presented in Table 2. Temporal features confirm our suppositions about the link between the time spent by a users adding contributions in Flickr, either uploading images or giving favorites to other user's photos, and the quality of his contributions. We observe a positive correlation for the features referring to the number of different days a user has been active in Flickr (*different_upload_days* and *different_favorited_days*) and a negative correlation for features that relate to the length of pauses between contributions (e.g. *avg_upload_time_delay_days*, *avg_favorited_time_delay_days*).

4 CREDIBILITY SCORE PREDICTION

4.1 PROBLEM DEFINITION

In most of the works that deal with predicting credibility, such as the credibility of tweets [3], credibility is viewed as a classification problem. In those scenarios, two (credible / not credible) or several credibility classes are formed. In the MUCKE framework, user credibility estimates are used to filter or rerank a list of retrieved items according to the users who produced them. We are interested in a fine-grained score that will allow us to rank users based on their credibility estimates. Considering that we have a single list of ground truth credibility scores, learning to rank approaches are not feasible. Given this limitation, we are not able to directly predict a ranking of users. In order to by-pass it, we treat the credibility score as a continuous variable (Y) and propose a regression problem in which we fit a model that learns to approximate the credibility score:

$$Y \approx f(X, \beta)$$

, where X is the feature vector and β are the model weights. We then used the regression model to predict credibility scores and rank users according to the predictions.

Unlike classical regression problems, our final goal is not to provide an approximation of the credibility score but to rank users according to their credibility estimates. This makes evaluation metrics usually used in regression problems(e.g. the mean squared root error) uninformative for our specific task. We directly evaluate the ranking obtained from the predicted scores to that given by the manual credibility scores. Following a similar procedure used to evaluate individual features, we use the Spearman rank correlation measure to test a new ranking. When comparing multiple classifiers, we are interested in the one that maximizes the correlation between the manual rank and the predicted rank:

$$\underset{m}{\operatorname{argmax}}\, Spearman(Y^m_{pred}, Y_{man})$$

Y^m_{pred} is the prediction vector corresponding to model m. The same evaluation measure is used when comparing different feature subsets selected to train the same classifier.

4.2 EXPERIMENTS

When building the training set for the regression experiments, we encounter a few cases of missing values. These may be caused by technical problems or bad responses from the Flickr API, by an user who removed are made private a part of his data. We first address this issue by imputing the missing value using the mean of value of the respective feature. Missing values account for less than 1% of our data. Due to large differences of magnitude between features, we then perform a L2 normalization. Although this does not affect ensemble models, it has a strong impact on the ability to learn of linear models. For predicting the credibility score, we test 9 models coming from 3 families of approaches:

- *linear models:* Linear Regression, Ridge (linear least squares with L2 regularization), Lasso (linear Model trained with L1 prior as regularizer), Elastic Net (linear regression with combined L1 and L2 priors as regularizer), Lars (Least Angle Regression model).
- *support vector machines:* SVR (epsilon-Support Vector Regression with rbf kernel)
- *ensemble models:* Extra Trees Regressor, Random Forest Regressor, Gradient Boosting Regressor

Due to the small size of our dataset, we evaluate each model in a leave-one-out-cross-validation (LOOCV) fashion. We select each time a different user and train a model on the remaining users. We do this for all the users in our evaluation dataset, keeping the prediction for the test user. Finally, we compare the predictions vector to the manual credibility scores. For each model, we tune the parameters on a randomly selected validation set in which we put 10% of the users in our dataset. We consider as baseline the best individual feature in terms of Spearman correlation.

Model	**Feature Family**					
	Metadata	*Contacts*	*Favorites*	*Groups*	*Photosets*	*All*
Linear Regression	0.144	0.107	0.116	-0.064	0.192	0.149
Ridge	0.145	0.108	0.116	-0.064	0.192	0.15
Lasso	0.151	0.105	0.126	-0.123	0.177	0.188
Elastic Net	0.145	0.108	0.116	-0.064	0.192	0.15
Lars	0.146	0.075	0.11	-0.064	0.192	0.138
SVR	0.161	0.102	0.109	031	0.179	0.178
Extra Trees Regressor	0.152	0.03	0.092	0.052	0.166	0.321
Random Forest Regressor	0.164	0.053	0.071	0.031	0.23	0.326
Gradient Boosting Regressor	0.184	0.056	0.1	0.039	0.283	**0.345**

TABLE 3: MODEL AND TRAINING SET COMPARISON FOR PREDICTING A USER CREDIBILITY SCORE. RESULTS ARE REPORTED IN TERMS OF THE SPEARMAN CORRELATION BETWEEN THE PREDICTED SCORES AND THE MANUAL CREDIBILITY SCORES.

Looking at the results presented in Table 3, we can observe that ensemble models outperform all other models regardless of the features used for training. In fact, they are the only one that manage to rise above the baseline. In a recent paper comparing 179 classifiers from 17 families over 121 datasets [5], Fernandez-Delgado et al. find that the family of features that gives the best performances on average over all the datasets is the ensemble family and the best individual classifier is Random Forests. We obse The best configuration is given when using a Gradient Boosting Regressor (GBR) model trained on the full feature set. In this case, we observe a 30% relative improvement over the best individual feature. We observe a similar behavior, with the exception that in our case, Random Forest is the best performing model. When comparing individual feature families, we observe that

only when training a model on photosets features, we obtain a correlation score higher than the one given by the best individual features. This result is not surprising, considering that the best features comes from this feature family. In total, we get only four configurations in which the baseline correlation score is surpassed. Although in the previous section we saw that individual features are poorly correlated with the manual credibility scores, we are able to learn a regression model that clearly offers a better credibility estimate than any of the features. This result validates the use of regression models for predicting a score which is later used for ranking.

4.3 FEATURE IMPORTANCE

In Section 3.3, we looked at the correlation between individual features and the manual credibility scores. We propose here two other methods for analyzing the usefulness of features in estimating credibility scores. The first one is given by a property of tree ensemble methods in which the depth of a feature used as a decision node in a tree can be used to assess the relative importance of that feature with respect to the predictability of the target variable. We extract the feature importance from the learned GBR model. The second one represents a feature ranking given by the weights leaned by a linear model. For this we chose Lasso, the best performing linear model (i.e. Lasso)

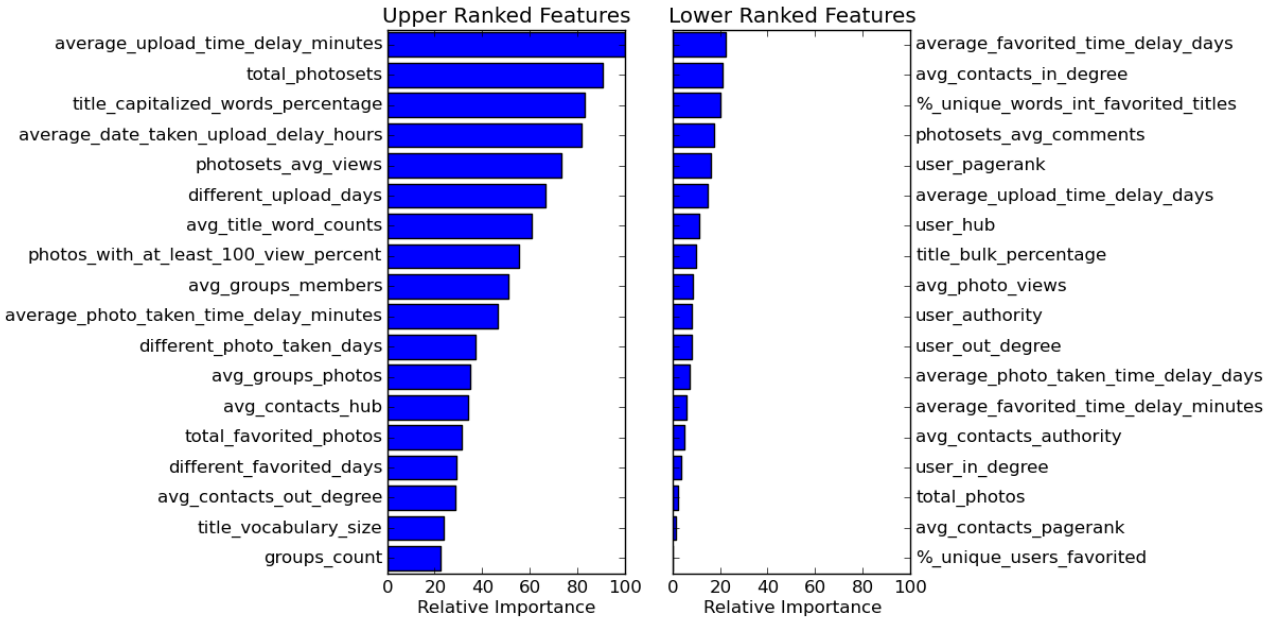

FIGURE 4: FEATURES RANKED ACCORDING TO THE FEATURE IMPORTANCE SCORE PROVIDED BY THE GBR MODEL.

In Figure 4, we see the ranked list of features according to their role in training the GBR model. We normalize the importance scores by giving the most important feature a score of 100 and then relating other feature to it. We observe that only 9 features out of 35 have an importance score higher than 50% of the score associated with the best feature. This indicates that selecting only top features for training may improve the predicted scores. We compare this ranking to the one introduced by the Spearman correlation score, shown in Table 2. Although the two rankings share

some similarities, we also observe a few big differences. We first notice that in the GBR feature importance ranking, the *avg_upload_time_delay_minutes* has the highest value, whereas in the Spearman ranking it was placed fourth. Also, photosets features play a lesser role in the model's decision than when directly looking at the correlation with the manual credibility scores. Similar to the results presented in Table 2, contacts features prove to be less relevant for estimating credibility.

In Figure 5, we can observe the impact of feature selection on prediction performances. We chose to test best performing ensemble model (i.e. GBR) and the best linear model. We train each of these models with top k ranked features according to each of the three feature ranking methods introduced above. We test with k ranging from 1 to 36 (the complete set of features). This way, we obtain 6 different configurations.

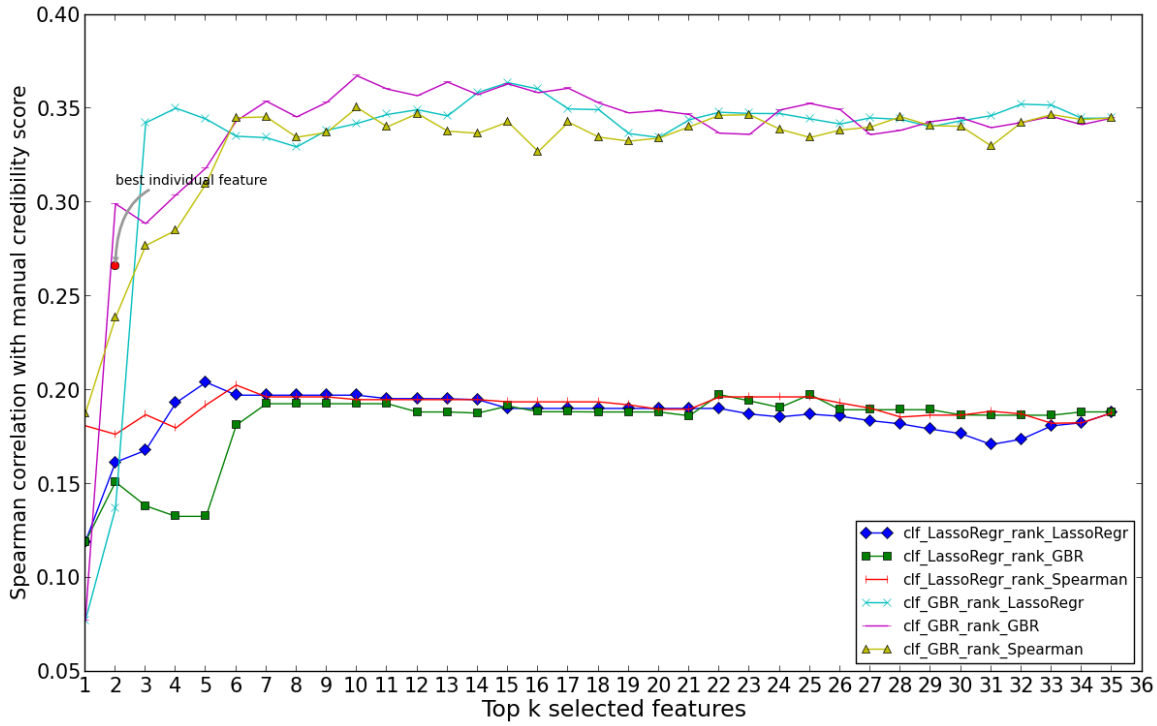

FIGURE 5: IMPACT OF FEATURES RANKING METHODS ON MODEL LEARNING.

We first observe that ensemble methods are better than the linear ones at almost any feature cut-off point. Also, as expected, the GBR model suffers from higher variability. An interesting result is that we can pass surpass the best individual feature by only training a GBR model with a couple of features (the first two features according to the Spearman ranking and the first three features according the the GBR or Lasso rankings). Feature selection also helps improving the best Spearman score obtained by a classifier trained on all features (0.345). There are several configurations that score higher and the best one achieves a correlation score of **0.367**. This is obtained by a GBR model trained on the top 10 features ranked according to the GBR feature importance scores.

5 CONCLUSION

In this work, we investigated the use of context features in estimating the credibility of Flickr users. We mined the features from various data sources in which a Flickr user has contributions, such as Flickr groups, photo favorites, a user's photosets or a user's contacts network. For the set of extracted features, we described the data acquisition process and tested their usefulness as individual credibility estimators. We also defined a credibility prediction problem, in which we learn regression models that provide better credibility estimators than the individual features. We find that, although individual context features are weak indicators for credibility, by choosing the appropriate regression model and the right set of features for training we are able to predict a credibility score that has considerably better correlated to the manual credibility score than any of the individual features.

All experiments were performed on a new dataset specifically built to help us evaluate potential indicators for credibility but also to serve as a training dataset on which we can compare multiple learning models and features. We presented the motivation behind the need for such a dataset and our methodology used for the creation of the dataset.

References

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980. ACM, 2007.

[2] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, pages 51–60. ACM, 2009.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[4] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009.

[5] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[6] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[7] B. Ionescu, A. Popescu, M. Lupu, A. L. Gınsca, and H. Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[8] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, and B. Loni. Div400: a social image retrieval result diversification dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 29–34. ACM, 2014.

[9] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[11] J. Kustanowitz and B. Shneiderman. Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[13] D.-R. Liu, Y.-H. Chen, W.-C. Kao, and H.-W. Wang. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites. *Information Processing & Management*, 49(1):312–329, 2013.

[14] X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 315–316. ACM, 2005.

[15] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *Proceedings of the first workshop on Online social networks*, pages 25–30. ACM, 2008.

[16] R. A. Negoescu and D. Gatica-Perez. Analyzing flickr groups. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 417–426. ACM, 2008.

[17] S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida. Analysis and improvement of hits algorithm for detecting web communities. *Systems and Computers in Japan*, 35(13):32–42, 2004.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[19] A. Pal, S. Chang, and J. A. Konstan. Evolution of experts in question answering communities. In *ICWSM*, 2012.

[20] A. Pal, R. Farzan, J. A. Konstan, and R. E. Kraut. Early detection of potential experts in question answering communities. In *User Modeling, Adaption and Personalization*, pages 231–242. Springer, 2011.

[21] J. J. Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online Submission*, 2005.

[22] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[23] T. Tsikrika, J. Kludas, and A. Popescu. Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. *IEEE MultiMedia*, 19(3):0024, 2012.

[24] C. Ye and O. Nov. Exploring user contributed information in social computing systems: quantity versus quality. *Online Information Review*, 37(5):752–770, 2013.