A Simple Language Independent Approach for Distinguishing Individuals on Social Media

Guangyuan Piao
Department of Computer Science, Maynooth University
Maynooth, Co Kildare, Ireland
guangyuan.piao@mu.ie

ABSTRACT
Nowadays, the large-scale human activity traces on social media platforms such as Twitter provide new opportunities for various research areas such as mining user interests, understanding user behaviors, or conducting social science studies in a large scale. However, social media platforms contain not only individual accounts but also other accounts that are associated with non-individuals such as organizations or brands. Therefore, distinguishing individuals out of all accounts is crucial when we conduct research such as understanding human behavior based on data retrieved from those platforms. In this paper, we propose a language-independent approach for distinguishing individuals from non-individuals with the focus on leveraging their profile images, which has not been explored in previous studies. Extensive experiments on two datasets show that our proposed approach can provide competitive performance with state-of-the-art language-dependent methods, and outperforms alternative language-independent ones.

CCS CONCEPTS
• Human-centered computing → Social media; Social networking sites; • Computing methodologies → Supervised learning by classification.

KEYWORDS
Account Classification, Deep Learning, Social Media Analysis

ACM Reference Format:

1 INTRODUCTION
Social media platforms such as Twitter have been widely used in different research areas to study users from various perspectives in a large scale based on the big data generated by user activities on those platforms. For example, research areas such as predicting substance usage [10], mining user interests [14, 15], and understanding user visiting behaviors [16] or personalities [7]. Although those studies usually assume all accounts retrieved from social media platforms are individuals, McCorriston et al. estimated that 9.4% of accounts on Twitter are non-individual ones (e.g., brands or organizations) [11]. Therefore, distinguishing individual users from retrieved accounts on social media platforms is crucial for studying different user behaviors such as mining user interests or understanding substance usage, and deriving conclusions out of those studies.

Previous studies for distinguishing individuals from non-individuals can be classified into two categories based on whether an approach is language dependent or independent. For example, language-dependent approaches utilize textual information, such as social posts and/or the profile description (biography) of a user in addition to a set of statistical features, e.g., the number of followers and/or followers on Twitter, to classify whether a given account belongs to an individual. As one might expect, the profile description of an account can provide crucial information to distinguish individuals. For example, we can assume that an account belongs to an individual if its profile description contains words such as "my", "I", "Dad", "Mum", etc. However, this line of approaches depends on language and the majority of the previous works have been focused on English users. Although English is the most popular language on Twitter, it is used in only 32% of all Twitter messages\(^2\). In contrast to relying on textual information, recent studies [3, 4] have proposed leveraging statistical features for classifying accounts on social media platforms such as Twitter.

Our focus in this paper falls into the second category, i.e., language-independent approaches for classifying individual accounts. To this end, we leverage the visual content of a user (i.e., profile image), which is critical information but has not been explored in previous studies. The intuition behind our approach is that the profile image of a user should be a good indicator for the classification of accounts. Our main contributions include:

- We propose a simple Language-Independent Individual Classification approach (Section 3), named LIIC, to classify social media accounts into individuals and non-individuals, with the focus on leveraging their profile images.

- We evaluate our approach with several state-of-the-art approaches using two datasets with ground truth labels in Section 4, and show that LIIC can achieve competitive performance in classifying Twitter accounts compared to

---

\(^1\)https://twitter.com/home

\(^2\)shorturl.at/sTVZ9
language-dependent approaches.

• Through an ablation study in Section 5, we further reveal that profile images are indeed an important indicator for classifying individual user accounts, which have not been explored in previous studies.

2 RELATED WORK

In this section, we review related works which are classified into language-dependent and language-independent ones.

Language-dependent approaches. This line of approaches exploits textual content such as social posts or profile descriptions of users for feature engineering or learning latent representations via deep learning approaches [5, 20, 21]. For example, Oentaryo et al. used content, social, and temporal features and investigated several machine learning approaches such as random forests and gradient boosting [6], and showed that the gradient boosting classifier provides the best performance [13]. Wood-Doughty et al. proposed using a character-based Convolutional Neural Network (CNN) [9] to learn the representation of a user’s name and incorporated profile features such as the ratio of followers to friends together for classifying individual accounts [20].

Language-independent approaches. This line of approaches uses statistical features such as the social network structure and the posting frequency of a user without relying on textual content for classifying individual accounts [3, 4, 18, 19]. For example, Tavares et al. used a naive Bayes classifier with features related to the time distribution between social posts [18]. More recently, Daoudi et al. proposed a set of comprehensive features incorporating profile and activity related ones, and used gradient boosting regression trees and random forests for classifying individual accounts on Twitter [3, 4].

Despite the appreciable body of previous studies, the profile image of an account, which might be a critical indicator for the classification, has not been explored. In this work, we close the gap and focus on leveraging profile images for classifying individuals and use other profile related features only if those images are not retrievable or are default ones. Our approach can be considered as one of the language-independent approaches as this approach does not require textual content such as social posts or profile descriptions.

3 LIIC: LANGUAGE INDEPENDENT INDIVIDUAL CLASSIFIER

In this section, we introduce our Language Independent Individual Classifier (LIIC) and its components in Section 3.1, and provide the training details of LIIC in Section 3.2.

3.1 LIIC Architecture

Figure 1 illustrates an overview of the LIIC architecture. LIIC uses three types of input such as the profile image, screen name, and profile features of a user, and leverages three different types of neural networks to learn the representations of each input for the binary classification of the target account (individual or non-individual).

The main assumption of LIIC is that the profile image of an account should be a good indicator for distinguishing individuals.
embeddings. Finally, \( v_f \) is fed into a dense layer to output the final representation \( v_s \in \mathbb{R}^{128} \) for the given screen name as follows:

\[
v_s = f(W^{(f)}v_f + b^{(f)})
\]

where \( W^{(f)} \) and \( b^{(f)} \) are a weight matrix and a bias term.

**Profile feature representation.** Finally, we also extract a set of features (13 in total) from user profile information which can be retrieved via the Twitter API. Table 1 shows the description about those features. As some of the features such as the number of followers or friends can have different scales compared to other features which can make the training of our model difficult, we first scaled those features using a logarithmic transformation, i.e., \( v_{k'} = \log_{10}(v_k) \). Here, \( v_k \in \mathbb{R}^{13} \) denotes the initial values of those features. \( v_{k'} \) is then used as an input to a dense layer to output \( v_f \in \mathbb{R}^{128} \) as shown below using Equation 4.

\[
v_f = f(W^{(k')}v_{k'} + b^{(k')})
\]

where \( W^{(k')} \) and \( b^{(k')} \) are a weight matrix and a bias term.

**Concatenation.** After obtaining the latent representations of the screen name and the set of features of a user (i.e., \( v_s \) and \( v_f \)), those two representations are concatenated together into \( v_c \in \mathbb{R}^{256} \) as follows.

\[
v_c = [v_s, v_f]
\]

**Fusion.** As our intuition is that the profile image of a user account is critical for distinguishing individuals from non-individuals, we deliberately pay full attention for the profile image information if possible, and only use the other information (i.e., screen names and profile features) when the profile image has not been changed from the default one or the image cannot be retrieved via the given image URL from the Twitter API. This can be formulated as follows:

\[
v_{\text{fused}} = \alpha \cdot v_p + (1 - \alpha) \cdot v_c
\]

\[
\alpha = \begin{cases} 0, & \text{if default profile image or not available} \\ 1, & \text{otherwise} \end{cases}
\]

where \( v_{\text{fused}} \in \mathbb{R}^{256} \) is the final vector that is used for predicting the score of being classified as an individual.

**Prediction.** Finally, a prediction score \( \hat{y} \) is calculated using Equation 8, and the target account can be classified as 1 (individual) if \( \hat{y} \geq 0.5 \) or 0 otherwise,

\[
\hat{y} = \sigma(W^{(\text{fused})}v_{\text{fused}} + b_{\text{fused}})
\]

where \( \sigma \) denotes the sigmoid function \( \sigma(x) = \frac{1}{1 + e^{-x}} \), and \( W^{(\text{fused})} \) and \( b_{\text{fused}} \) are a weight matrix and a bias term.

### 3.2 Training

We used Tensorflow 2.3.0 to implement LIIC. For training LIIC, we use the Adam update rule [8] and a batch size of 1,024 to train the model on the training set with a learning rate of 0.001 to minimize the binary cross entropy loss \( \mathcal{L} \), which is defined as follows.

\[
\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)]
\]

where \( y_i \) and \( \hat{y}_i \) denote the ground truth and predicted labels for \( i \)-th instance, respectively. \( N \) refers to the total number of instances for training.

To resolve the overfitting problem, we divided the training set and used 1/3 of it as our validation set, and used the rest of the data for training. An early stopping strategy is adopted with the validation set, which stops the training if there is no improvement of accuracy on the validation set.

### 4 EVALUATION

In this section, we discuss the datasets (Section 4.1) and the set of compared methods (Section 4.2) for evaluating LIIC followed by the experimental results in Section 4.3.

#### 4.1 Datasets

Here, we describe the training set used for training LIIC and two other datasets for testing. Table 2 shows the statistics about the three datasets.

The training dataset is from Wood-Doughty et al. [20]. Instead of manual labeling of individuals and non-individuals, the authors constructed the dataset using an automated method based on weak supervision for the discovery and labeling of those accounts. For example, the authors in [20] identified Twitter lists\(^*\) containing non-individual accounts which include terms such as "companies" or "businesses", and those containing individual accounts which include key terms such as "friends" and "families" using a search engine. By using this approach, a large dataset can be obtained without manual labeling effort which is time-consuming. From manual

investigation of 200 accounts out of the large dataset obtained using this approach, the authors in [20] found that the accuracy is also high (98%). We used the same dataset and crawled all accounts that are accessible via the Twitter API except those ones that either do not exist or do not allow access anymore at the time of our experiments. After all, the training set contains 186,874 accounts/instances where 85.6% (159,989) are individual accounts and 14.4% (26,885) are non-individual ones. The idea is that using this large dataset with a small portion of noisy labels is able to train better deep learning approaches compared to using a small manually labeled dataset as shown in [20].

We constructed the first test set – Test (Ours) in Table 2 – based on the same idea from Wood-Doughty et al. [20], and retrieved 989 Twitter accounts. Afterwards, we further manually investigated and adjusted incorrect labels where 10 (1%) account labels have been adjusted. Among the 989 accounts, 63.1% (624) are individual accounts and the rest (365) are non-individual ones.

The second test dataset – Test (Humanizer) in Table 2 – is from McCorriston et al. [11] where all the accounts are manually annotated using Amazon Mechanical Turk\(^2\). After filtering those accounts that are not accessible, the dataset contains 15,809 accounts in total where 88.9% (14,050) are individual accounts and the rest (1,759) are non-individual ones.

### 4.2 Compared Methods

To evaluate the performance of LIIC, we compare LIIC with the following baseline and state-of-the-art approaches.

- **Majority Class** is a straightforward baseline method which always predicts the majority class, i.e., *individual* in our case.

- **Humanizer** [11] uses three types of features such as post content features (e.g., words and hashtags for each class), stylistic features (e.g., the average number of words used per tweet), and structural and behavioral features based on how the account interacts with others (e.g., ratio of retweets to tweets), and trained a LibSVM [1] classifier using those features. We used the implementation of the authors in [11]\(^6\).

- **Demographer** [20] uses a character-based CNN to learn the representation of a user’s name, and uses profile-based features such as the ratio of followers to friends as well as the presence of pronouns such as “my” and “our”. We used the implementation of the authors in [4] which is publicly available\(^7\).

- **RandomForest** [4] uses a set of language-independent features, and a Random Forest classifier is used for classifying different account types. We extracted the same set of features that are used in [4], and trained a Random Forest classifier using the training set with three-fold cross-validation for optimizing hyperparameters.

### 4.3 Results

We evaluate the performance of aforementioned methods in terms of the overall accuracy of classification, and precision, recall, and F\(_1\) score of both individual and non-individual labels. Table 3 and 4 show the results on our and the humanizer datasets, respectively. The last column of each table indicates whether a method is language independent.

As we can see from Table 3, all methods outperform the baseline approach – **Majority Class**. LIIC provides competitive performance compared with Demographer despite LIIC is language independent and does not require any profile description or tweet. Overall, LIIC and Demographer achieve an accuracy of 0.93 followed by Humanizer and RandomForest. When the focus is the classification of individual, LIIC with a precision of 0.97 and F\(_1\) score of 0.95 provides the best performance followed by Demographer. Demographer performs better in terms of precision (0.91) when the classification focus is non-individuals compared to LIIC (0.88) while LIIC performs better in terms of both recall and F\(_1\) score.

The results of the humanizer dataset in Table 4 show similar trends with Table 3. For instance, both LIIC and Demographer achieve an accuracy of 0.94 followed by RandomForest (0.90). LIIC with an F\(_1\) score of 0.97 outperforms Demographer when the focus is the classification of individual. When the focus is classifying non-individuals, LIIC with an F\(_1\) score of 0.74 also performs better than Demographer while Demographer provides a higher precision (0.91) over LIIC (0.88).

The performance on the two datasets indicates that we can achieve comparable performance for distinguishing individual accounts in a language-independent manner using LIIC.

### 5 ABLATION STUDY

To investigate the effectiveness of the components of LIIC such as the ones for profile images, screen names, and profile features, we conducted an ablation study with some variants of LIIC listed below.

- **LIIC\(_img\)** uses \(V_p\) instead of \(V_{fused}\) in Figure 1 to predict the label \(\hat{y}\), which does not use the screen name or statistical features of a user such as the number of friends. This can be seen as an image classifier.

- **LIIC\(_screenname+features\)** uses the concatenated vector \(V_c\) instead of \(V_{fused}\) in Figure 1 for predicting the label \(\hat{y}\), which does not use the profile image of a user.

---

\(^1\)https://www.mturk.com/  
\(^6\)http://networkdynamics.org/resources/software/humanizr/  
\(^7\)https://bitbucket.org/mdredze/demographer/src
Table 3: Performance of individual classification using compared methods on our dataset with the best-performing scores in bold (except the baseline – Majority Class). Majority Class, RandomForest, and LIIC are language independent as they are not using textual information such as profile descriptions or posts.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Individual</th>
<th></th>
<th>Non-individual</th>
<th></th>
<th>Language Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F₁</td>
<td>Recall</td>
<td>F₁</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.63</td>
<td>0.63</td>
<td>1.00</td>
<td>0.77</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Humanizer [11]</td>
<td>0.81</td>
<td>0.79</td>
<td>0.95</td>
<td>0.87</td>
<td>0.88</td>
<td>0.58</td>
</tr>
<tr>
<td>Demographer [20]</td>
<td>0.93</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.77</td>
<td>0.76</td>
<td>0.93</td>
<td>0.84</td>
<td>0.82</td>
<td>0.50</td>
</tr>
<tr>
<td>LIIC</td>
<td>0.93</td>
<td>0.97</td>
<td>0.92</td>
<td>0.95</td>
<td>0.88</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 4: Performance of individual classification using compared methods on the humanizer dataset [11] with the best-performing scores in bold (except the baseline – Majority Class). The results of Humanizer is not applicable as it is trained using the humanizer dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Individual</th>
<th></th>
<th>Non-individual</th>
<th></th>
<th>Language Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F₁</td>
<td>Recall</td>
<td>F₁</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.89</td>
<td>0.89</td>
<td>1.00</td>
<td>0.94</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Humanizer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Demographer</td>
<td>0.94</td>
<td>0.95</td>
<td>0.98</td>
<td>0.96</td>
<td>0.82</td>
<td>0.55</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.90</td>
<td>0.92</td>
<td>0.98</td>
<td>0.95</td>
<td>0.63</td>
<td>0.30</td>
</tr>
<tr>
<td>LIIC</td>
<td>0.94</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
<td>0.71</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 5: Performance of variants of LIIC by removing different components.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy</th>
<th>Individual</th>
<th></th>
<th>Non-individual</th>
<th></th>
<th>Language Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>LIIC</td>
<td>0.93</td>
<td>0.97</td>
<td>0.92</td>
<td>0.95</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>LIIC_img</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>LIIC_screenname+features</td>
<td>0.79</td>
<td>0.77</td>
<td>0.94</td>
<td>0.85</td>
<td>0.84</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>LIIC_features</td>
<td>0.72</td>
<td>0.72</td>
<td>0.91</td>
<td>0.80</td>
<td>0.72</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>LIIC_screenname</td>
<td>0.74</td>
<td>0.72</td>
<td>0.96</td>
<td>0.82</td>
<td>0.85</td>
<td>0.35</td>
</tr>
<tr>
<td>Humanizer</td>
<td>LIIC</td>
<td>0.94</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>LIIC_img</td>
<td>0.94</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>LIIC_screenname+features</td>
<td>0.92</td>
<td>0.93</td>
<td>0.98</td>
<td>0.95</td>
<td>0.72</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>LIIC_features</td>
<td>0.90</td>
<td>0.91</td>
<td>0.98</td>
<td>0.95</td>
<td>0.64</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>LIIC_screenname</td>
<td>0.88</td>
<td>0.92</td>
<td>0.95</td>
<td>0.94</td>
<td>0.46</td>
<td>0.31</td>
</tr>
</tbody>
</table>

- LIIC_features relies on the set of features only with a fully connected dense layer, and uses $V_f$ instead of $V_{fused}$ in Figure 1 for prediction.
- LIIC_screenname relies on the screen name of a user only with a bidirectional GRU, and uses $V_s$ instead of $V_{fused}$ in Figure 1 for prediction.

Table 5 shows the results of those variants of LIIC on both our and the humanizer datasets. Overall, LIIC achieves the best performance in terms of the accuracy and F₁ scores for both individual and non-individual. This indicates that considering all components of LIIC is useful to achieve the best performance. It is worth noting that LIIC_img which can be seen as an image classifier also provides good performance and outperforms the rest of those variants. This shows that profile images alone indeed provide a good clue for distinguishing individuals, and LIIC_img can be generally applied to other social media platforms as it does not rely on features that might be tied to a specific platform. The results in Table 5 also show that LIIC_screenname+features outperforms LIIC_features and LIIC_screenname. This indicates that incorporating both screen name and features is useful compared to considering each separately.

6 CONCLUSIONS

In this paper, we presented LIIC which is a language-independent approach for classifying individual accounts on social media platforms such as Twitter. Despite of the language independence, our results show that LIIC can achieve competitive performance compared to other state-of-the-art methods. In addition, the ablation study in Section 4 indicates that the profile image of an account indeed is a good indicator for distinguishing individuals, while using all components of LIIC provides the best performance. For
future work, we will investigate using different CNN models such as customized smaller models compared to VGG16 for extracting the profile image features and its impact on performance. Our datasets and code can be found here.8

REFERENCES


8https://github.com/parklize/twitter-account-classification