2021 International Conference on Computational Performance Evaluation (ComPE)
North-Eastern Hill University, Shillong, Meghalaya, India. Dec 1-3, 2021

Towards Emotional Intelligence: Analysis of Static Facial Features in LinkedIn Profile Pictures

Quynh T. Nguyen
Department of Mathematics and Statistics
Langara College
Vancouver, BC, Canada
and
Deptartment of Business Administration and Management
Dai Nam University
Hanoi, Vietnam
qnguyen@langara.ca
nguyentqbio@gmail.com
ORCID: 0000-0002-3265-3386

Raouf N.G. Naguib
School of Mathematics, Computer Science and Engineering
Liverpool Hope University
Liverpool, UK
naguibr@hope.ac.uk
r.naguib@ieee.org
ORCID: 0000-0001-6807-7993

Michael Loo
Department of Mathematics and Statistics
Langara College
Vancouver, BC, Canada
mlo@langara.ca

***Abstract* — In an attempt to distil information conveyed by LinkedIn users' pictures, in addition to the usual information provided in their profiles, we explored a large public dataset of 10,610 Australian LinkedIn users. The dataset contained in excess of 50 parameters, of which 22 were dedicated to picture features. The study confirmed that *K*-means clustering and Principal Component Analysis (PCA) are viable techniques for the classification of users, based on facial feature extraction and analysis. Furthermore, the study demonstrated that reduction in feature dimensionality, using PCA, yielded a significant improvement in computational time and resource consumption.**

Keywords — K-means clustering, Principal Component Analysis, Picture classification, Facial features, Social media profiles.

I. Introduction

LinkedIn is known as a platform for professional social networking. It is increasingly used by young graduates to virtually build their professional and personal profile and, subsequently, to enhance their chances of success in job hunting [1]. Moreover, nascent professionals usually seek ways to make themselves standout from the crowd in their corresponding industry, and thus become more noticeably to prospective employers [2]. To this end, and aside from educational backgrounds and experiences, posting a professional profile picture is an equally important endeavour. This study is therefore focused on classifying types of profile pictures, using existing datasets.

In terms of data availability, a dataset, originally provided by Andrew Truman in 2019, can be publicly accessed and includes nearly 10,610 anonymous LinkedIn profiles that comprise more than 50 parameters [3]. These are divided into four categories, namely, profile (where number of followers and age are the 2 main parameters), job features (e.g., historical number of jobs, service duration for each position, as well as in current job), demographics (e.g., gender, ethnicity and nationality), and profile picture with numerical scores for quality (e.g., blurriness) and facial features (beauty, emotions, facial quality, head details, mouth details, skin details and smile).

Since all profile picture data are numerical, *K*-means and hierarchical clustering algorithms can be experimented with for classification purposes. However, the dataset contains in excess of 50 features and would thus heavily consume resources when experimenting with various clustering techniques. Hence, we endeavoured to use an efficient technique to extract important features and to subsequently use *K*-means clustering to produce classifications. In an attempt to reduce resource consumption, we additionally monitored and compared the time taken to compute *K*-means clustering from the original dataset and from a Principal Component Analysis (PCA)-reduced dataset.

II. Methods

A. K-means clustering

The *K*-means algorithm is a partition clustering method that endeavours to uncover *K* hard clusters. Initially, *K* centroids are selected, where *K* is specified by the user. Data points are individually assigned to the nearest centroid, and respective collections of points form clusters. This subsequently enables the centroid of each cluster to be updated based on the collection of points attributed to it, and the process is continuously repeated until no further changes incur in the clusters [4].

Given a dataset $D = y_1, \ldots, y_n$, the objective of the *K*-means algorithm is to optimise the criterion by which the distance between the objects in a cluster and their respective cluster centroid is minimised. This can be mathematically expressed as follows:

$$\underbrace{min}_{\{cen_k\}, 1 \le k \le K} \sum_{i=1}^{k} \sum_{y \in Ck} \pi_y dist(y, cen_k)$$

where:

K: User-defined number of clusters

π_y: Weight of y

cen_k: Centroid of cluster k

978-1-6654-3656-4/21/$31.00 ©2021 IEEE

The function *dist* computes the distance between object *y* and centroid, cen_k. In this study, the Euclidean distance function is used to compute such distance:

$$dist = (y - cen_k)^2$$

In order to validate the quality or goodness of a cluster, the Silhouette score is used, where distances need to be calculated for each observation that belongs to a cluster, *k*. as follows:

$$S = \frac{(b-a)}{max(a,b)}$$

where:

a is the mean intra-cluster distance, i.e., the distance between an observation and the rest of the data points within the same cluster.

b is the nearest cluster distance, i.e., the distance between an observation and all other data points of the next nearest cluster.

The Silhouette score falls within the range [-1, 1]. A score of 1 means that the clusters are very compact and clearly separated; a score of 0 means that they are overlapping or very close to a decision boundary; a score below 0 means that points have been assigned to the wrong clusters.

B. Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction technique applied to datasets that have a large number of features/variables. It is used when variables are numerical and within the same scale. Therefore, data with different scales would be standardised. PCA results in fewer variables that are weighted linear combinations of the original ones, while retaining the majority of the information of the original dataset [5].

Initially, a large-dimension dataset is assessed to ascertain that the data are in the same scale. If not, the dataset is scaled and, subsequently, a covariance matrix for the features/variables is populated. The eigenvalues and eigenvectors for the covariance are computed in order to sort the variables into *k* components, from the largest to smallest, in order to form a new dataset that has a lower dimension than the original.

III. Data Processing

The publicly-available Kaggle dataset was compiled by Andrew Truman in 2019 and consists of data pertaining to 10,610 LinkedIn anonymous Australian professionals. Various inaccuracies and redundancies in the data were detected and processed. For example, the "Blur" and "Beauty" features of candidates appeared repeatedly under different labels. Equally, since under Australian laws, individuals below the age of 13 are not allowed to work, these were filtered out. Also, any features stored as a string format were discarded from the analysis since this study focused on numerical data only.

To detect multivariate outliers, a Mahalanobis distance with a χ^2 test cut-off of 84.03 was used. In addition, features that had low correlations were removed. This resulted in a final clean dataset consisting of 10,610 instances with 40 features.

In order to understand the concepts of blurriness, beauty and smile, their respective explanations are given below:

- Blurriness (or Blur feature), as the name entails, is a measure of the level of blurriness in a picture.
- Beauty is the assessment of facial beauty based on anthropometric, non-permanent and acquisition characteristics [6]. By the same token, facial symmetry and aesthetically pleasing proportions can be deduced from a picture and are expressed through a parameter known as the "Golden Ratio" [7].
- The smile feature was determined through a picture analysis of various muscles in the facial region. These were split into different frames, and within each frame analysed, specific measures performed on the teeth and soft tissue surrounding the mouth were undertaken, leading to a determination of the smile feature.

As mentioned earlier, the clean dataset had a size of 10,610 instances with 40 features. However, only 22 features related to profile pictures. Hence, those features were extracted for dimensionality reduction and clustering analysis. To identify whether PCA could be applied, pairwise correlations were applied and the pairs with the highest correlation coefficients were selected, as outlined in Table I. Details of values for the 22 features/variables are given in Appendix 1. These high-correlation pairs indicate that PCA can be appropriately used since there is a linear association pattern in the dataset.

TABLE I. PAIRWISE CORRELATION COEFFICIENTS

Pair	Correlations		
	Name of features/variables	***Coef.***	***p-values***
1	Blur_gaussian and face_quality	-.065	< 0.01
2	Emo_happiness and emo_neutral	-0.91	< 0.01
3	Emo_happiness and mouth_open	0.56	< 0.01
4	Emo_happiness and smile	0.78	< 0.01
5	Emo_neutral and mouth_close	0.55	< 0.01
6	Smile and mouth_open	0.65	< 0.01

IV. Data Analysis and Discussion

A. K-means clustering results

The original data was used to generate similarity points according to features of profile pictures, using a similarity metric based on Euclidean distance. Cluster analysis can be visually exhibited through the dissimilarity matrix and Hopkins' statistic [8, 9]. Here, the red and blue colours indicate high and low similarity, respectively, as shown in Fig. 1. Thus, from the visual exploration of this matrix, it can be seen that a number of similar clusters exist in the dataset. The Hopkins' statistic produced a value H $\approx$ 0.841. In consequence, the null hypothesis at the $\alpha = 0.05$ level was rejected since the critical value for H is $\approx$ 0.5.

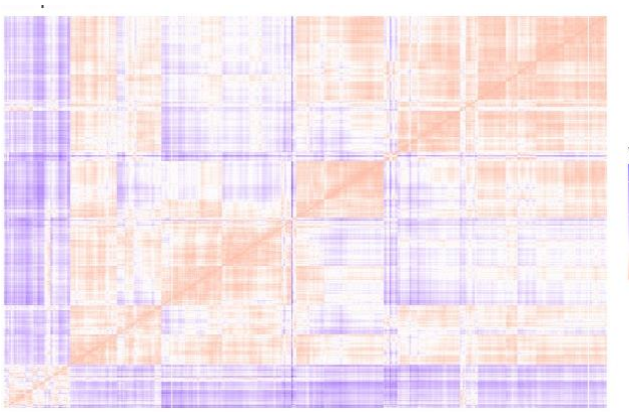

Fig. 1. Dissimilarity Matrix

Since this method requires knowledge of the optimal number of clusters in advance, the value of *K* had to be determined prior to executing the clustering algorithm. Three different methods were used in this study to identify the optimal value of *K*, namely, the Elbow, Silhouette Width and Gap Statistic methods. To visualize the suggested optimal *K*, three plots of the aforementioned methods were produced. The Elbow method suggested an optimal value of *K* of 3 clusters, while the optimal value of *K* suggested through Silhouette Width and Gap Statistic was 4 and 2 clusters, respectively.

We endeavoured to classify the data with those values, i.e., with $K = 2$, 3 and 4. We also used the Silhouette score to identify how well the data points were clustered, in other words, we assessed the quality of clusters pertaining to the different optimal values of *K*. The average Silhouette score for $K = 2$, 3 and 4 was 0.36, 0.22 and 0.24, respectively. On this basis, the optimal value of *K* was determined to be 2.

Consequently, in order to profile these two clusters, their centroids and corresponding distance between them and the various data points had to be examined by concatenating the average of each feature and selecting the most important feature where the differences between the average for each feature in the two clusters are large. The results demonstrated that perceived emotional state, mouth structure and smile appeared to be the most important features for the classification of LinkedIn profile pictures.

B. Pricinpal Component Analysis and K-means Clustering Results

In this study, PCA was performed on scaled numerical data and components that were significant in explaining variances in the original dataset were identified.

The Scree plot in Fig. 2 shows that the first four components explain 48% of the total variance in the subset data. Beyond the 4th component, any additional component does not contribute to the explanation of variation by a significant value. Hence, the first four principal components were selected to profile the profile picture features of LinkedIn users.

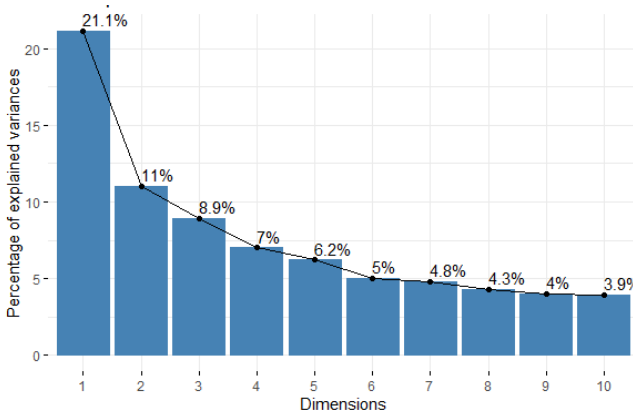

Fig. 2. Scree Plot Showing the Total Variance of 10 Principal Components

Confirmation of the above result was corroborated through the elbow line plot of Fig. 3, which yielded a similar result.

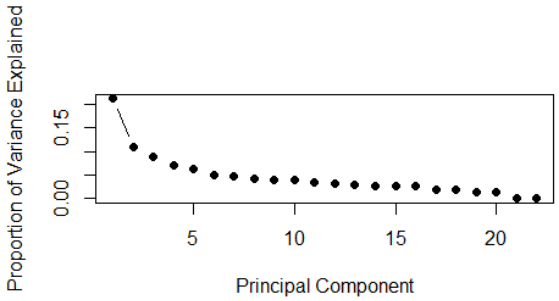

Fig. 3. Elbow Line Plot showing the Proportion of Variance

Fig. 3. shows that components beyond the 5th one do not significantly contribute towards the explanation of proportion of variance. Hence, based on these conclusions, 4 principal components were selected.

We endeavoured to understand which variables bore heavily on those four components by identifying eigenvalues for each of them. By checking the coefficients associated with each component, it could be gleaned that:

- Principal Component 1 may be labelled as "*Facial Expression*". The following variables bore heavily in it: *smile*, neutral emotion (*emo_neutral*), happiness emotion (*emo_happiness*), *mouth_open* and *mouth_closed*.
- Principal Component 2 may be labelled as "*Skin Quality*". The following variables bore heavily in it: *skin_ health*, *skin_stain*, *skin_acne* and *skin_dark_circle*.
- Principal Component 3 may be labelled as "*Blurriness*". This is because *face_quality* and *blur_gaussian* bore heavily in this component.
- Principal Component 4 may be labeled as "*Head Position*" since *head_roll* and *head_yaw* bore heavily in it.

Following the selection of these 4 components, a reduced dataset was formed. The optimal *K* procedure was repeated again on that dataset, and all approaches, including the elbow, silhouette width and gap statistic equally suggested an optimal value of $K = 2$.

C. Performance of K-means clustering

Table II outlines the evaluation of the performance of *K*-means clustering on, both, the original dataset and the PCA-reduced dataset. The procedures were performed on two machines with the same configuration.

TABLE II. PROCEDURES AND DURATIONS OF PERFORMING *K*-MEANS CLUSTERING

Step	***Clustering on the original dataset***	***Clustering on the reduced dataset using PCA***
1	Perform Hopkins' statistic and visualisation on reduced dataset [10,610 x 22] – Duration: 7,250 seconds.	Pearson correlation and covariance matrices – Duration: 0.3 seconds. PCA dimension reduction (checking the eigenvalues to decide how many components to select) – Duration: 1.2 seconds. Perform Hopkins' statistic and visualisation on reduced dataset [10,610 x 4] – Duration: 6,129 seconds.
2	Select optimal *K* clusters using 3 tests – Duration: 151 seconds.	Select optimal *K* clusters using 3 tests – Duration: 122 seconds.
3	Compute *K*-means clustering based on the number of suggested *K* – Duration: 5.5 seconds.	Compute *K*-means clustering based on number of suggested *K* – Duration: 4.5 seconds.
4	Evaluate quality of clusters using Silhouette scores – Duration: 62.7 seconds.	Evaluate quality of clusters using Silhouette scores – Duration: 27 seconds.
5	Profiling users using averages of features for *K* clusters.	Profiling users using averages of features for *K* clusters.
Total[1]	7,250 + 151 + 5.5 + 62.7 = 7,469.2 seconds.	0.3 + 1.2 + 6,129 + 122 + 4.5 + 27 = 6,284 seconds.

For step 1, in order to check the validity of the original dataset and explore the possibility of using *K*-means clustering, Hopkins' statistic and visualisation took 7,250 seconds to yield the outputs. This test was performed on the dataset with a dimension of 10,610 instances x 22 variables.

For the reduced dataset using PCA, initially Pearson correlation and covariance matrices were computed with a duration of 0.3 second. Subsequently, PCA was computed. The process to compute all principal components and visually create a scree plot to select the number of principal components took 1.2 seconds. In this study, the reduced dataset had a size of 10,610 instance x 4 principal components. Hopkins' statistic and visualisation took 6,129 seconds to yield the outputs.

For step 2, The three aforementioned different methods to select optimal *K* were used and the time taken to process these 3 methods was 151 seconds on the original dataset and 122 seconds the PCA-reduced dataset.

Step 3 shows durations of 5.5 and 4.5 seconds for the calculation of the *K*-means clustering for both datasets, respectively.

In step 4, the durations to compute the Silhouette scores for the *K*-means clustering results on the original and reduced datasets were 62.7 seconds and 27 seconds, respectively. The total time taken to yield the result using *K*-means clustering for the original dataset was 7,469 seconds, while the total time for the reduced dataset was 6,284 seconds.

V. FINDINGS AND CONCLUSIONS

In order to assess the demeanour conveyed by a person's LinkedIn profile picture, given a large user dataset, the *K*-means clustering technique was utilised. The meaningful interpretation of such *K*-means clusters was achieved through PCA to extract the important features and compute the clustering technique on a reduced feature dataset. For comparison purposes, *K*-means clustering was also performed on the original LinkedIn profile picture dataset, with each picture comprising 22 features.

The detailed analysis carried out demonstrates that PCA is an appropriate method to reduce picture feature dimensionality, while retaining its important features. The clustering approach applied to the full large dataset appeared to consume significantly more time than on the PCA-reduced dataset. Therefore, it can be concluded that applying PCA prior to computing *K*-means clustering would result in time and resource savings.

Results generated by the *K*-means clustering technique were subsequently combined with user demographic and other LinkedIn information (e.g., number of followers, average days worked in last 10 years and days spent in current job). This suggested two overall categories of classification: Category 1, which included users who uploaded good quality pictures (i.e., achieved a high score for overall beauty according to the analysis undertaken), in combination with a number of followers higher than the corresponding mean of 1,225 in the dataset (note that the maximum number of followers for one user in the dataset was 35,056), while Category 2 comprised individuals with a greater amount of experience but their profile pictures yielded low scores due to their blurriness.

ACKNOWLEDGEMENT

We would like to thank our assistants, Harshad Subhash, Jaspreet Singh and Kirandeep Kaur for providing support in experimenting with the *K*-means clustering technique.

REFERENCES

[1] M. Cubrich et al., "Examining the Criterion-Related Validity Evidence of LinkedIn Profile Elements in an Applied Sample", Computers in Human Behavior, vol. 120, p. 106742, Jul. 2021.

[2] E. Basak and F. Calisir, "Uses and Gratifications of LinkedIn: An Exploratory Study", in *Proceedings of the World Congress on Engineering*, vol. 2, pp. 2-4, 2014.

[3] A. Truman, "Andrew Truman | Datasets Novice," Kaggle, 2019. https://www.kaggle.com/killbot/datasets (accessed Jul. 22, 2021).

[1] Duration of Step 5 was not taken into consideration due to user profiling being conducted by experts and not machine.

[4] J. Wu, "Cluster Analysis and *K*-means Clustering: An Introduction", in *Advances in K-means Clustering: a Data Mining Thinking*, Springer-Verlag, 2012.

[5] G. Shmueli, P. Bruce, I. Yahav, N. Patel, and K. Lichtendajhl, "Chapter 4: Dimension Reduction," in *Data Mining for Business Analytics*, John Wiley & Sons, Inc, 2018.

[6] A. Dantcheva, A. and J.L. Dugelay, J. L, "Assessment of Female Facial Beauty Based on Anthropometric, Non-permanent and Acquisition Characteristics", *Multimedia Tools and Applications*, 74(24), pp. 11331-11355, 2015.

[7] E.P. Prokopakis, I.M. Vlastos, V.A. Picavet, G. Nolst Trenite, R. Thomas, C. Cingi and P.W. Hellings, "The Golden Ratio in Facial Symmetry", *Rhinology*, 51(1), pp. 18-21, 2013.

[8] A.L. Nogueira and C.S. Munita, "Quantitative Methods of Standardization in Cluster Analysis: Finding Groups in Data", *Journal of Radioanalytical and Nuclear Chemistry*, 325, pp. 719-724, 2020.

[9] G.R. Cross and A.K. Jain, "Measurement of Clustering Tendency", IFAC Proceedings Volumes, 15(1), 315-320, 1982.

APPENDIX 1

List of Picture Variables and their Respective Clusters

No	Variable name	Mean	Min	Max	Cluster 1	Cluster 2	Distance[2]
1	Beauty	0.02	-3.39	2.95	-0.054	0.150	-0.204
2	blur_gaussian	-0.03	-0.43	3.54	-0.114	0.114	-0.228
3	emo_anger	-0.05	-0.17	10.94	-0.117	0.063	-0.18
4	emo_disgust	-0.05	-0.22	8.19	-0.094	0.046	-0.14
5	emo_fear	-0.07	-0.16	11.52	-0.099	-0.009	-0.09
6	emo_happiness	0.05	-1.68	0.74	0.638	-1.038	1.676
7	emo_neutral	0.01	-0.60	2.11	-0.567	1.093	-1.66
8	emo_sadness	-0.06	-0.18	10.36	-0.132	0.070	-.0202
9	emo_surprise	-0.06	-0.18	11.22	-0.136	0.086	-0.222
10	face_quality	0.05	-1.94	0.86	0.163	-0.167	0.33
11	head_pitch	0.03	-6.13	4.46	0.160	-0.221	0.381
12	head_roll	0.00	-5.73	5.85	0.012	-0.034	0.046
13	head_yaw	0.01	-6.02	6.11	0.000	0.023	-0.023
14	mouth_closed	0.01	-0.86	1.27	-0.504	0.972	-1.476
15	mouth_mask	-0.05	-0.13	10.69	-0.061	-0.040	-0.021
16	mouth_open	0.02	-1.09	0.98	0.545	-0.955	1.5
17	mouth_other	-0.05	-0.28	4.56	-0.098	0.037	-0.135
18	skin_acne	-0.03	-0.64	3.83	-0.081	0.060	-0.141
19	skin_dark_circle	-0.02	-0.61	4.08	-0.050	0.038	-0.088
20	skin_health	0.02	-0.75	3.86	0.031	-0.012	0.043
21	skin_stain	-0.03	-0.81	3.06	0.011	-0.119	0.13
22	smile	0.03	-1.75	1.13	0.632	-1.087	1.719

[2] Distance: distance between 2 centroids of 2 clusters