**Twitter Users Classification via Sparse PCA**

Barak Yichye, Department of Industrial Engineering (Master's student), BGU.
Dan Vilenchik, Department of Communication Systems Engineering, BGU.

**Overview.** Microblogging services such as Twitter have become an important feature of the daily life of millions of users. In this work we study the problem of user classification in Twitter and suggest a quantitative *un*supervised learning method based on Principal Component Analysis (PCA). To perform the task we use non-textual simple features such as the number of follower, number of tweets, etc. The main contribution of our work is:

- We propose a *generic* approach that may be applied in a straightforward manner to other social networks (such as LinkedIn, Instagram, etc). Our approach is generic in the sense that the user-profile statistics that we use are common (or very similar) across many social networks, e.g. the number of followers or the number of likes. We ignore almost all content-related features, as content may be very different in different networks: text, images, videos.

- We introduce a new concept which we call the *"semantic dimension"* of the problem. PCA is one of the most popular and widely used methods in various data mining applications for the purpose of dimension reduction, visualization and feature transformation. This task is performed by picking the top $r$ PC's and projecting the data onto the subspace spanned by them. The parameter $r$ is usually chosen according to the total variance explained by that set. For these tasks to be accomplished successfully, i.e. in a meaningful way, it is (almost) necessary that the obtained PC's will be sparse. A common practice to achieve this goal is to zero out the entries of lowest absolute value in every PC. In our work we suggest a new methodology to perform this task, and a way of validating the result. We suggest to use sparse PCA instead of standard PCA. We identify the sparsity parameter $k$ (the number of allowed non-zeros in every PC) as the *semantic* dimension of the problem, alongside $r$ which is the *algebraic* dimension. The validity of the choice of $k$ is obtained by comparing the variance explained by the top $r$ PC's and the top $r$ $k$-sparse PC's.

- We suggest a new quantitative measure of *robustness* of the classification. To compute the robustness score we perform a *truncated* crawl of the social network. In this crawl we ignore users with high expression of any of the newly derived features (i.e. with a large projection on any of the leading sparse PC's). We compute the sparse PC's of the "truncated" matrix and compare them to the sparse PC's obtained in the non-truncated crawl. If the PC's are similar in both crawls, we say that the classification is robust, in the sense that it may be useful for various types of (sub-)networks: for example users from a specific region, students in a certain school, etc.

- A byproduct of our work is a perceptron for *spam detection* which takes a different approach than existing spam detection models in social networks. The main two differences are ($a$) we use only structural features (as mentioned above), and in particular no NLP is performed; ($b$) our approach is unsupervised and hence labeled data for training is not required. We tested our classifier on a set of 164 accounts, 69 spam and 95 legitimate, taken from [GK14]. Our classifier obtained precision rate of 94.3% and recall rate of 97.1%. The F1 score 95.7%.

**Our Results.** To sample Twitter we implemented a crawler that crawled the social network graph in a BFS manner exploiting the public APIs provided by Twitter. The crawling rate was about 25,000 users per day (there are limitations posed by the Twitter API), and we collected a total of 284,758 active Twitter accounts. For each account we collected the set of features

described in Table 1. The attributes in Table 1 represent data about the user's activity in the social plane (followers, following, re-tweets) and statistics about the user activity in the content plane (number of tweets, text vs urls, etc). Since the features have different scales, we had to normalize the data to unit variance, as common in such cases (see for example [LPC$^+$04]).

Using the 284,758 Twitter profiles we generated the $12 \times 12$ correlation matrix $\hat{\Sigma}$, computed its leading eigenvectors (PCs), and sorted them according to a decreasing order of eigenvalues. The eigenvalue $\lambda_i$ of the $i^{th}$ PC is proportional to the percentage of variance it explains. In Figure 3, one can see that the first three PCs account respectively for 18%, 16.2% and 13% of the total variance, totalling about 50% of the variance. A similar result is obtained from sparse PCA, see Figure 4; Figures 1 and 2 show side-by-side the top three PC's and the top three 4-sparse PC's.

The new features (or labels) induced by the top three sparse PC's are:
$\mathbf{PC_1}$: This dimension includes the number of followers, number of re-tweets by other users, and the likes given to the user's tweets. These aggregated attributes identify the characteristic of being a popular user in the social network, or measure of the user's celebrity. The top ten Twitter accounts in $\mathbf{PC_1}$-measure in our sample are all pop-music teen idols like Justin Bieber.
$\mathbf{PC_2}$: This dimension has two attributes in positive sign: number of tweets that contain only text, and number of tweets that I re-tweet. In negative sign: number of tweets that contain a url, and number of hashtags in my tweets. The positive values are typical of human twitter accounts, and the negative ones are more typical of robot accounts, which tend to be spam accounts. Indeed this PC is our perceptron spam detector. The top Twitter accounts in our sample in $\mathbf{PC_2}$-measure are bot-accounts like *prayerballoons* and *3XasianIdolVids*.
$\mathbf{PC_3}$: This dimension includes the number of tweets and tweet rate, likes given to others, and the number of other users mentioning. These attributes measure the extent to which a user is a content provider. The top accounts in our sample here include news providers *littlebytesnews*, video gaming support *XboxSupport*, and an American teen content provider *ChelseaAMusic*.

**Conclusions.**

- We were able to derive new meaningful features from the raw set of 12 features. The feature $\mathbf{PC_1}$ is a popularity measure, $\mathbf{PC_2}$ measures the spam potential of the Twitter account, and $\mathbf{PC_3}$ a content provider measure.

- We can see the clear benefit of performing a *semantic* dimension reduction, and not only an algebraic one. The two principal components $\mathbf{PC_2}, \mathbf{PC_3}$, which relate to the user's activity in the network, are far better feature-wise separated in the sparse PCA case than in the standard PCA. Using the sparse version we get a clear distinction between the two types of activities: spam potential vs non-spam content (like news). See Figures 5 and 6.

- The choice of $k = 4$ is indeed justified. The scree plots in Figure 3 and 4 reconfirm that the top three 4-sparse eigenvectors cover roughly the same amount of variance as the original PC's, and therefore we may conclude that the correct semantic dimension of the observed phenomenon is 4, while its algebraic dimension is 3. More generally, one may wonder whether there is a universal constant for the semantic dimension of user-classification in social networks. Comparing to the YouTube user classification task carried out in [CCL10] also using PCA, also there the number of large entries in the first two PC's is indeed 4 and 5. Is this pure chance? We leave this as an intriguing direction for future research.

| Var | Attribute | Description |
|---|---|---|
| $x_1$ | NumOfTweets | Total number of tweets |
| $x_2$ | NumOfFollowers | Total number of users following me |
| $x_3$ | NumOfFollowing | Total number of users I follow |
| $x_4$ | LikesGivenToOthers | Number of tweets that I like |
| $x_5$ | NumOfTxt | Total number of tweets that contain only text |
| $x_6$ | NumOfUrl | Total number of tweets that contain URLs |
| $x_7$ | NumOfMyRT | Number of other users' tweets that I re-tweet |
| $x_8$ | NumOfOthRT | Total number of retweets by other users to the user's tweets |
| $x_9$ | TweetsPerDay | Total number of tweets divided by lifetime (in days) |
| $x_{10}$ | NumOfUserMent | Number of other users that are mentioned in the tweets |
| $x_{11}$ | NumOfHashTag | Number of hashtags # that are referenced in the tweets |
| $x_{12}$ | LikesGivenToMe | Number of likes that my tweets received |

Table 1: Feature details

| PCA1 | | | |
|---|---|---|---|
| features | PC1 | PC2 | PC3 |
| NumOfTweets | -0.01 | 0.28 | -0.42 |
| NumOfFollowers | 0.38 | 0.02 | -0.07 |
| NumOffolowing | 0.06 | 0.05 | -0.22 |
| LikesGivenToOthers | -0.03 | 0.30 | -0.17 |
| NumOfTxt | -0.07 | 0.56 | 0.07 |
| NumOfUrl | 0.07 | -0.47 | -0.35 |
| NumOfMyRT | -0.06 | 0.22 | 0.31 |
| NumOfOthRT | 0.65 | 0.08 | 0.07 |
| TweetPerDay | -0.03 | 0.34 | -0.42 |
| NumOfUserMent | 0.00 | 0.17 | -0.41 |
| NumOfHashTag | 0.05 | -0.29 | -0.41 |
| LikesGivenToMe | 0.65 | 0.09 | 0.07 |

Figure 1: Top three PCs

| SPCA1 | | | |
|---|---|---|---|
| features | PC1 | PC2 | PC3 |
| NumOfTweets | 0.00 | 0.00 | 0.60 |
| NumOfFollowers | 0.38 | 0.00 | 0.00 |
| NumOffolowing | 0.07 | 0.00 | 0.00 |
| LikesGivenToOthers | 0.00 | 0.00 | 0.44 |
| NumOfTxt | 0.00 | 0.57 | 0.00 |
| NumOfUrl | 0.00 | -0.61 | 0.00 |
| NumOfMyRT | 0.00 | 0.32 | 0.00 |
| NumOfOthRT | 0.65 | 0.00 | 0.00 |
| TweetPerDay | 0.00 | 0.00 | 0.59 |
| NumOfUserMent | 0.00 | 0.00 | 0.31 |
| NumOfHashTag | 0.00 | -0.44 | 0.00 |
| LikesGivenToMe | 0.65 | 0.00 | 0.00 |

Figure 2: Top three sparse PCs, $k = 4$

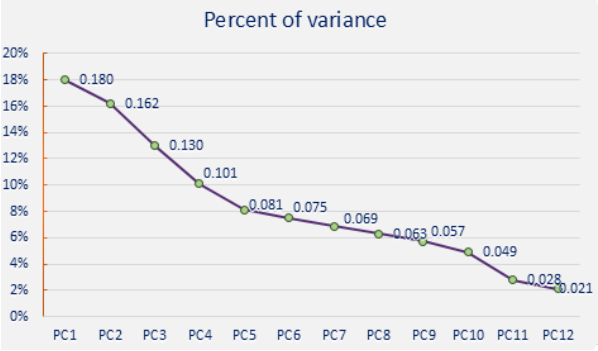


Figure 3: Scree plot for PCA

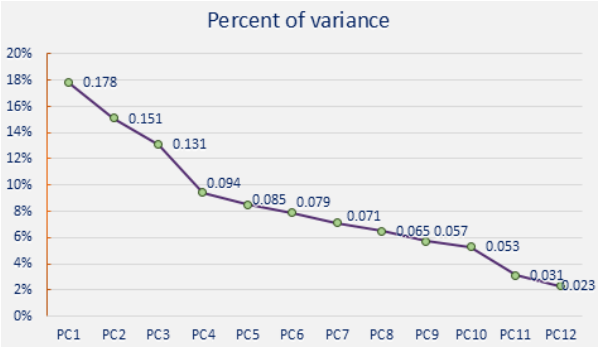


Figure 4: Scree plot for Sparse PCA, $k = 4$

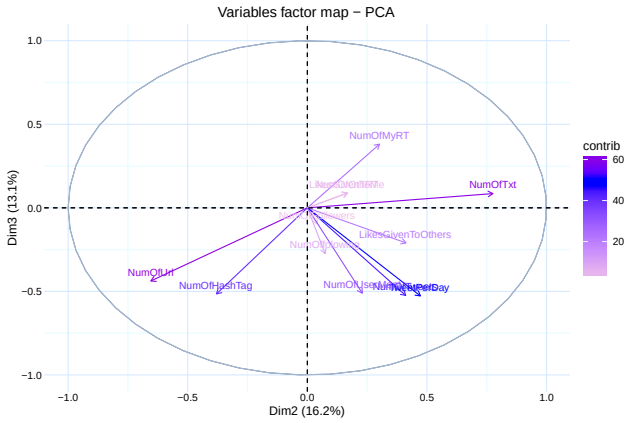


Figure 5: Factor map for standard PCA

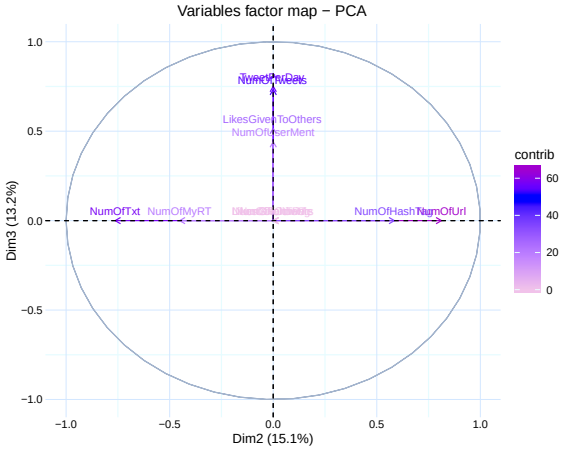


Figure 6: Factor map for Sparse PCA

## References


[CCL10] C. Canali, S. Casolari, and R. Lancellotti. A quantitative methodology to identify relevant users in social networks. In *Business Applications of Social Network Analysis (BASNA), 2010 IEEE International Workshop on*, pages 1–8, Dec 2010.

[GK14] A. Gulec and Y. Khan. Feature selection techniques for spam detection on twitter. Technical report, Electronic Commerce Technologies (CSI 5389) Project Report, SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, University of Ottowa, 2014.

[LPC+04] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. *SIGMETRICS Perform. Eval. Rev.*, 32(1):61–72, June 2004.