DBMI 2016

Twitter User Classification via Sparse PCA

Barak Yichye*, Dan Vilenchik, BGU**

- Twitter is one of the leading micro-blogging service (~300M users)
- A fertile platform for content-marketing, advertising but also spamming
- User classification is a fundamental task in social networks

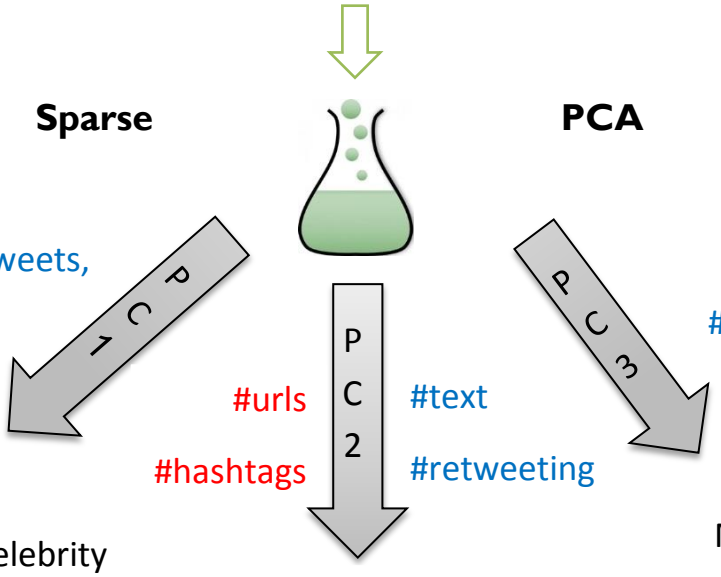
Our contribution:

- Unsupervised-learning method for user classification based on **sparse PCA**
- Standard PCA is used to reduce the **algebraic** dimension
  - Our approach – use **sparse PCA** to reduce **semantic dimension**
  - Benefits: enhanced **interpretability** of the new feature space
- We use non-textual features ⇨ method **transferable** across social networks

Methodology:

Crawling Twitter (300,000 accounts)

12 "raw" features: # tweets, # followers, # likes, # of urls in tweets



Measure of fame/celebrity

#5 in our sample

PC2 is a perceptron for spam detection: Human vs Bot account

Measure of content provider

#4 in sample: Xbox support

* Department of Industrial Engineering, BGU (Master student) ** Department of Communication Systems Engineering, BGU.