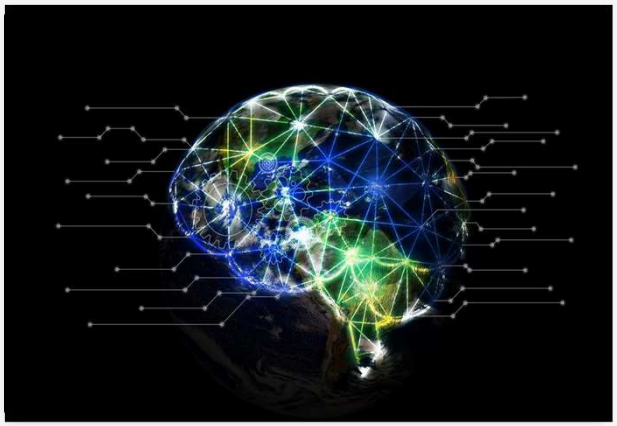


Large Scale Social Media-based Analysis of Human Behavior and Decision Making

SHIMEI PAN
UMBC
(SHIMEI@UMBC.EDU)

Acknowledgement

This is a joint work with my PhD students


Tao Ding


Arpita Roy


Doa Yakut Kilic

and my collaborator:


Dr. Warren Bickel, Virginia Tech.

Social Media-based Behavior Analysis

Social media contain rich and diverse behavior evidences that are indicative of who we are

Main Advantages

- **Large scale**: includes the behaviors of millions of social media users
- **Comprehensive:** contains a large number of personal and social factors (e.g., individual traits, family and community context)
- **Longitudinal**: follows behaviors of users continuously over a long period of time (e.g., years);
- **Organic**: data are automatically, continuously collected in an open and natural environment

Main Challenges (1)

Heterogeneous User Data → Need to combine them to paint a comprehensive picture

Data associated with a typical social media user

- Text : tweets or status updates on Facebook
- Image: profile pictures and images shared
- Like: books, movies, TV shows, music, stores, brands, places that a social media user likes
- Social network: friends and followers
- Demographics: social media profile
- …

Main Challenges (2)

Large Feature Space → Curse of Dimensionality

- Text (including symbols, hashtags and emojis)
  - Unigrams/bigrams: millions of features
- Image
  - Thousands/millions of pixels (depending on resolution)
- Like
  - Millions of different things (books, movies, stores, brands) and people (authors, scientists, celebrities)
- Social network
  - Social graph with millions of nodes and links

Main Challenges (3)

Small ground truth dataset

- Survey-based behavior assessment
  - Lengthy questionnaire
    - E.g., Five 5 personality assessment: IPIP-50 (5o items), IPIP-300 (300 items)
- Hard to scale
    - Typical size : a few hundred to a few thousand people

Small ground truth + large feature space →

Learned models can easily overfit the data!

Key: Unsupervised User Modeling (User Embedding)

- Goal:
  - To employ unsupervised or self-supervised learning to derive a small number of (e.g. a few hundreds) latent user features to characterize the behavior and decision making process of an individual based on raw social media data
- Advantages:
  - A large amount of training data are available
  - Perform both feature learning and dimension reduction simultaneously

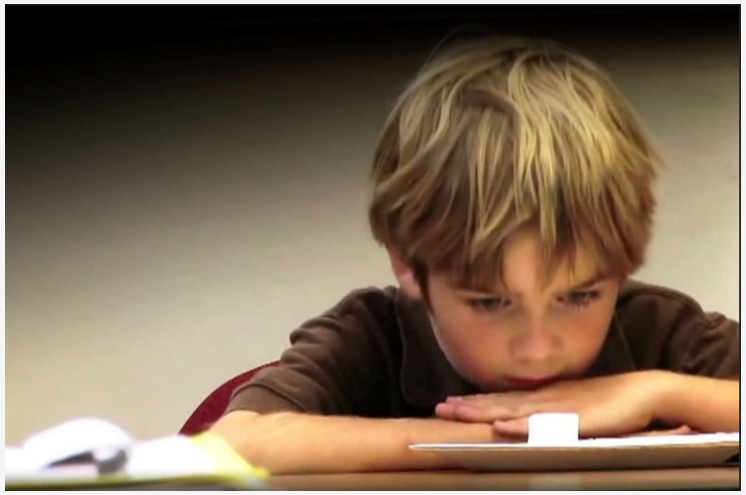
Case Study

- Modeling ***Delay Discounting*** from social media *likes*

- Predicting ***Substance Use*** from social media *posts* and *likes*

Case study1 : Delay Discounting

- A behavioral measure of impulsivity to quantify the human tendency to choose a smaller, sooner reward over a larger, later reward
- The Marshmallow Test (Mischel et al. 1970, Mischel et al. 1972)



Why Study Delay Discounting?

- Delay discounting is a hot topic in economics and behavior science
  - Pitting the demands of long-term goals against short-term desires is among the most difficult tasks in human decision making
- High delay discounting rate is linked to
  - pathological gambling
  - credit card default
  - poor academic performance
  - substance abuse
  - …

Research Question: Can We Infer Delay Discounting from Social Media Likes?

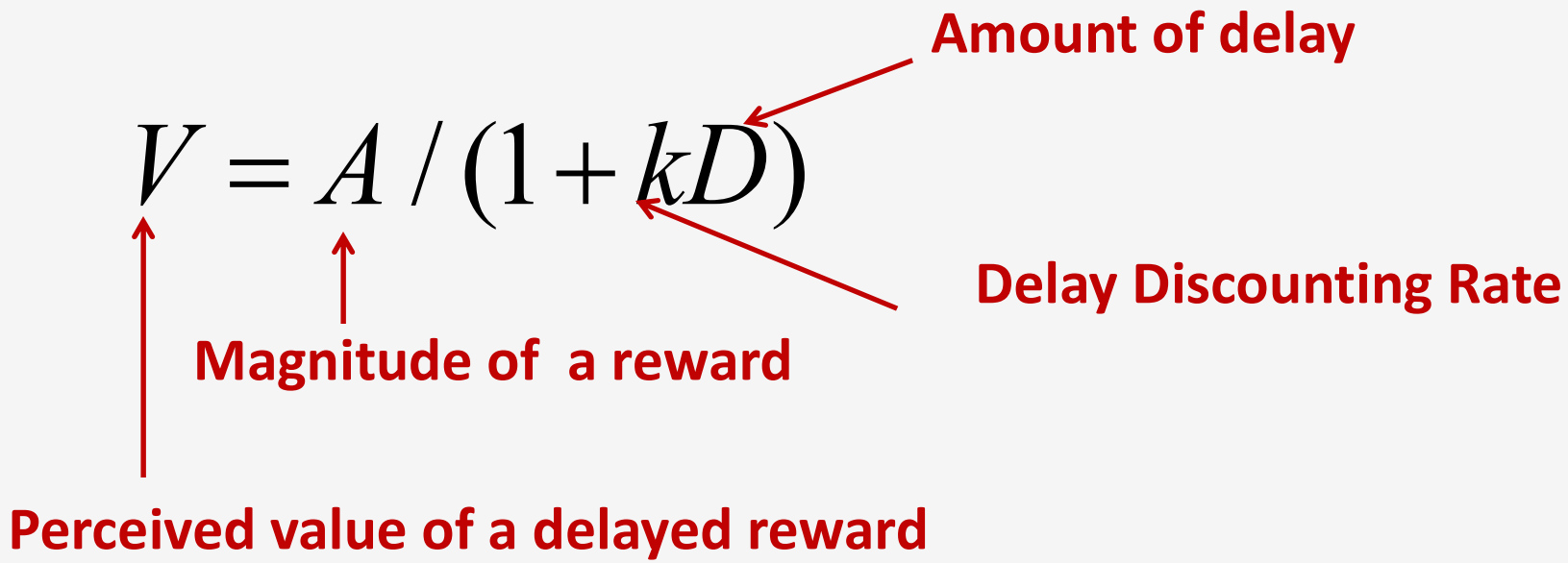
- Social media Likes represent one of the most generic digital footprints left by people on social media
  - A user can indicate his likes of almost anything such as a book, a movie, a song, a brand, a store, a hobby, a statement, a person (e.g., an author, scientist, actor, singer) …
- Previous research has shown that social media Likes contain rich evidences that are indicative of who we are (Kosinski et al. 2013)
  - Personality, ethnicity, religious and political views, intelligence, happiness, age, gender …

Dataset Description

- Raw social media data
  - 11 million Facebook users
  - 9.9 million unique Like Entities (LEs)
  - 1.8 billion user-like pairs
  - Average Likes per user: 161
  - Average Likes received per LE: 182
- Delay discounting ground truth data
  - 3508 people with both Likes and delay discounting ground truth

Modeling Delay Discounting

- A hyperbolic delay discounting model



- Delay discounting Rate (DDR)
  - *Small DDR → small discount for future reward (tomorrow person)*
  - *Large DDR → steep discount for future reward (today person)*

How to Obtain DDR Ground Truth?

- Delay discounting task (Stillwell et al. 2012)
  - Each FB user was presented with 15 different immediate monetary rewards (e.g., $1000, $950, $900 … $100)
  - The future reward is always $1000
  - The delays were between 1 week and 5 years
  - Multi-item delay discounting questionnaire such as
    - $ 950 now or $1000 in 1 year
    - $ 900 now or $1000 in 1 year
    - $ 850 now or $1000 in 1 year
    - $ 800 now or $1000 in 1 year
    - $ 750 now or $1000 in 1 year
    - $ 700 now or $1000 in 1 year …
  - Obtain "indifference value" for each delay (e.g., $700 above)
  - Compute $k_i=(A-V_i)/V_i\ D_i$ ($k_{1yr}=(1000-700)/1000*700=0.00043$)
  - *Normalize* $(k_i)=log\ (k_i)$ *(e.g.,* $log(0.00043)=-3.37$*)*
  - *DDR= average of normalized* (k_i) *for all the delays*

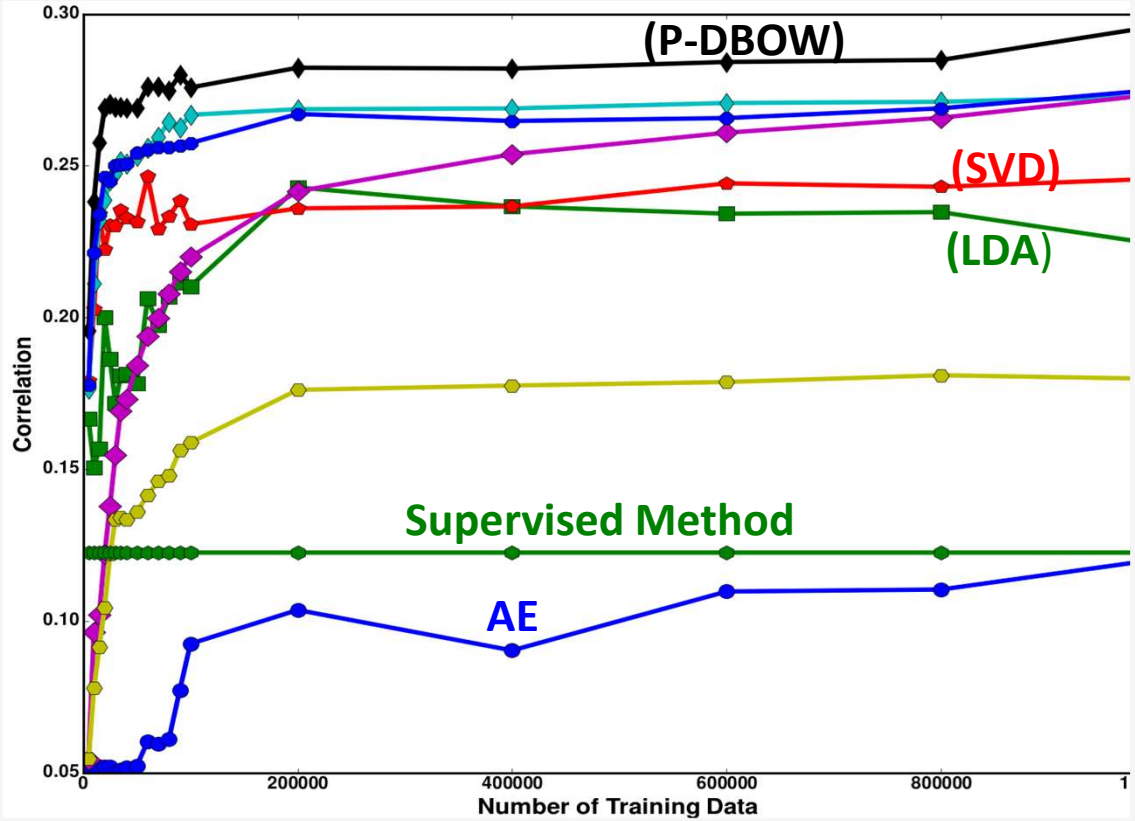
User Like Embedding

- Goal: to employ unsupervised/self-supervised learning to obtain a representation of user behavior and traits based on all the social media Likes of a user
- Input: all the social media Likes of a user
  - A sparse vector with a large number of features (e.g., one for each unique LE)
- Output: a latent user like representation
  - A dense vector representation with a small number of features (e.g., a few hundred features)

User Like Embedding Methods

Method	Inference	Stages	Aggregation	Local Context	Interpretable
SVD	count	1-stage	direct inference	No	No
LDA	count	1-stage	direct inference	No	Yes
Autoencoder	prediction	1-stage	direct inference	No	No
U-CBOW	prediction	2-stage	Average	Yes	No
U-SG	prediction	2-stage	Average	Yes	No
U-GloVe	count	2-stage	Average	Yes	No
P-DM	prediction	1-stage	direct inference	Yes	No
P-DBOW	prediction	1-stage	direct inference	No	No

Evaluating DDR Prediction



The improvement of the best model (P-DBOW) over the model without user embedding is 123%

Understanding Like Embedding and Delay Discounting

Correlation analysis with Bonferroni correction

- Control for confounding variables such as age and gender

Topic ID	Correlation	Representative Likes
Positive Correlation		(Favored more by a today person)
141	+	*2Pac, Wiz Khalifa, Ludacris, Dr. Dre, Tyga ...*
430	+	*wake up in middle of night, look at clock, yes I still have time to sleep! OH, I GET IT! (Teacher walks away) Dude, i STILL dont get it ...*
431	+	*Ciara, R. Kelly, Tyrese Gibson, Kelly Rowland ...*
014	+	*The Tattoo Page, Kat Von D, Inked Magazine...*
051	+	*Eminem, Association football, Corsair, Logitech Gaming, AMD Gaming ..*
Negative Correlation		(Favored more by a tomorrow person)
494	−	*Wikileaks, BBC Earth, Ferris Bueller's Day Off, Earth hour ...*
250	−	*Star Trek, The Shawshank Redemption, The Lord of the Rings (film), Start Wars ...*
481	−	*NPR, The Daily Show, The Colbert Report, The Onion, Barack Obama ...*
159	−	*The Lord of the Rings, The Lord of the Rings Trilogy, Lord Of the Rings, The Hobbit ...*
405	−	*George Takei, Ricky Gervais, Peter Jackson, Bill Nye The Science Guy, Ian McKellen ...*

Case Study 2: Predicting Substance Use

- Goal: to predict substance use based on multiple types of social media data
  - Likes → User Like Embedding (ULE)
  - Posts → User Post Embedding (UPE)
  - Likes+Posts → Multi-view User Embedding (MUE)

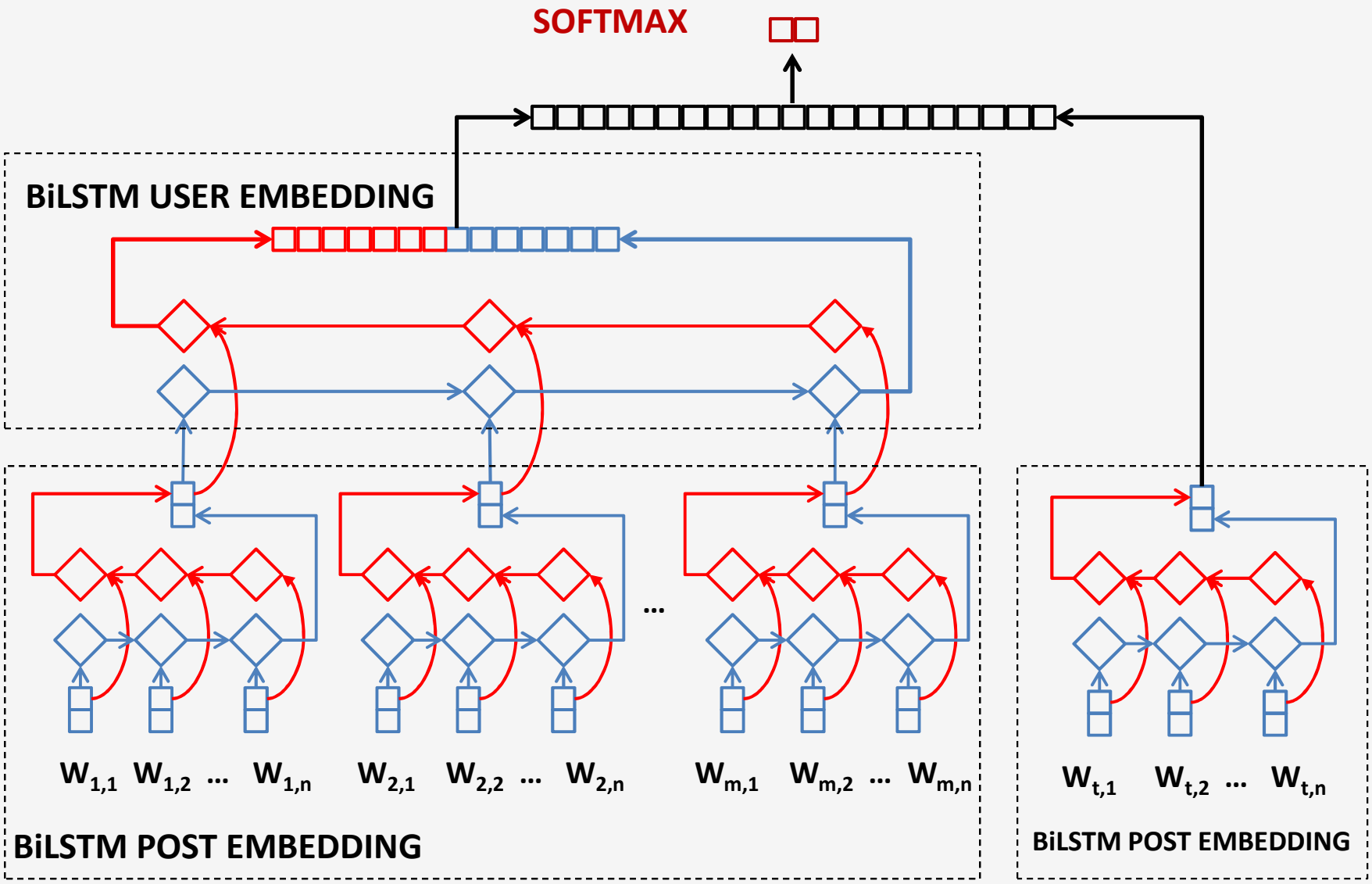
Dataset

Dataset	users	AvgUserLikes	AvgUserPosts	Usage
Likes	5,138,857	184	NA	Single View Feature Learning
LikesSUD	3,508	267	NA	Single View SUD Prediction
Status Update	106,509	NA	143	Single View Feature Learning
StatusSUD	1,231	NA	195	Single View SUD Prediction
LikeStatus	54,757	232	220	Multi-View Feature Learning
LikeStatusSUD	896	277	219	Multi-View SUD Predication

User Post Embedding (UPE)

- Goal: to learn a latent representation based on all the social media posts of a user
- Methods
  - SVD
  - LDA (UserLDA, PostLDA_Word, PostLDA_Doc)
  - Post_DM
  - User_DM
  - Post_DBOW
  - User_DBOW
  - AUT

User Post Embedding: AUT



Multi-view User Embedding (MUE)

- Goal: to combine user like embedding (ULE) with user post embedding (UPE)

- Methods:
  - Canonical Correlation Analysis (CCA)
  - Deep CCA (DCCA)

Canonical-Correlation Analysis (CCA)

- Given two vectors X and Y, CCA tries to find vectors a and b, so that a'X and b'Y that are maximally correlated. → linear transformation

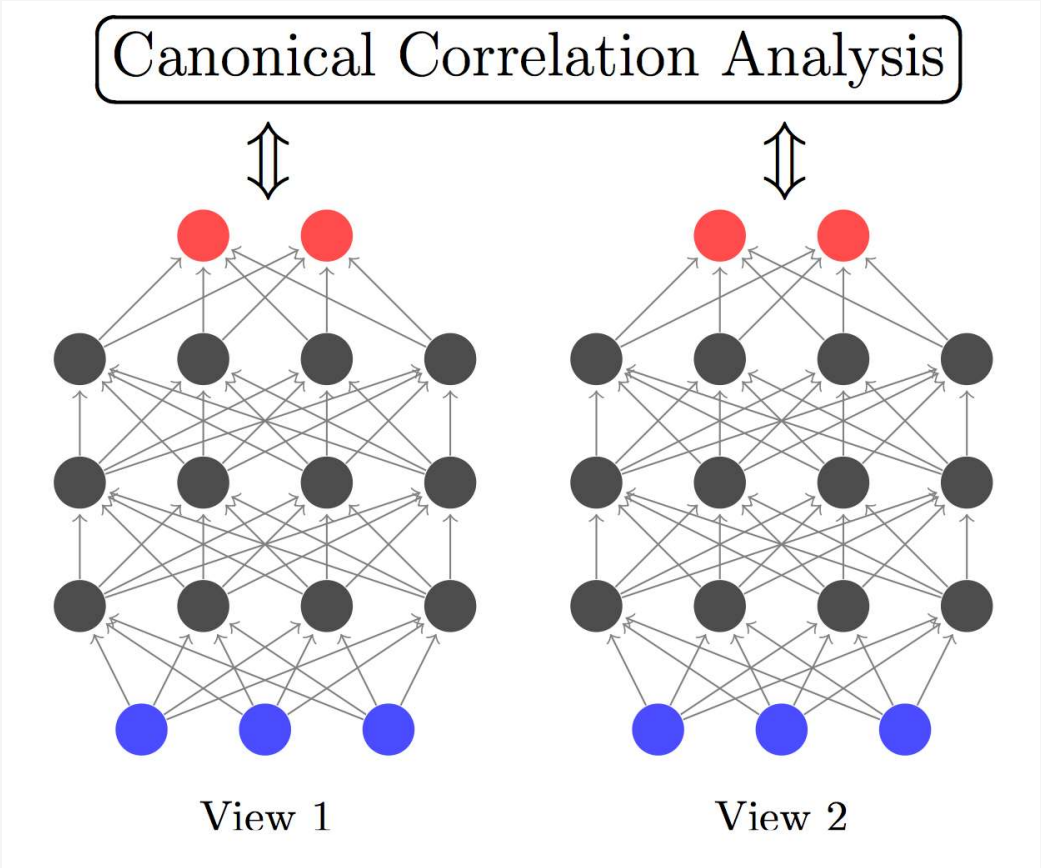
$$(a^*, b^*) = \arg\max_{a,b} corr(a'X, b'Y)$$

$$= \arg\max_{a,b} \frac{a' \sum_{XY} b}{\sqrt{a' \sum_{XX} ab' \sum_{YY} b}}$$

- $\Sigma_{X,X}$: covariance (X,X), $\Sigma_{Y,Y}$: covariance (Y,Y), $\Sigma_{X,Y}$: cross-covariance (X,Y)

Deep Canonical-Correlation Analysis (DCCA)

- Goal: to learn highly correlated deep architectures
- A non-linear extension of CCA

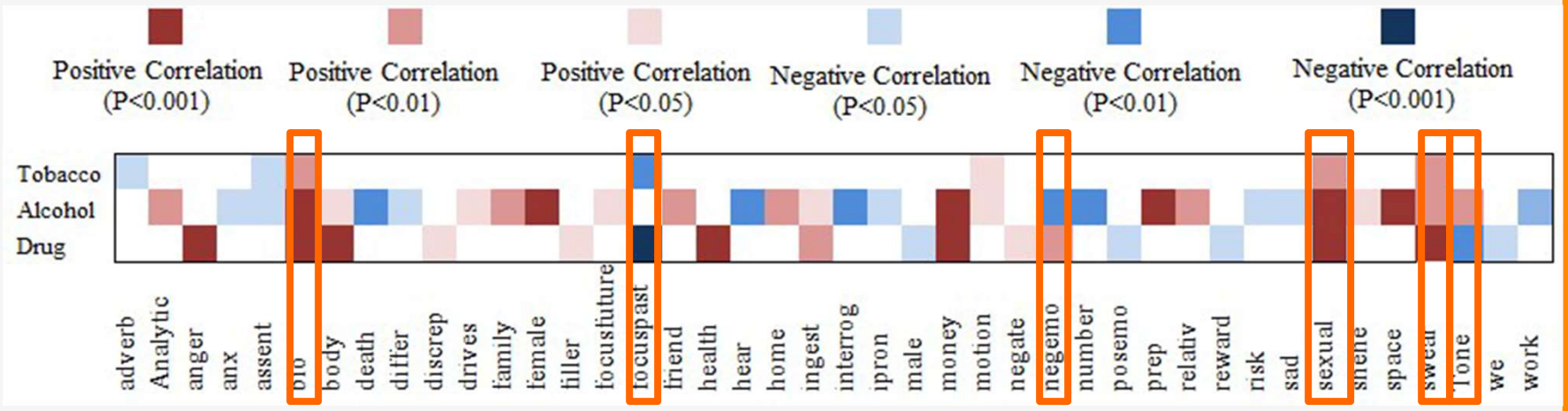


Predicting SUD: Results

Models	Tobacco (ROC AUC)	Alcohol (ROC AUC)	Drugs (ROC AUC)
NO PRE-TRAINING	0.685	0.669	0.662
Konsinski 2013	0.73	0.70	0.65
UPE ONLY	0.802	0.768	0.819
ULE ONLY	0.787	0.795	0.791
CCA	**0.855**	**0.811**	**0.844**
DCCA	0.774	0.781	0.737

* ROC AUC is the Area Under the Receiver Operating Characteristic (ROC) Curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity).

Predicting SUD Based on Word Usage



Conclusions

- Social media contains rich, diverse behavior evidence that can be used to model individual behavior and decision making

- The raw feature space is very large → curse of dimensionality → unsupervised feature learning is the key to success

Contact

Dr. Shimei Pan (shimei@umbc.edu)

Thnnk You!