Name Embeddings and Online News Analysis

Speaker: Junting Ye
Department: Computer Science
Advisor: Prof. Steven Skiena

Outline

- **Overview**
- **Name Embeddings**
  - Nationality Classification
  - Ethnicity & Gender Embeddings
- **Quality Analysis of News and Social Media**
  - Motivation
  - *MediaRank* Overview
  - Progress
- **Future Work**

Overview

News Analysis	• V. Kulkarni, **J. Ye**, S. Skiena, W. Wang, *Multi-modal Models for Political Ideology Detection of News Articles*, Under review.
Name Embeddings	• **J. Ye**, S. Skiena, *The Secret Lives of Names? Public Name Embeddings and Lifespan Modeling*, Working paper. • **J. Ye**, S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, S. Skiena, *Nationality Classification using Name Embeddings*, in Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM), Nov. 2017, pages 1897- 1906.
Opinion Spam Detection	• **J. Ye**, S. Kumar, L. Akoglu, *Temporal Opinion Spam Detection by Multivariate Indicative Signals*, the 10th International AAAI Conference on Web and Social Media (ICWSM), May 2016, pages 743-746. • **J. Ye**, L. Akoglu, *Discovering opinion spammer groups by network footprints*, in Proceed- ings of the 14th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Sep. 2015, pages 267-282.
Others	• H. Chen, X. Sun, **J. Ye**, S. Skiena, *Dynamics of Restaurant Reviews: Sites, Ratings, and Topics*, Under review. • **J. Ye**, L. Akoglu, *Robust Semi-Supervised Learning on Multiple Networks with Noise*, in Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Melbourne, Australia, Jun. 2018.

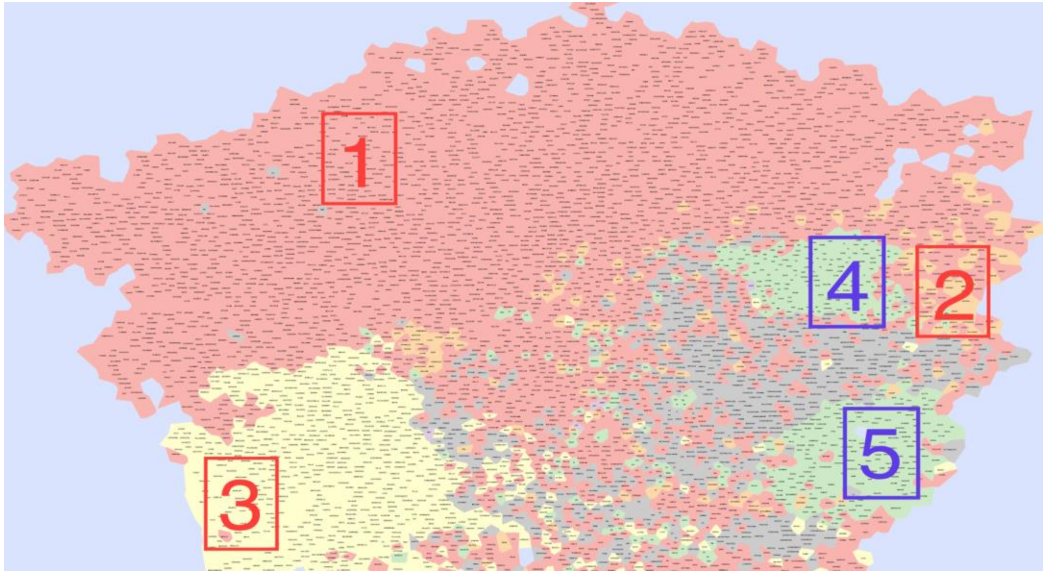
Outline

- **Overview**
- **Name Embeddings**
  - Nationality Classification
  - Ethnicity & Gender Embeddings
- **Quality Analysis of News and Social Media**
  - Motivation
  - *MediaRank* Overview
  - Progress
- **Future Work**

Name Embeddings

Using our machine learning algorithm, each name part (**first or last name**) is represented by a 100-dimention **vector** (i.e. embedding).

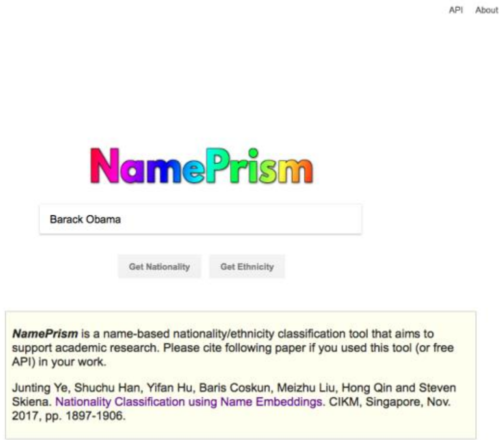
When projecting 100-dimention to 2-dimention:



Name Embeddings

Input (examples)	Gerda_Zavada@ Roxana Carmen, Adina Margine, Radoi Seicaru, Drînd Ramona,… Chilap_ja@ leung Ja, Chow Iris, Ken Ja, Betty Cheung, Chan Stone, Donna Tang, … balbirsingh@ Krishan Singh, Neeraj Kumar, Pankaj Bawa, Vijay Kumar, …
Objective Function (negative sampling)	$\log \sigma({v'_{w_O}}^{\top} v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^{\top} v_{w_I}) \right]$
Labels	Positive: name part pairs in the same list Negative: random name part pairs
Output	Distributed representation of name parts

NamePrism: A nationality classifier



NamePrism is a name-based nationality/ethnicity classification tool that aims to support academic research. Please cite following paper if you used this tool (or free API) in your work.

Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin and Steven Skiena. Nationality Classification using Name Embeddings. CIKM, Singapore, Nov. 2017, pp. 1897-1906.

Our API* has been supporting **100+** research projects from social science, economics, etc..

Research Project Goal	Research Group	Country
“working on ***racial representation*** in historical bureaucracies”	Haas School of Business, UC Berkley	U.S.
“determine if ethnic group size impacts national ***cabinet diversity***”	Department of Political Science, Washington University in St. Louis	U.S.
“promote the ***contributions of Iranian Americans*** to members with-in and outside of the Iranian community living in America.”	Iranian Americans' Contributions Project	U.S.
“determine if ethnicity plays a part/plays no part in whether a written evidence submitted to a ***Parliamentary Inquiry is accepted or rejected***”	Parliamentary Digital Service	UK
“working on a study on the network effects for long term ***unemployed***”	German Institute for Employment Research	Germany
“unveiling the origins of French citizens in order to study ***discrimination*** in several areas of the French society”	Laboratoire Interdisciplinaire Sciences Innovations Sociétés (LISIS)	French
“Investigate whether hosts on Airbnb get ***discriminated*** based on their ethnicity”	Stockholm School of Economics	Sweden

*: www.name-prism.com

Media Coverage

- WIRED Magazine;
- Irish Tech News;
- TyN Magazine;
- 24 Heures;
- ….

Gender & Ethnicity Classification

Embedding	White	Black	API	Hisp.	Avg.
Retweet	0.92	0.20	0.57	0.64	0.58
Mention	0.93	0.22	0.61	0.71	0.62
Follower	0.94	0.31	0.77	0.86	0.72
Followee	0.92	0.27	0.72	0.81	0.68
Followee*	0.94	0.31	0.77	0.84	0.72
Friends	0.93	0.28	0.74	0.81	0.69
NonFriends	0.92	0.26	0.71	0.82	0.68
Aggregated	0.93	0.32	0.76	0.83	0.71
Aggregated*	**0.94**	**0.33**	**0.79**	**0.86**	**0.73**
Email	**0.96**	**0.47**	**0.83**	**0.87**	**0.78**

Table 3: Ratios of same-ethnicity names among nearest neighbor (i.e. k = 1). *Aggregated* gets promising performance It achieves comparable performance on *White* and *Hispanic*. *Black* names are harder task because they only take up 3.5% of all labels. (API: Asian and Pacific Islander)

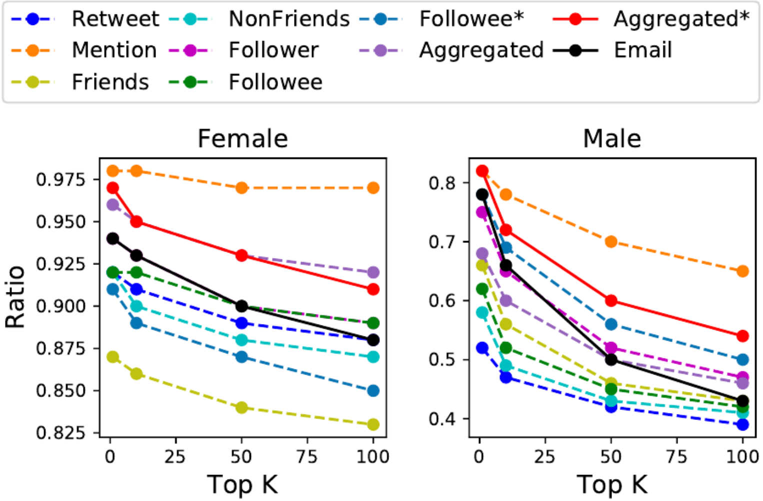


Figure 2: Ratio of same-gender names among top k nearest neighbors ($k \in [1, 10, 50, 100]$). *Mention* performs the best (avg. on female: 0.94, male: 0.74). *Aggregated* outperforms *Email* (female (avg.): 0.94 vs. 0.91, male: 0.67 vs. 0.59). Performance of male names are generally lower than female, because there are far less male names (29% vs. 71%).

Outline

- **Overview**
- **Name Embeddings**
  - Nationality Classification
  - Ethnicity & Gender Embeddings
- **Quality Analysis of News and Social Media**
  - Motivation
  - *MediaRank* Overview
  - Progress
- **Future Work**

Quality Analysis of News and Social Media

Motivation

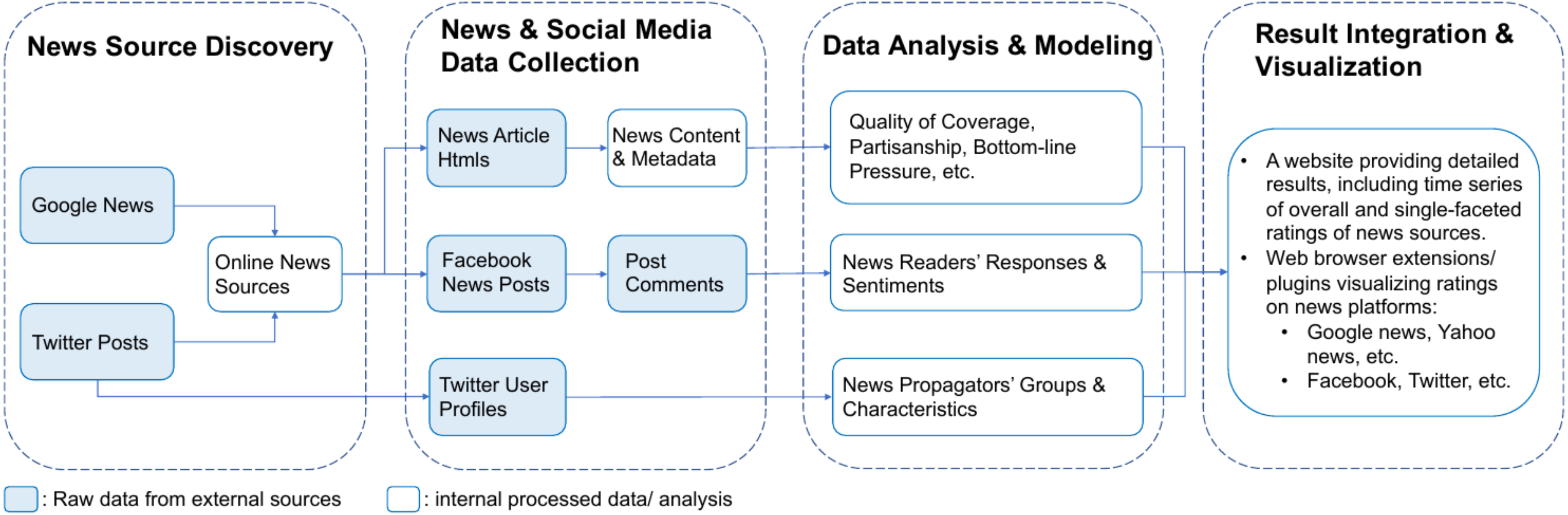
- Fake news went viral in 2016 election
  - Pizzagate of Hillary Clinton
  - Pope endorse Donald Trump
  - ISIS leader calls for American Muslim voters to support Hillary Clinton
  - Donald Trump sent his own plane to transport 200 stranded marines in 1991
  - …

- Impact of fake news on social media
  - 62% U.S. adults get news on social media in 2016 [1]
  - 15% recall seeing fake news headlines [1]
  - Popular fake news shared more times and faster on Facebook than mainstream news [2]

[1]: [H. Allcott & M. Gentzkow, Journal of Economic Perspectives, 2017]

[2]: [S. Vosoughi, Science, 2017]

MediaRank



News Source Discovery

Google News

Twitter Posts

Online News Sources

News & Social Media Data Collection

News Article Htmls

News Content & Metadata

Facebook News Posts

Post Comments

Twitter User Profiles

Data Analysis & Modeling

Quality of Coverage, Partisanship, Bottom-line Pressure, etc.

News Readers’ Responses & Sentiments

News Propagators’ Groups & Characteristics

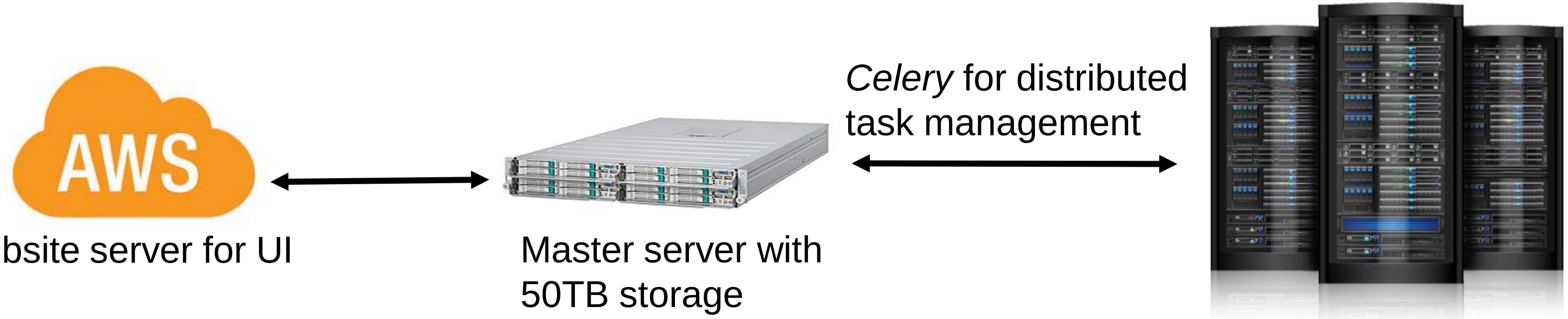
Result Integration & Visualization

- A website providing detailed results, including time series of overall and single-faceted ratings of news sources.
- Web browser extensions/ plugins visualizing ratings on news platforms:
  - Google news, Yahoo news, etc.
  - Facebook, Twitter, etc.

▨: Raw data from external sources □: internal processed data/ analysis

Figure 1: Four major components of *MediaRank* system.

MediaRank: System Overview



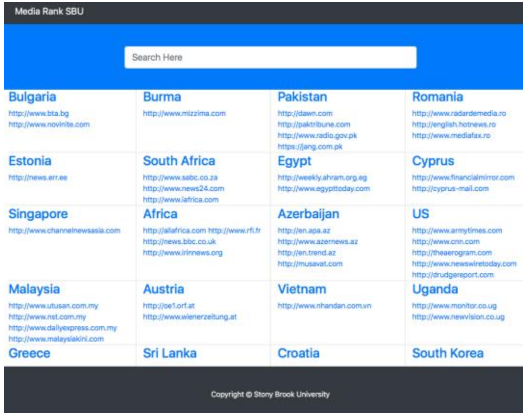
OpenStack for virtualization;
Ansible for cluster management

Celery for distributed task management

Website server for UI

Master server with 50TB storage

Cluster of 85 workers



Media Rank SBU

Search Here

Bulgaria http://www.bta.bg http://www.novinite.com	**Burma** http://www.mizzima.com	**Pakistan** http://dawn.com http://paktribune.com http://www.radio.gov.pk https://jang.com.pk	**Romania** http://www.radardemedia.ro http://english.hotnews.ro http://www.mediafax.ro
Estonia http://news.err.ee	**South Africa** http://www.sabc.co.za http://www.news24.com http://www.iafrica.com	**Egypt** http://weekly.ahram.org.eg http://www.egypttoday.com	**Cyprus** http://www.financialmirror.com http://cyprus-mail.com
Singapore http://www.channelnewsasia.com	**Africa** http://allafrica.com http://www.rfi.fr http://news.bbc.co.uk http://www.irinnews.org	**Azerbaijan** http://en.apa.az http://www.azernews.az http://en.trend.az http://musavat.com	**US** http://www.armytimes.com http://www.cnn.com http://theaerogram.com http://www.newswiretoday.com http://drudgereport.com
Malaysia http://www.utusan.com.my http://www.nst.com.my http://www.dailyexpress.com.my http://www.malaysiakini.com	**Austria** http://oe1.orf.at http://www.wienerzeitung.at	**Vietnam** http://www.nhandan.com.vn	**Uganda** http://www.monitor.co.ug http://www.newvision.co.ug
Greece	**Sri Lanka**	**Croatia**	**South Korea**

Copyright © Stony Brook University

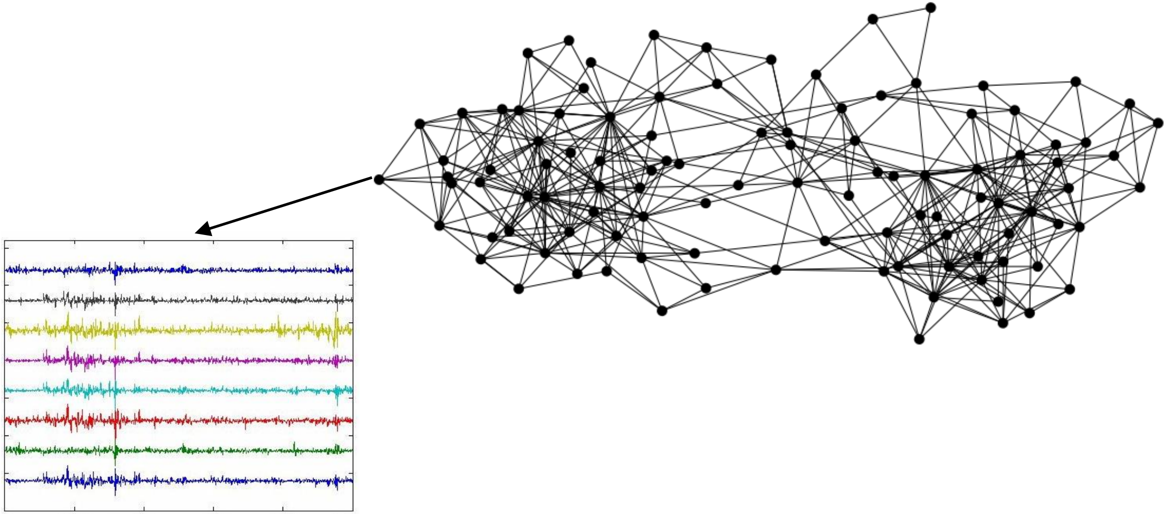
	Source	*Html*	*T_Post*	*T_User*	*FB_Post*	*FB_Comment*	*FB_Like*
Total	71K	224M	375M	30M	51M	714M	3055M
Daily	0.15K	1.1M	2.5M	10K	150K	2.1M	–

Table 1: Sizes of datasets from last 6 months and they are growing on daily basis. *T*: Twitter, *FB*: Facebook.

MediaRank: News Analysis

- Independent Signals
  - Social Media
  - Monetization
  - Political Bias
  - Quality of the Coverage
  - Duplicate Articles
  - Popularity
  - Readability

- Relations
  - Hyperlinks
  - Common News Reader



Timeline for Following Year

- Aug. 2018 ~ Dec. 2018:
  - Investigating political bias and monetization;
  - Leading a team of two PhD and three master students on computing remaining signals and building reliable system;
- Jan. 2018 ~ May. 2018:
  - Modeling heterogeneous signals;
  - Publishing papers and defend thesis;

Q & A