Analyzing Linguistic Differences Between Owner and Staff Attributed Tweets

Daniel Preoțiuc-Pietro, Bloomberg – AI Group

Rita Devlin Marier, Bloomberg – News

Bloomberg Engineering

TechAtBloomberg.com

Motivation

User-level predictive tasks are very successful:

Age (Rao et al. 2010 ACL)
Gender (Burger et al. 2011 EMNLP)
Location (Eisenstein et al. 2010 EMNLP)
Personality (Schwartz et al. 2013 PLoS One)
Impact (Lampos et al. 2014 EACL)
Political Orientation (Volkova et al. 2014 ACL)
Mental Illness (Coppersmith et al. 2014 ACL)
Occupation (Preoțiuc-Pietro et al. 2015 ACL)
Income (Preoțiuc-Pietro et al. 2015 PLoS One)
Account Type (McCorriston et al. 2015 ICWSM)

User-level representations are used to improve results on a variety of tasks:

Sarcasm (Amir et al. 2016 CoNNL, Oprea&Magdy Next talk)
Comment moderation (Pavlopoulos et al. 2017 EMNLP)
Stance detection (Benton et al. 2018 EMNLP)

Motivation

However, all these methods make a tacit assumption:

Tweets posted from an account are from a single person

TechAtBloomberg.com

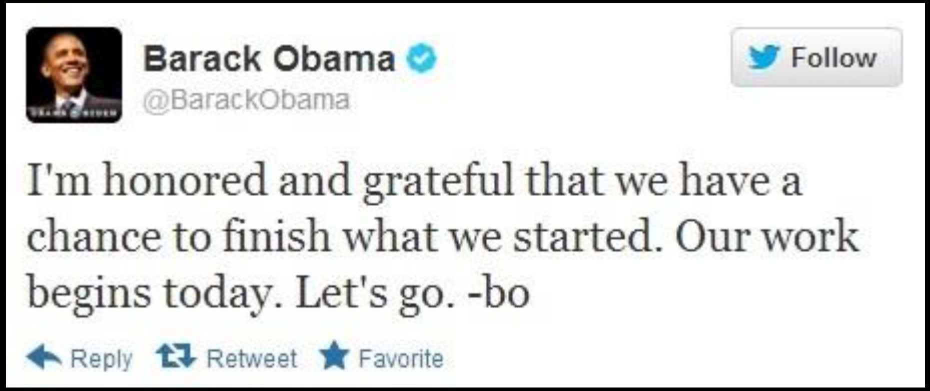
Bloomberg

Engineering

Motivation

One account = one person does not always hold

1. Politicians have staffers post their content



Signed by Barack Obama
Likely posted by Barack Obama

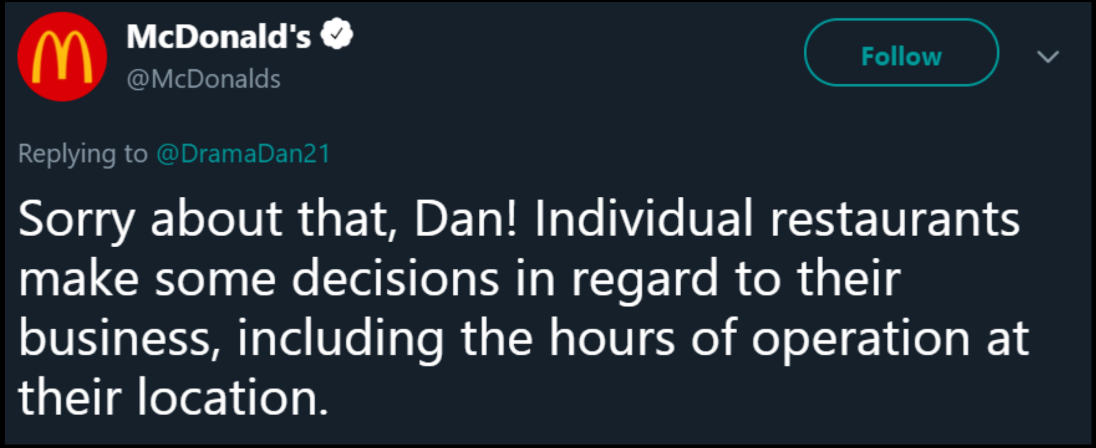
Not Signed by Barack Obama
Likely posted by staff

Motivation

One account = one person does not always hold

2. Companies use their account handles for marketing and customer support

Marketing tweet



Customer support tweet

Aim

The aim of our paper is to study differences between **types** of users posting from the same account

Case study:
- Twitter
- U.S. politicians
- Owner vs. staff attributed

Data – Acquisition

TechAtBloomberg.com

Bloomberg Engineering

Data – Acquisition

Identify accounts which use a signature using a regex

- e.g., “tweets by me are signed …”

Manually annotated

- If the account uses the convention
- The signature of the account (e.g., -bo)

Largest subgroup are U.S. politicians – 147 accounts

- We only use these accounts to control for the topic

Disclaimer:

- Users may use the signature deceitfully
  — Albeit, little to gain and a lot to lose
- We refer to the task as author vs. staff **signed**

Bloomberg

Engineering

Data – Processing

Downloaded most recent 3,200 tweets from each account

- Retweets are removed

Search for signature in each tweet using a regex

- This is the label to predict
- If found, remove signature

Data set

- Size - 202,024 tweets in English
- Signed tweets - 4.8%
- Publicly available: https://github.com/danielpreotiuc/signed-tweets

Features

We experiment with traditional features to aid with our analysis

Tweet features

- **Length:** char, tokens
- **Type:** @-reply, URL
- **Time:** hour of day, day of week
- **Impact:** no. of retweets, no. of likes

Topics

- LIWC topics (Pennebaker et al. 1995)
- Word2Vec Clusters (Preoțiuc-Pietro et al. 2015 ACL)

Sentiment & emotion predictions

- Positive or negative sentiment (Mohammad and Turney, 2013)
- Six Eckman emotions

Unigrams

Prediction

Binary classification task

Logistic Regression with Elastic Net regularization

Evaluated using ROC AUC (Area under the Curve)

- Data is class imbalanced (95 – 5)
- Random performance is 0.50

Experimental setups

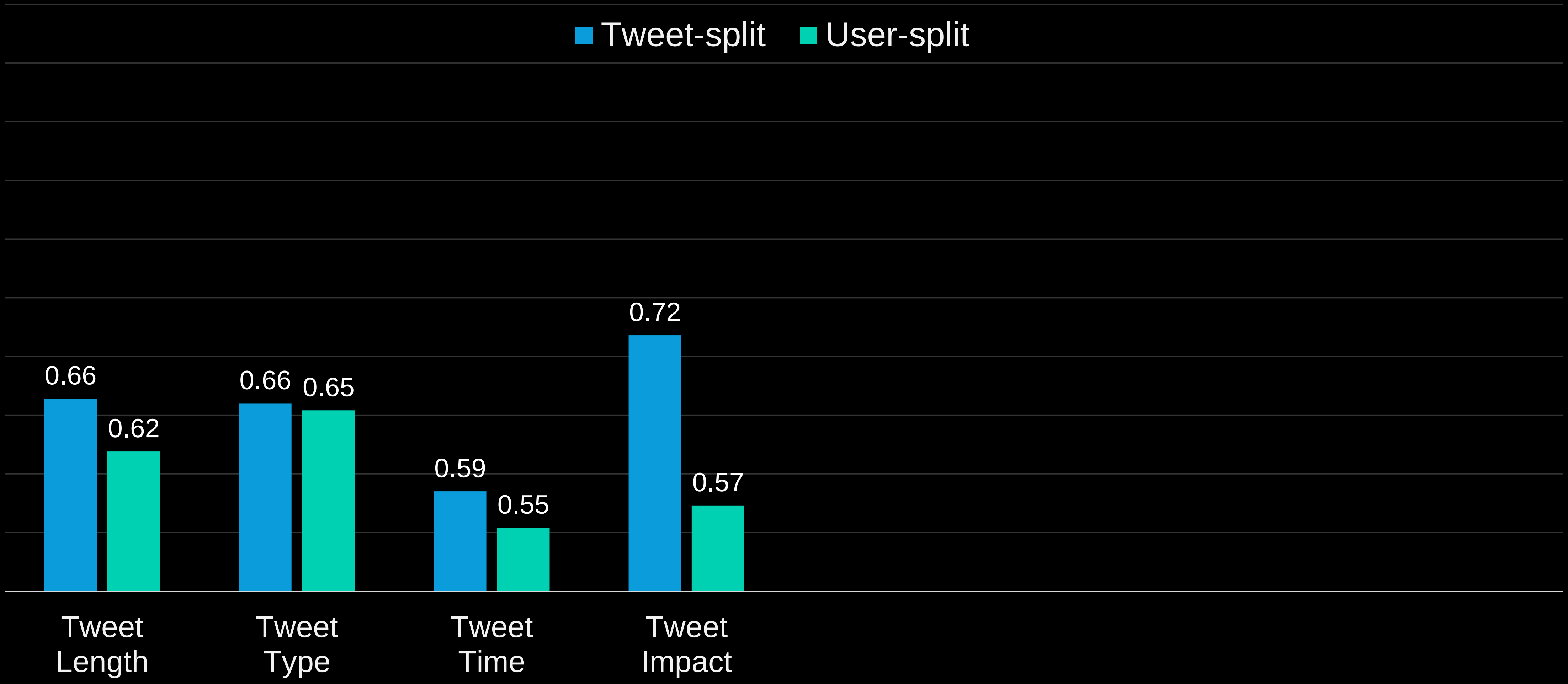
Tweet-split Train Test

User-split Train Test

Bloomberg

Engineering

Prediction

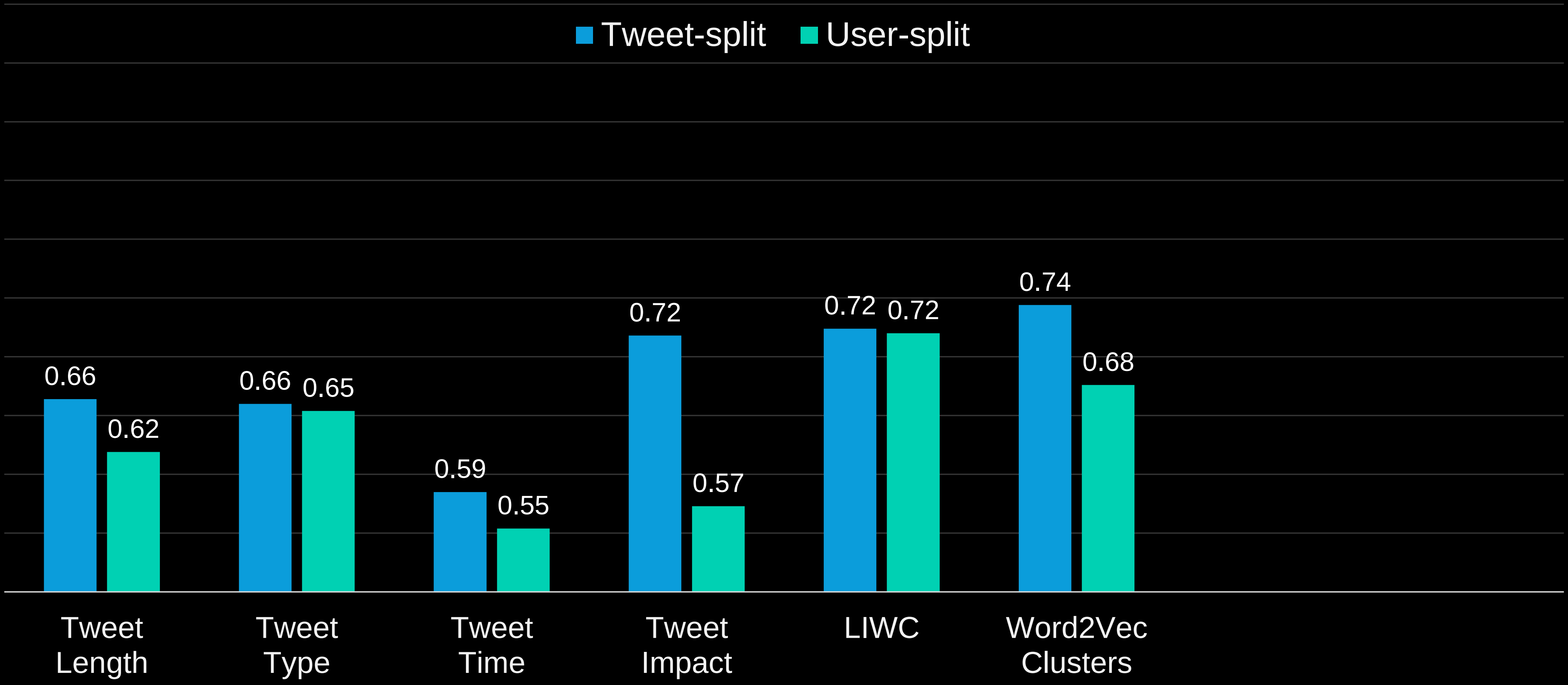


TechAtBloomberg.com

Bloomberg

Engineering

Prediction

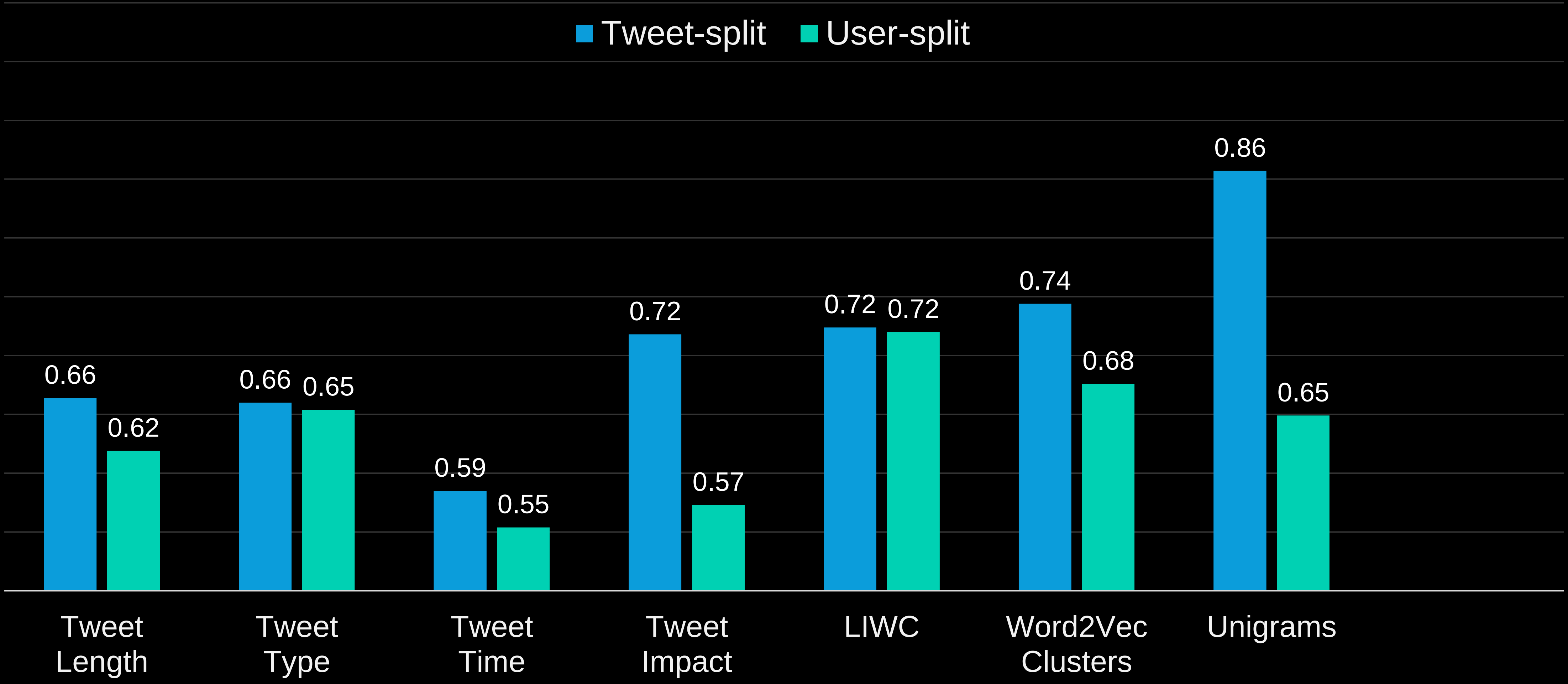


TechAtBloomberg.com

Bloomberg

Engineering

Prediction

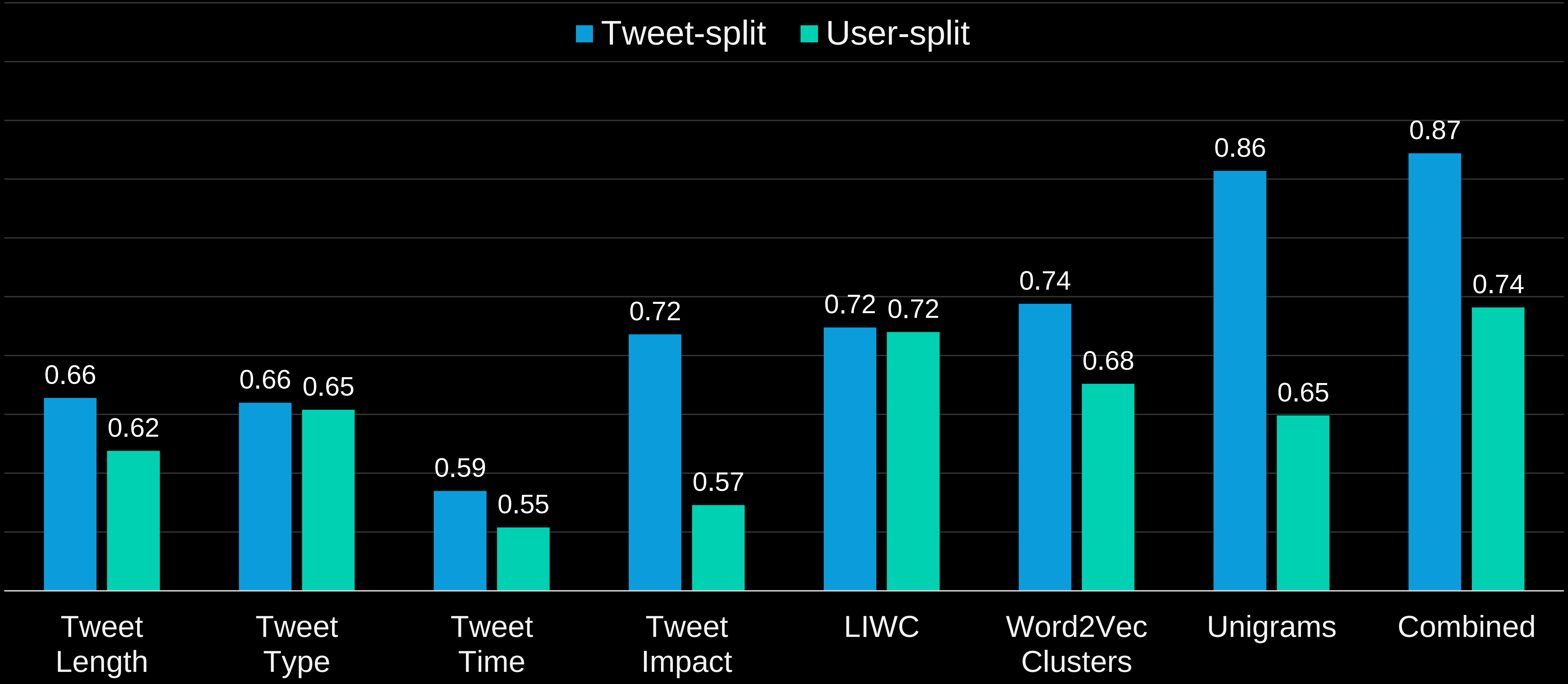


TechAtBloomberg.com

Bloomberg

Engineering

Prediction



TechAtBloomberg.com

Bloomberg Engineering

Analysis

Subsampled data from each account

- 1 signed – 9 unsigned
- Each account contributes at most 100 tweets

Aim is that no single user dominates

- the data set
- any label

Tweet Features – Mean Values

Feature	Owner	Staff
# Chars	105.4	102.4
# Tokens	23.2	21.4
Contains URL	45.7%	73.9%
@-Reply	4.2%	9.5%
Sent on Weekends	23.5%	20.7%
# Retweets	29.4	38.0
# Likes	82.3	79.1

* All differences between means shown in this table are significant at p .001, Mann-Whitney U test, Simes corrected

Analysis

- Signed tweets are more likely to be:
  — Longer
  — Sent on weekends
  — More liked, but less retweeted
  — Congratulations, condolences and support
  — More personal pronouns
  — More function words
  — More positive and negative sentiment
- No features are correlated with unsigned tweets:
  — More generic usage
- Other feature analysis in paper

Tweet Features – Mean Values

Feature	Owner	Staff
# Chars	105.4	102.4
# Tokens	23.2	21.4
Contains URL	45.7%	73.9%
@-Reply	4.2%	9.5%
Sent on Weekends	23.5%	20.7%
# Retweets	29.4	38.0
# Likes	82.3	79.1

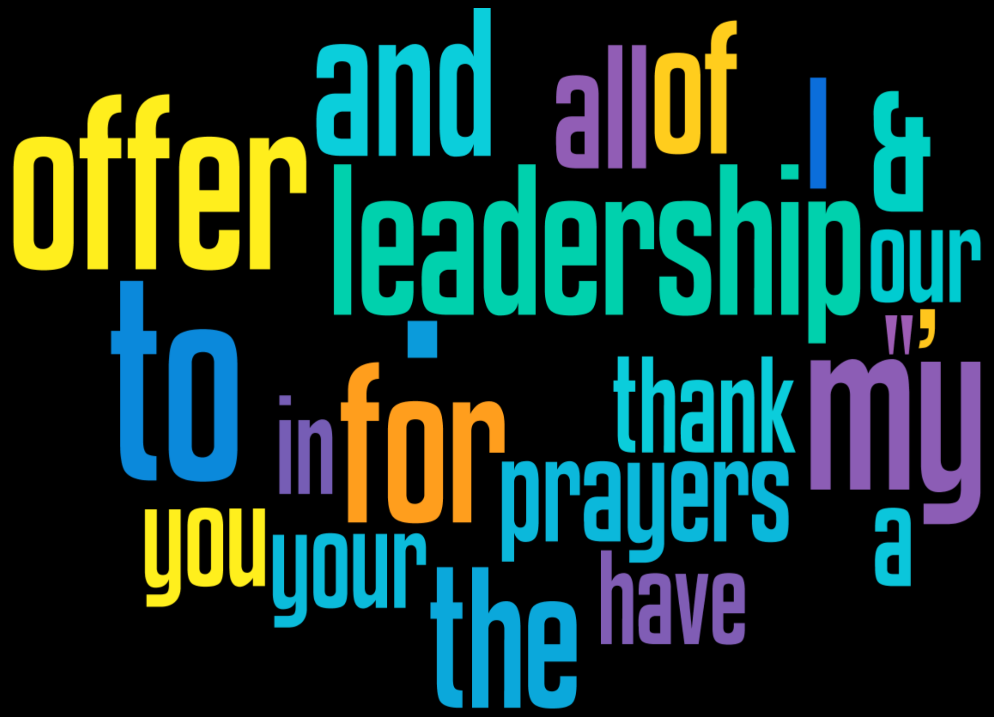
* All differences between means shown in this table are significant at p .001, Mann-Whitney U test, Simes corrected

TechAtBloomberg.com

Bloomberg
Engineering

Analysis

- Signed tweets are more likely to be:
  — Longer
  — Sent on weekends
  — More liked, but less retweeted
  — Congratulations, condolences and support
  — More personal pronouns
  — More function words
  — More positive and negative sentiment
- No features are correlated with unsigned tweets:
  — More generic usage
- Other feature analysis in paper



Unigrams

TechAtBloomberg.com

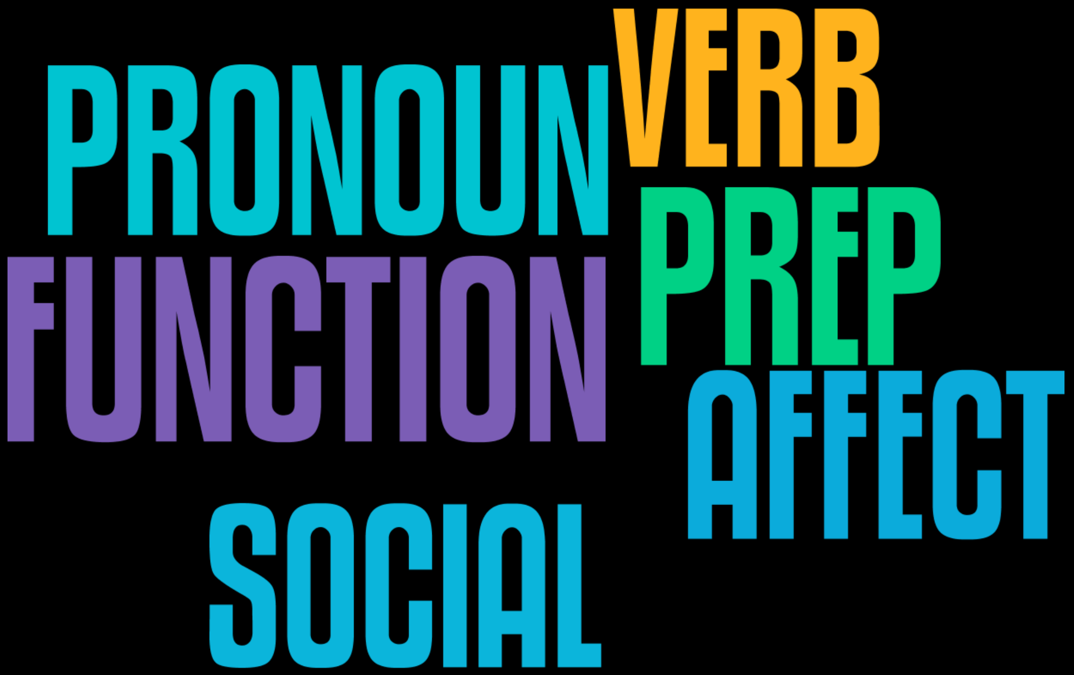
Bloomberg
Engineering

Analysis

- Signed tweets are more likely to be:
  — Longer
  — Sent on weekends
  — More liked, but less retweeted
  — Congratulations, condolences and support
  — More personal pronouns
  — More function words
  — More positive and negative sentiment

- No features are correlated with unsigned tweets:
  — More generic usage

- Other feature analysis in paper



LIWC Topics

Takeaways

Not all tweets from a single account are posted by a single person

New data set released for research

- https://github.com/danielpreotiuc/signed-tweets

We are able to predict type of author with good precision

- Even for accounts unseen in training
- Different features transfer better to unseen users

Use case results provide insight into the behavior of politicians

We are hiring:

- NYC – http://careers.bloomberg.com/job/detail/74022
- London – http://careers.bloomberg.com/job/detail/74154

TechAtBloomberg.com

Bloomberg

Engineering