

On Predicting Deletions of Microblog Posts

Mossaab Bagdouri
Douglas W. Oard

Computer Science
iSchool & UMIACS

{ University of Maryland }
College Park, USA

mossaab@umd.edu
oard@umd.edu

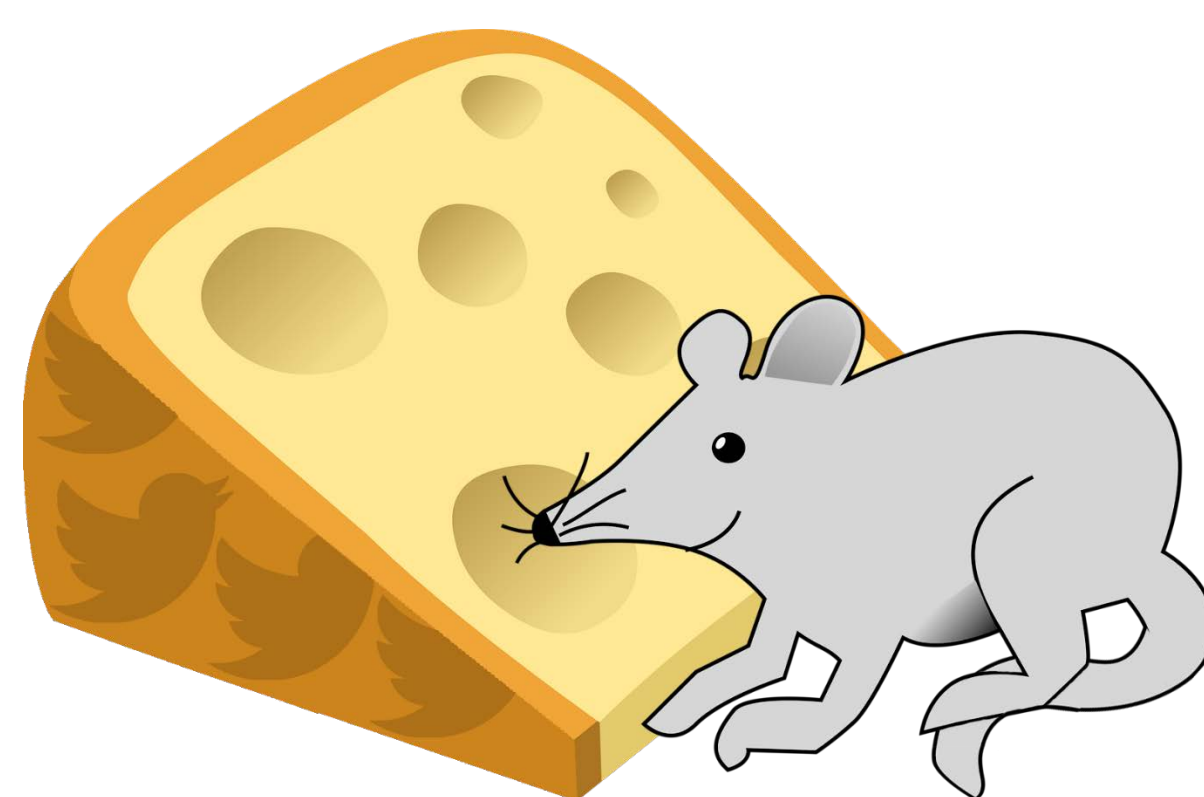


Introduction

Sponsored by Qatar National Research Fund (NPRP 6-1377-1-257) and by SIGIR Student Travel Grant

Why Predict Deletions?

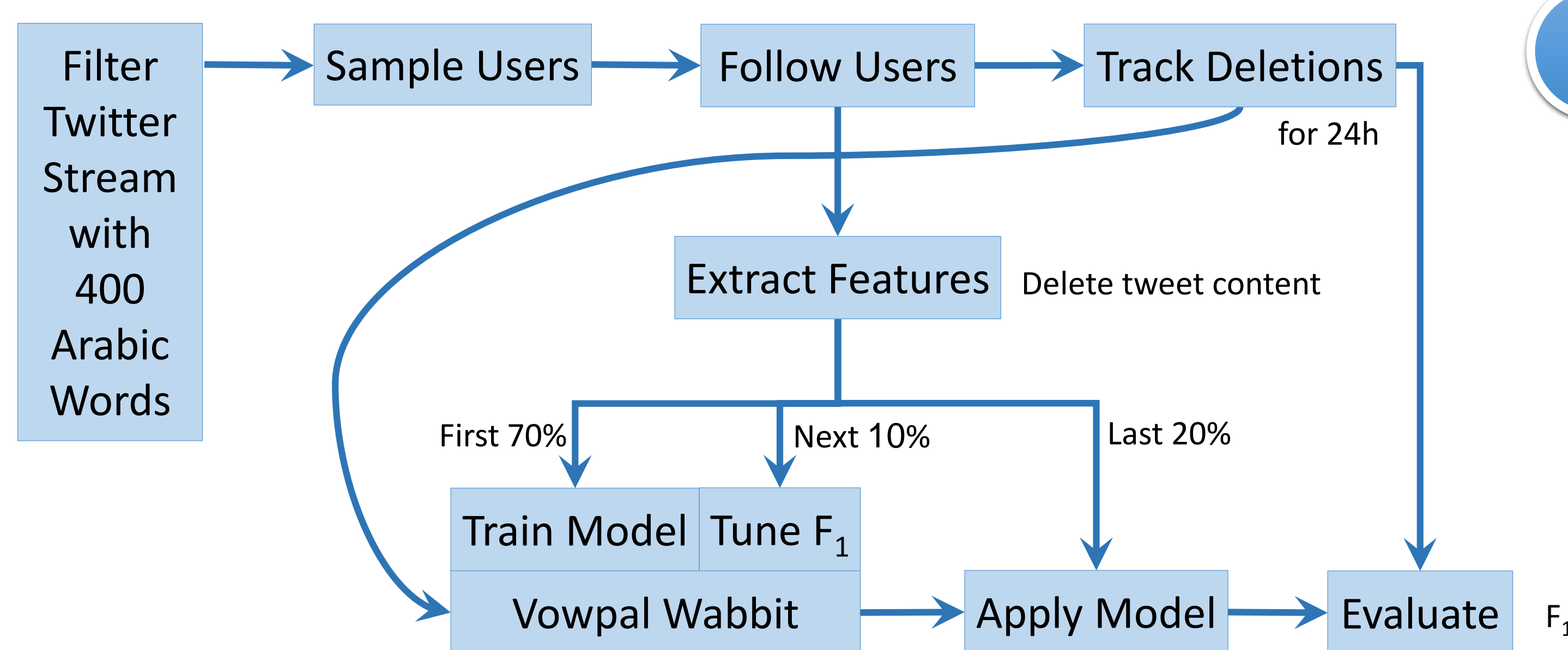
- Regret avoidance
- Censorship avoidance
- Collection persistence



How Do Deletions Occur?

- Delete own tweet
- Make a profile private
- Suspend an account
- Cascade RT deletions

Experiments

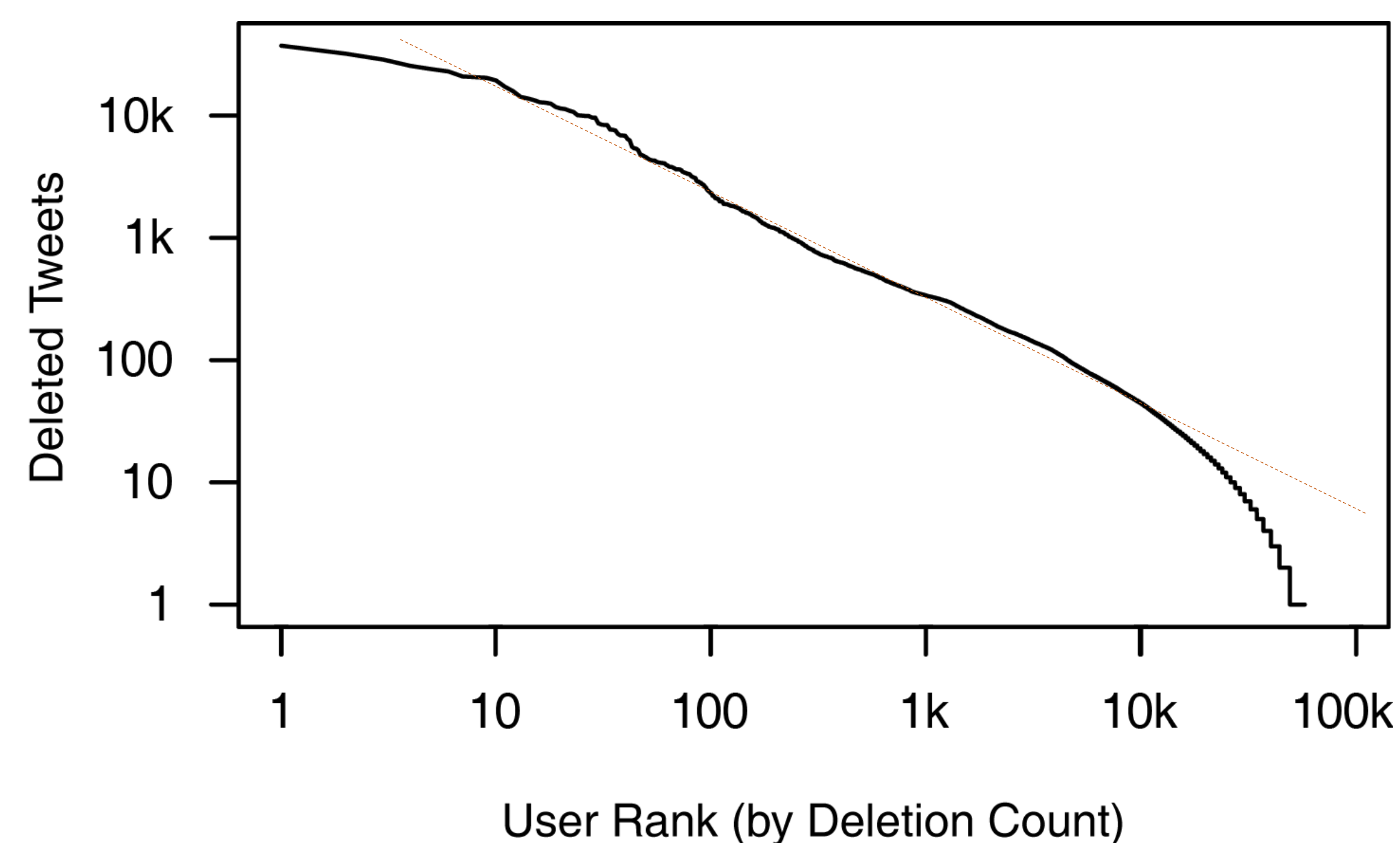
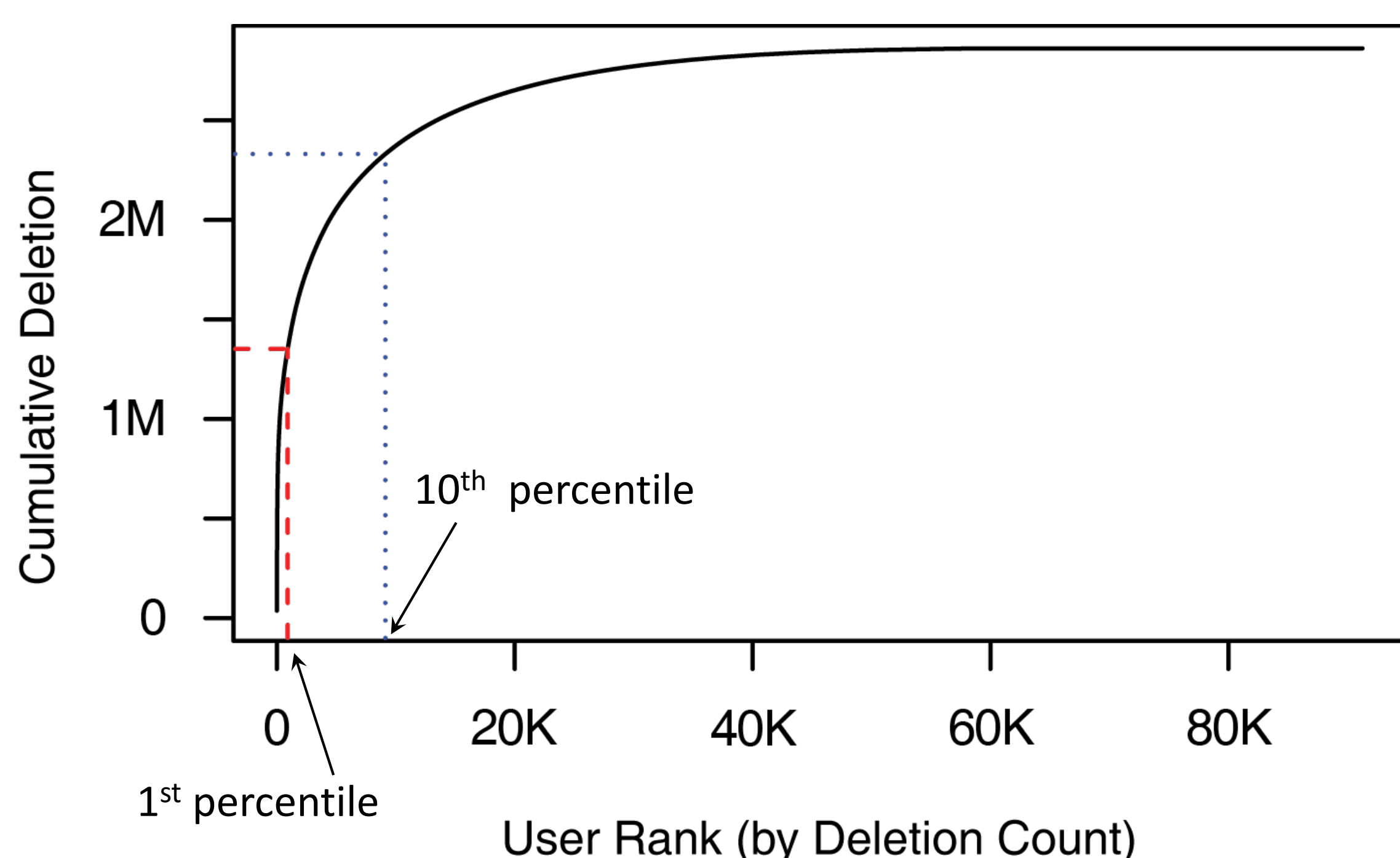


Data

	Feature Design	Evaluation
Streaming started	Oct 24, 2014	Dec 21, 2014
Streaming ended	Nov 21, 2014	Jan 22, 2015
Users followed	95,000	180,000
Users who tweeted	91,283	179,425
Number of tweets	80,8239,916	415,582,993
Labeled tweets	78,527,525	406,140,249
Deletion rate	3.64%	2.33%
Deletion rate by user	3.55%±9.15%	2.88%±7.47%

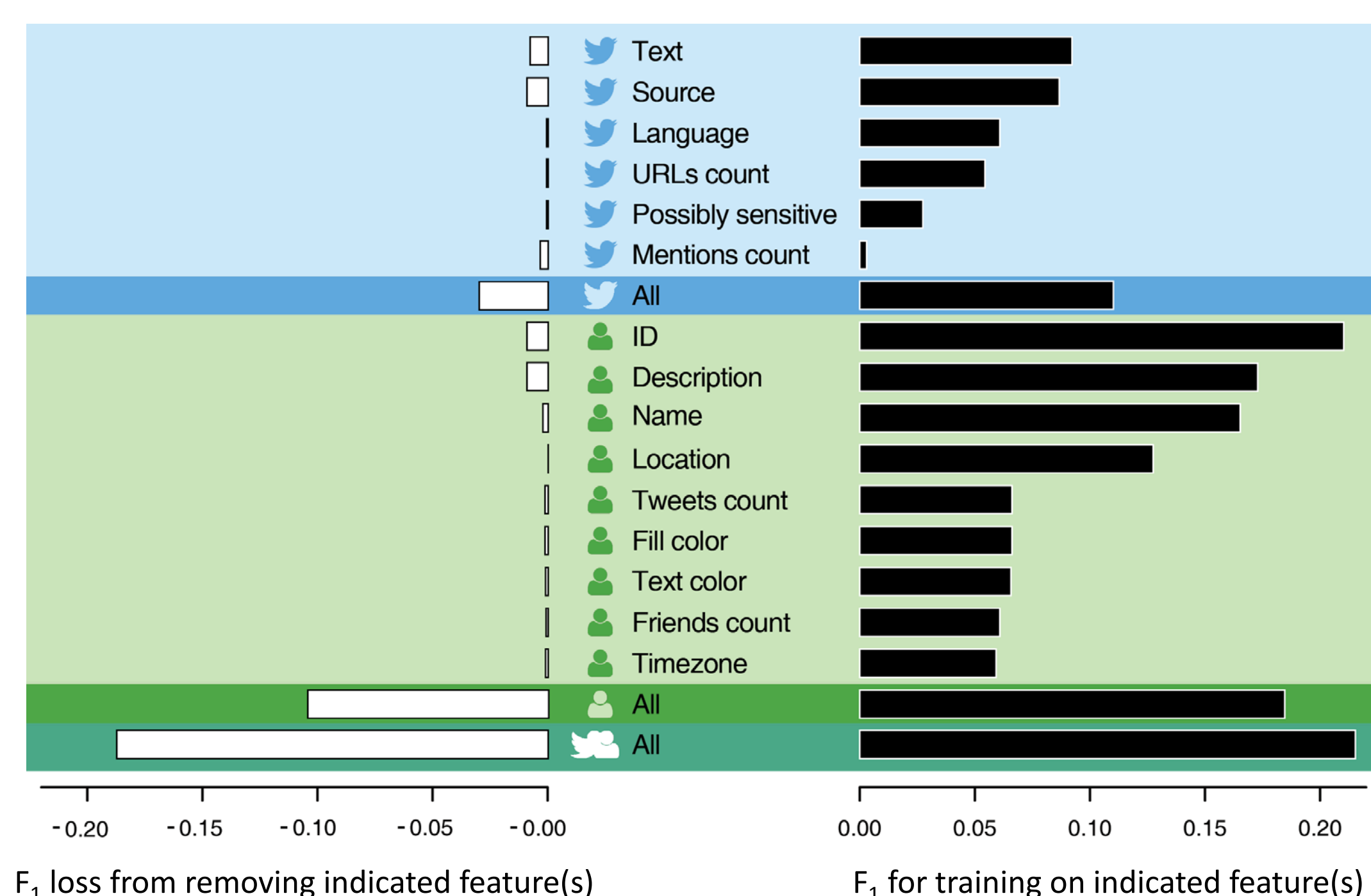
Naïve Features and Evaluation

- Petrovic et al. → $F_1 = 0.39$
- User ID: → $F_1 = 0.46$



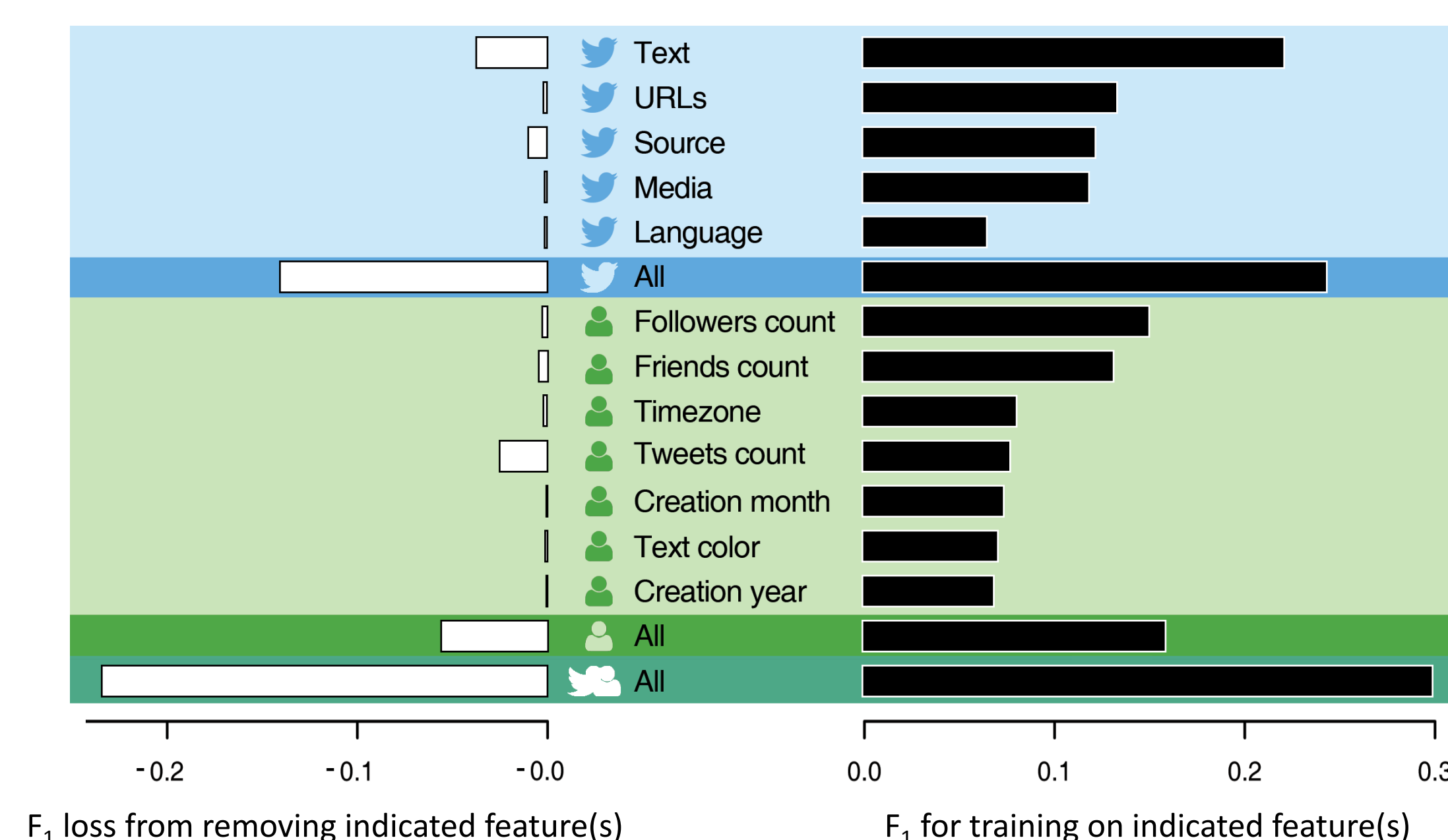
Excluding Retweets and Outliers

- Exclude Retweets (65% of deletions)
- Exclude 2% users (34% of non RT deletions)



Separating Users

- Goal: Neutralize the effect of user ID
- Training: 70% of users • Testing : 20% of users
- F_1 optimization: 10% of users



Conclusion

- User ID is a strong feature
- Different tasks ⇒ Different evaluation designs

Future Work

- Study different deletion types
- Study language-dependent features