# Inferring Gender from Twitter Data

**Xinyi Liu** [*1] **Mingren Shen** [*1] **Faust Shi** [*1]

## Abstract

The purpose of this document is to provide both the basic paper template and submission guidelines. Abstracts should be a single paragraph, between 4–6 sentences long, ideally. Gross violations will trigger corrections at the camera-ready phase.

## 1. Introduction

## 2. Relate Work

### 2.1. Twitter Gender Inference

Automatically inferring user gender from Twitter is heavily investigated by both academic socitey and industry because gender is one of the most important demographic property of the user. Generally, there are 3 main approaches used for deriving Twitter users gender information: (1) profile-based (2) content-based (3) hybrid (Beretta et al., 2015).

#### 2.1.1. Profile-based Gender Inference

Profile-based methods use the meta-data of the user's account in Twitter to help determine the gender of the users (Sloan et al., 2015). In Twitter, a user can show his name, description, location, followers and friends publicly. Although Twitter does not check the authenticity of profile information, several studies have proven that most Twitter users provide their real name and real gender in their public profile (Cesare et al., 2017). The simplest and best feature of profile information is users' first name. Previous studies have shown that by comparing name record from nation demographic survey, first name based gender classifier can achieve real good performance (Sloan et al., 2013; Mislove et al., 2011) .There are several mature services like genderize.io and packages [1] [2] inferring gender using only first name. For example `genderize.io` provides API that can be used to determine gender of a first name with the help of a database contains 216286 distinct names across 79 countries and 89 languages. Generally, the profile-based method is considered as the benchmark of gender inference due to the high efficacy it can achieve. For example, Liu et al. use first name as the main feature to infer gender in Twitter and they obtain the accuracy around 85% (Liu & Ruths, 2013).

[*]Equal contribution [1]University of Wisconsin,Madison , USA. Correspondence to: Xinyi Liu <xliu636@wisc.edu>.

*CS 760 Final Project*, UW-Madison, 2017. 

[1]`https://github.com/tue-mdse/genderComputer`

[2]`https://github.com/muatik/genderizer`

## Acknowledgements

We would like to sincerely thank Prof. Liang, Depratment of Computer Science at UW-Madison, for valuable suggestions to implement this project and Prof. Huang, Department of Geography at UW-Madison, for the help of collecting Twitter data-set and annotations of user genders.

### Software and Data

We provide all our data and program in Github and you can check them online `https://github.com/iphyer/cs760_TwitterDemographics`. We mainly use scikit-learn (Pedregosa et al., 2011) as our machine learning program library and processing data by Pandas (McKinney, 2015).

## References

Beretta, Valentina, Maccagnola, Daniele, Cribbin, Timothy, and Messina, Enza. An interactive method for inferring demographic attributes in twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 113–122. ACM, 2015.

Cesare, Nina, Grant, Christan, and Nsoesie, Elaine O. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 2017.

Liu, Wendy and Ruths, Derek. What's in a name? using first names as features for gender inference in twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, pp. 01, 2013.

McKinney, Wes. pandas: a python data analysis library. *see http://pandas. pydata. org*, 2015.

Mislove, Alan, Lehmann, Sune, Ahn, Yong-Yeol, Onnela, Jukka-Pekka, and Rosenquist, J Niels. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sloan, Luke, Morgan, Jeffrey, Housley, William, Williams, Matthew, Edwards, Adam, Burnap, Pete, and Rana, Omer. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):7, 2013.

Sloan, Luke, Morgan, Jeffrey, Burnap, Pete, and Williams, Matthew. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.