Twitter Hate Speech Analysis

ML4Geo

Robert Wright

29th July, 2021

Heidelberg University

Motivation

- *"Twitter Penalizes Record Number of Accounts for Posting Hate Speech"* [1]
- *"Twitter permanently suspends President Donald Trump"* [2]

  $\rightarrow$ use **machine learning** to analyse hate speech on Twitter

sometimes people say bad stuff on Twitter (image source)

Data

- dataset provided by Founta et al. [3]
- ~100,000 Tweets, 4% geotagged
- **downside**: only contains Tweet-IDs and labels (*abusive, hateful, normal, spam*)

  $\rightarrow$ **Twitter API** to download meta-info (e.g. text, coordinates, place)
- extensive **data preparation**

Project Proposal

- **Random Forest Classifier**
- single words as features (enormous number) $\rightarrow$ **overfitting**
- reduce dimensionality by taking 1000 (?) most frequent words
- aim: create (normalized) **heat map** of hate and investigate **spatial distribution**

Project Proposal

- **Random Forest Classifier**
- single words as features (enormous number) $\rightarrow$ **overfitting**
- reduce dimensionality by taking 1000 (?) most frequent words
- aim: create (normalized) **heat map** of hate and investigate **spatial distribution**

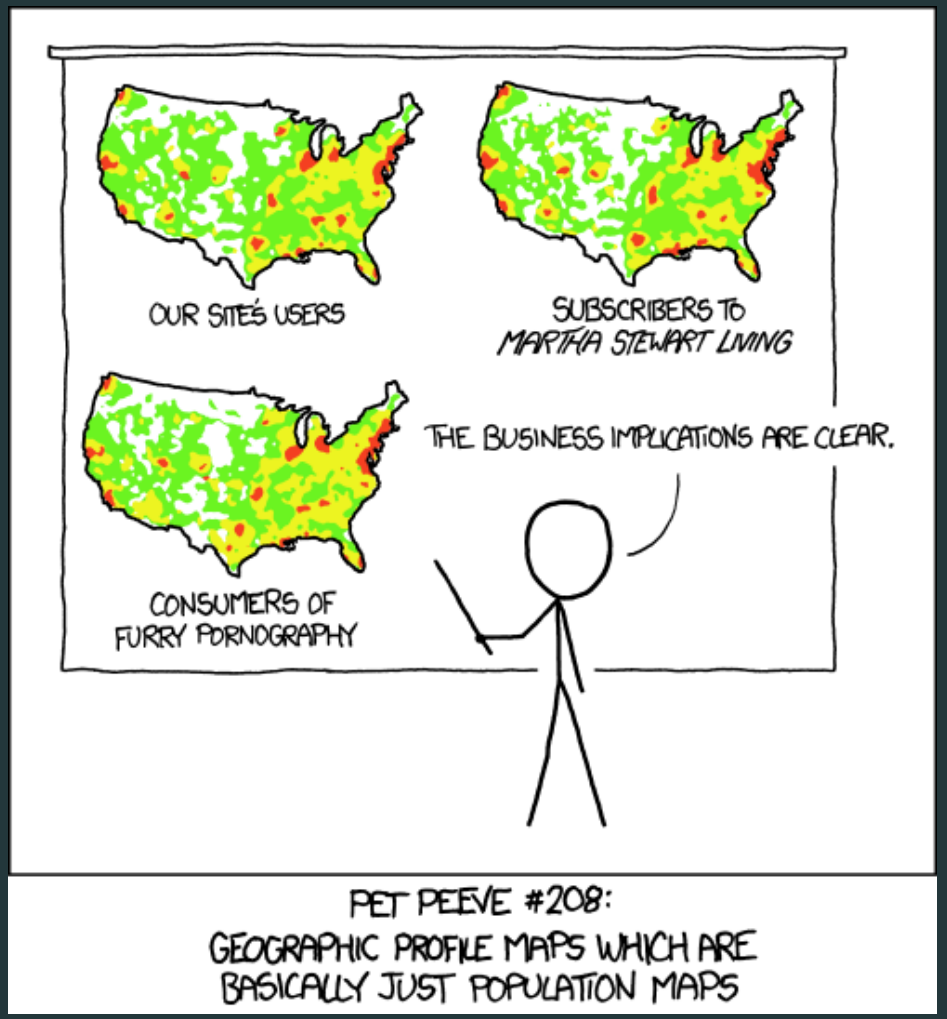

image source

References i

[1] Kurt Wagner. *Twitter Penalizes Record Number of Accounts for Hate Speech*. TIME. URL: https://time.com/6080324/twitter-hate-speech-penalties/ (visited on 07/28/2021).

[2] Ben Collins and Brandy Zadrozny. *Twitter permanently suspends President Donald Trump*. NBC News. URL: https://www.nbcnews.com/tech/tech-news/twitter-permanently-bans-president-donald-trump-n1253588 (visited on 07/28/2021).

References ii

[3] Antigoni-Maria Founta et al. “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior.” In: *arXiv:1802.00393 [cs]* (Apr. 15, 2018). arXiv: 1802.00393. URL: http://arxiv.org/abs/1802.00393 (visited on 07/23/2021).