KU LEUVEN

Understanding & Correcting Selection Bias in the Sentiments derived from Flemish tweets

Statistics Flanders
May 24 2022

Jehoul Astrid
Tonglet Jonathan

Table of content

1. Problem setting

- Surveys = costly, time-consuming, and subject to bias
- Social media = more representative of true opinion[1]

- **Twitter**: Academic Research API
  - Query tweets
  - No demographic attributes available

[1] Biffignandi et al., 2018

1.1 Twitter is a biased source of information

- Demographics of Twitter population differ from those of general population
  - 74.7% men[1]
  - Young people[2]

  ⇒ **Selection bias**

- Demographic attributes of Twitter users aren't available

  → Use machine learning to infer them

<u>Problem</u>: How to measure this bias and correct it?

1 https://www.xavierdegraux.be/sociale-netwerken-belgie-statistieken-2021
2 Vandendriessche et al., 2020

1.2 Research questions

- How can demographic labels be assigned efficiently and with minimal supervision to a sample of Twitter users?
- How does the population distribution of Flemish Twitter users differ from census data in terms of gender, age, and location?
- Which methods are best suited to correct the selection bias present in Twitter users datasets?

2. Approach

Target variables:

- Gender *{Male, Female}*
- Age category *{-18, 19-29, 30-39, 40+}*
- Location *{Antwerp, Limburg, Flemish Brabant, East Flanders, West Flanders, Brussels-Wallonia, Foreign countries}*

Inferring demographic attributes of Twitter users → Correcting the selection bias

2.1 Dataset

- Few public datasets (mainly English)

- 1,2M tweets and 28k user profiles
  - *Timeframe*: 2019-2020
  - *Language*: Dutch
  - *Geolocation*: Belgium

- **Hand-labeling**: costly, time-consuming, and not scalable
  - *Test set*: 2% labeled by 14 student annotators
  - *Training set*: alternative solution needed

    ⇒ **Weak supervision**

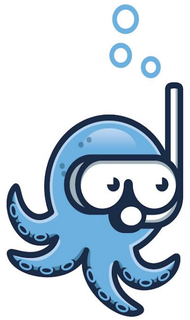
3. Demographic Inference

Keyword searches

Regular expressions

Third-party models

Machine Learning models



Gender, Age & Location

3.1 Heuristics & knowledge bases

Keyword searches & regular expressions

Age

- Keyword list *'twenties', 'grandpa'*
- Regular expressions

Gender

- Keyword list *'he/him', 'sister'*
- Dictionary of first names

Location

- Zip codes
- Town names (& W-Eu countries + capitals)

3.2 Third-party models (gender)

VGG-Face[1]: face detection + gender prediction

CLIP[2]: token assignment to image

- Woman 0.01
- Man 0.90
- Object 0.09

- Woman 0.01
- Man 0.24
- Object 0.75

[1] Parkhi et al., 2015; Serengil & Ozpinar, 2020, 2021
[2] Radford et al., 2021; https://github.com/openai/CLIP

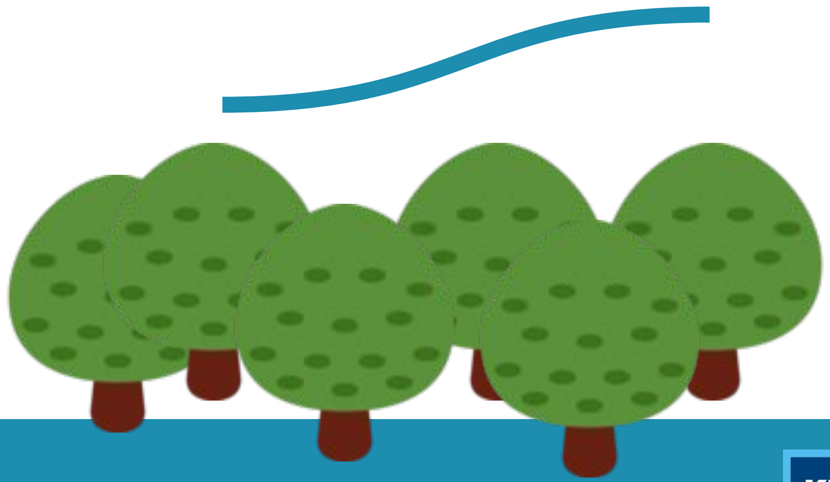
3.3 Machine Learning Classifiers

Features:

- Common terms in profile descriptions and tweets
- Topics discussed
- Celebrities followed (politicians, artists, football clubs, …)
- .nl/.be + account metadata

Models:

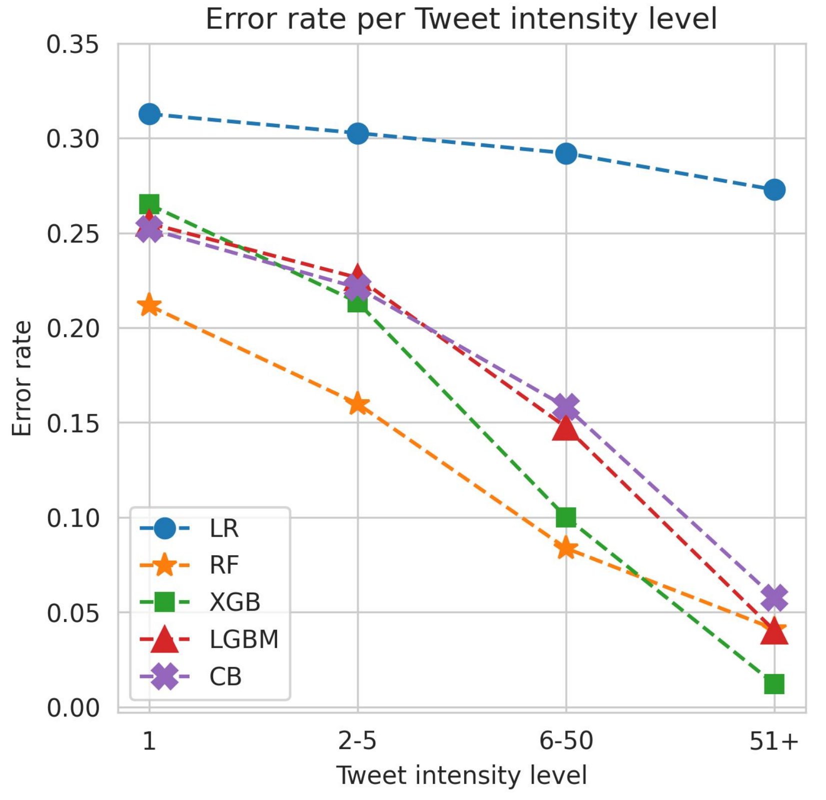
- Logistic regression: multi-class & ordinal
- Tree ensembles: RF, XGB, LGBM, and CB



4. Results

- Accuracy of the predictions:
  - Gender (2 categories): 92 %
  - Age (4 categories): 55%
  - Location (7 categories): 75%

- Better results on users with many tweets



4.1 Top features per predicted category

Female:

- Emojis: ✨🦋💕❤️🚺🌻👩🤍
- Description: fashion, lezen

Male:

- Description: cloud, software, developer, gamer, guy, echtgenoot/husband
- Follows: @ElevenSportsBEn/f, @KVCWesterlo

40+:

- Tweets about politics + mentioning
- Tweet content: @torfsrik, @groen, @kristofcalvo, @vlbelang, @phroose, @cdenv, @spa, @jdeceulaer, @bartdewever

4.1 Top features per predicted category

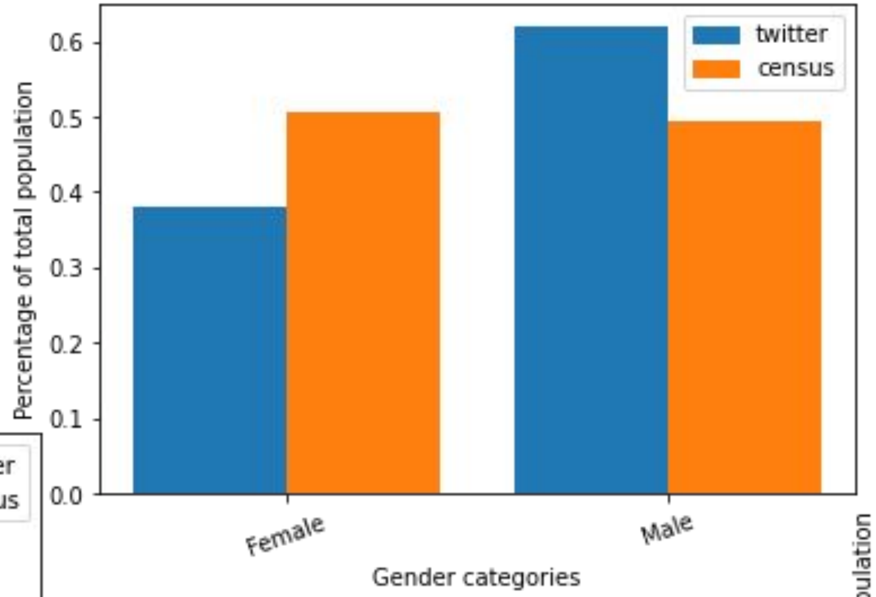
Foreign:

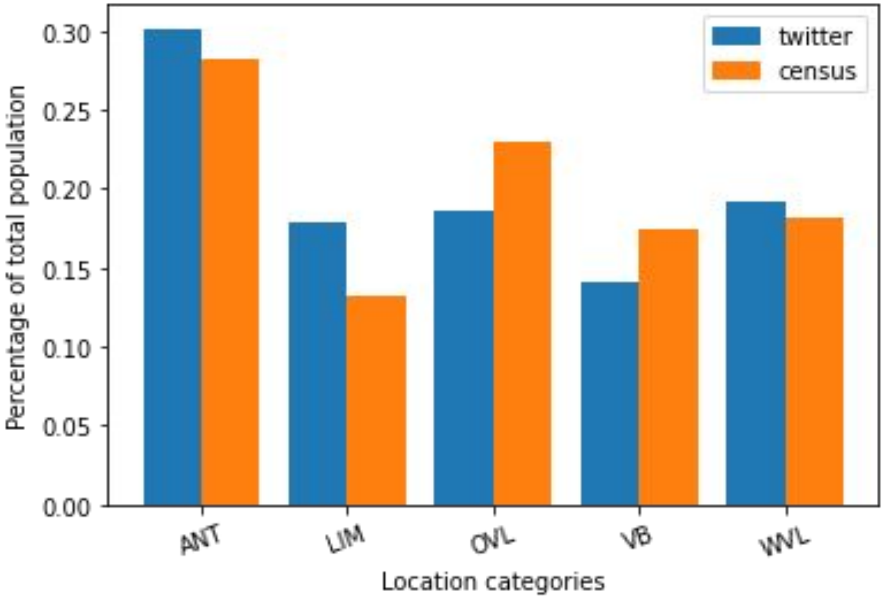
- .nl hyperlink in profile
- Follow Dutch celebrities/sports clubs

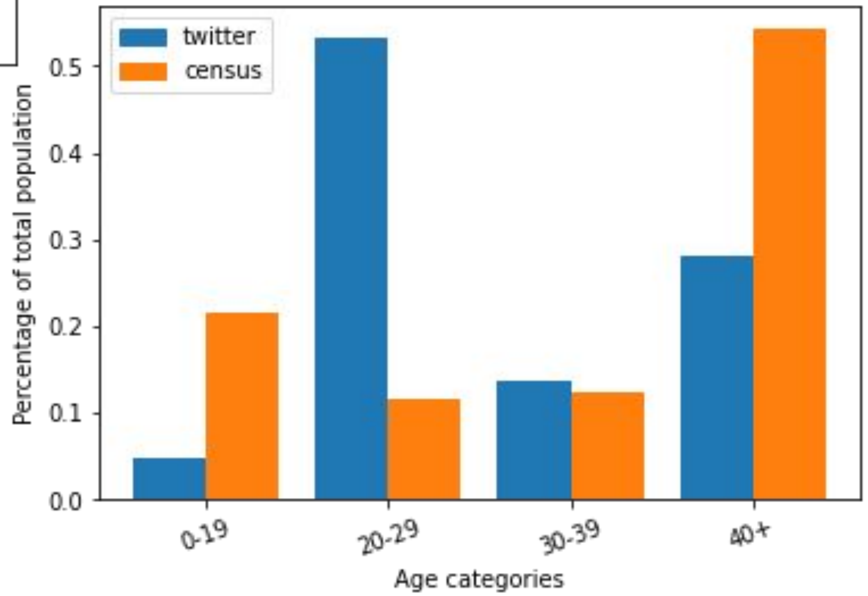
Other location categories:

- Antwerpen: @Stad_Antwerpen
- West-Flanders: @ClubBrugge
- East-Flanders: @UGent, @KAAGent
- Flemish-Brabant: @KULeuven, @PolitieLeuven
- Limburg: @KRCGenkOfficial

4.2 Compared to census







4.3 Limits

- No guarantee to get sufficient labels for all categories
  - Over-representation of users in their twenties
  - Hurts the performance

- We considered users with geolocated tweets only (41% of all users)[1]

[1] Sloan and Morgan, 2015

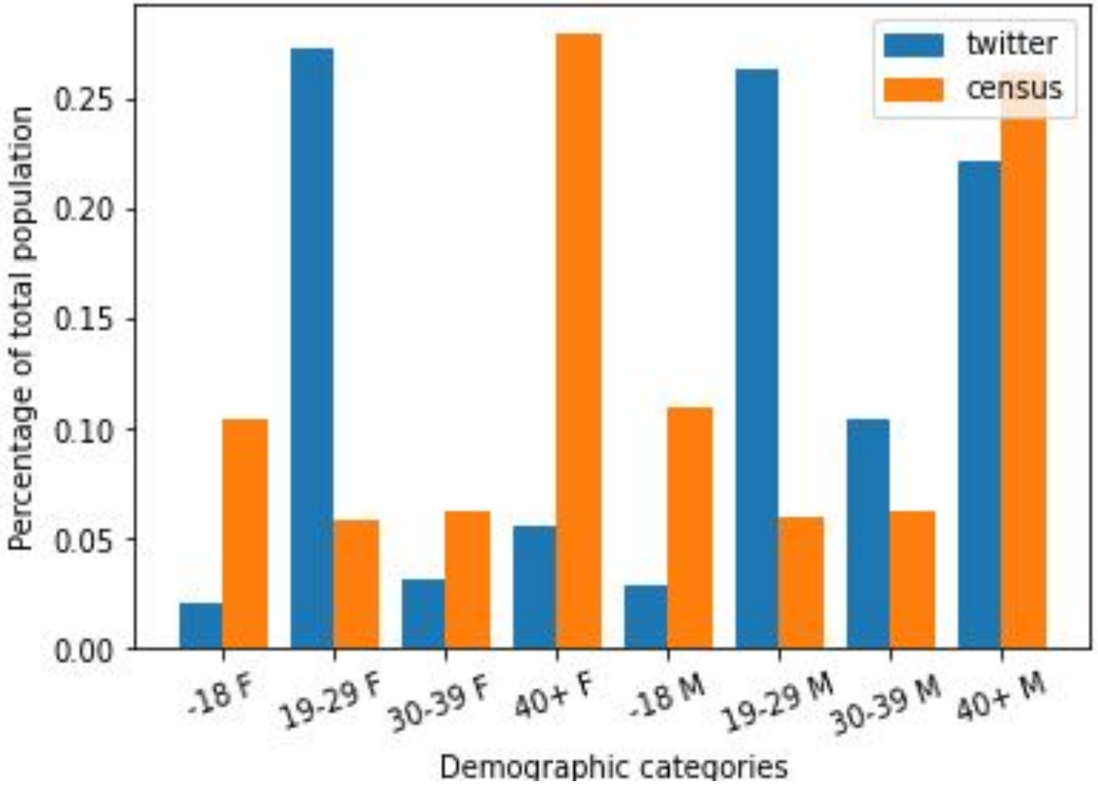
4.4 Future research

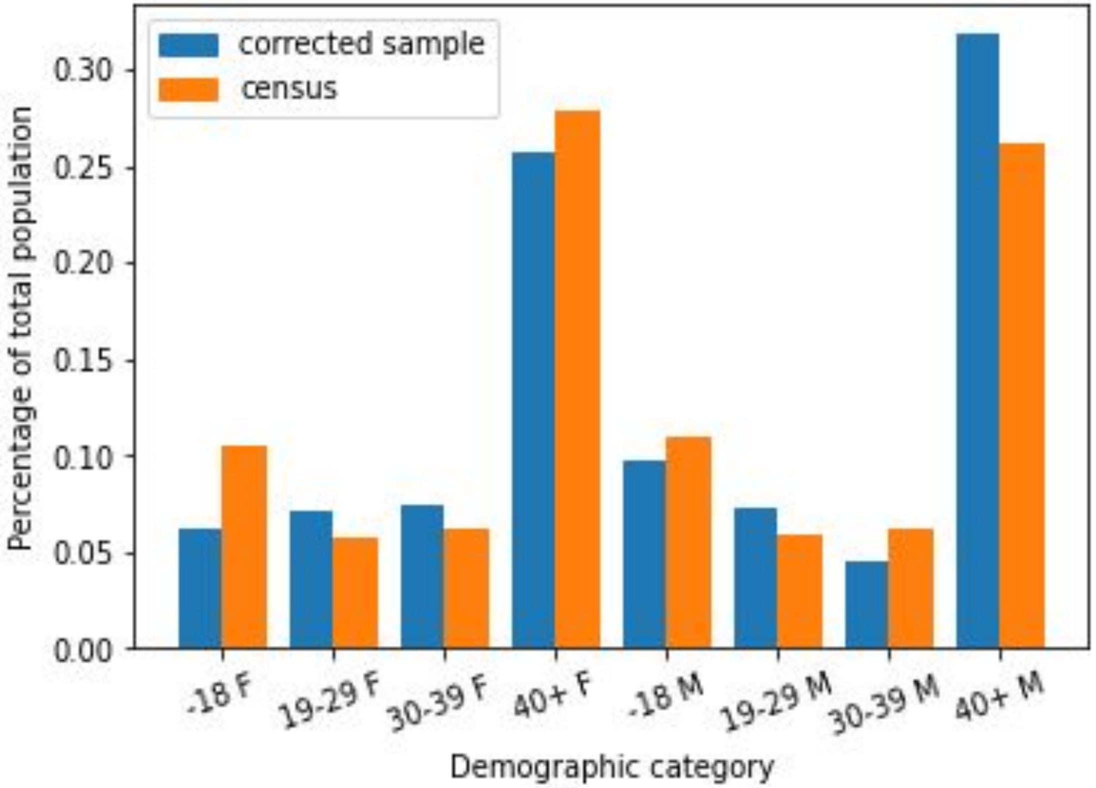
- Leveraging new attributes
  - Education level: High school, Bachelor, Master, …
  - Income level
  - More fine-grained age and location categories

- More advanced labeling models to improve coverage and accuracy

5. Correction methods

Resampling[1]





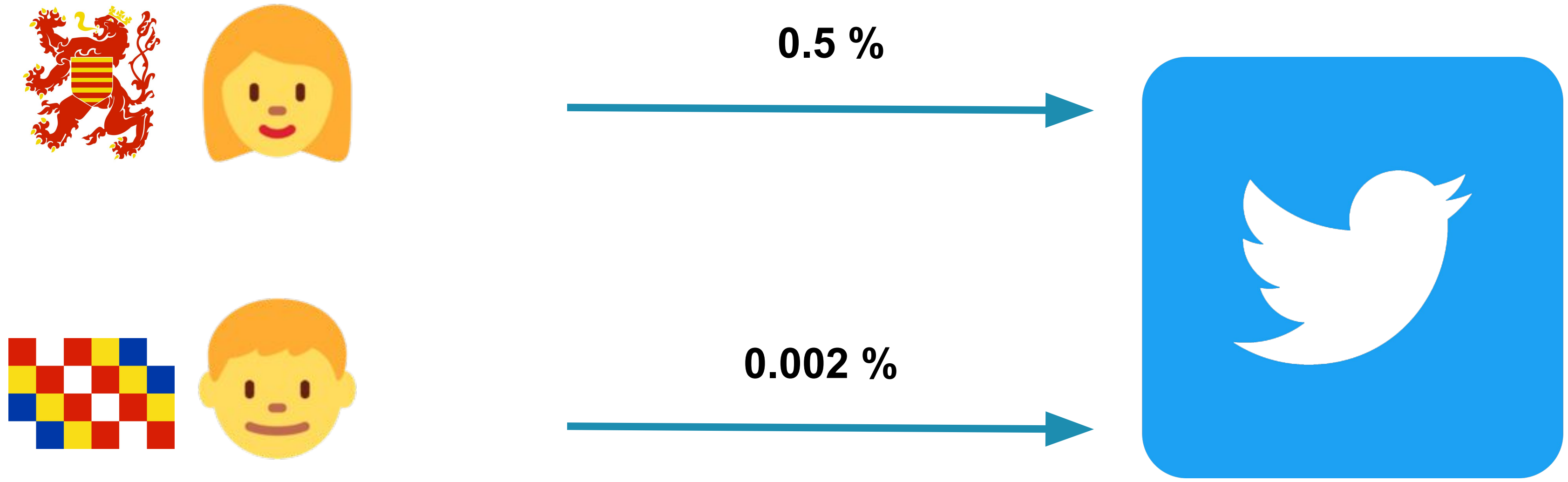
[1] Wang et al., 2020

5. Correction methods

Reweighting[1]

Computes probability that a demographic group joins Twitter

Assign weights based on inclusion probabilities



[1] Wang et al., 2019

6. Conclusion

- Demographic inference is successful for gender and location
- Age prediction is more challenging
- Resampling methods allow to remove the selection bias
- More experiments are needed for reweighting methods

Link to code https://github.com/jtonglet/Twitter-Selection-Bias/

Remarks & suggestions?

Thank you for your attention!

Bibliography

Biffignandi, S., Bianchi, A., & Salvatore, C. (2018). Can big data provide good quality statistics? A case study on sentiment analysis on Twitter data. In *Int. Total Surv. Error Workshop ITSEW-2018 DISM-Duke Initiat. Surv. Methodol.*

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)* (pp. 41.1-41.12). British Machine Vision Association Press. URL: https://dx.doi.org/10.5244/C.29.41. doi:10.5244/C.29.41.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (pp. 8748-8763).* PMLR.

Serengil, S. I., & Ozpinar, A. (2020). LightFace: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 23-27). IEEE. URL: https://doi.org/10.1109/ASYU50717.2020.9259802. doi:10.1109/ASYU50717.2020.9259802.

Serengil, S. I., & Ozpinar, A. (2021). Hyperextended LightFace: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-4). IEEE. URL: https://doi.org/10.1109/ICEET53442.2021.9659697. doi:10.1109/ICEET53442.2021.9659697.

Sloan L, Morgan J (2015) Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. PLOS ONE 10(11): e0142209. https://doi.org/10.1371/journal.pone.0142209

Bibliography

Vandendriessche, K., Steenberghs, E., Matheve, A., Georges, A., & De Marez, L. (2020). imec.digimeter 2020, Digitale trends in Vlaanderen. URL: https://www.imec.be/sites/default/files/inline-files/DIGIMETER2020.pdf.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference* (pp. 2056-2067).

Wang, Z., Yu, Z., Fan, R., & Guo, B. (2020). Correcting biases in online social media data based on target distributions in the physical world. *IEEE Access*, 8 , 15256-15264.