Tweet Gendeer Classification

AmirMohammad Azadi

Advisors: Dr. Sauleh Etemadi, Hadi Sheikhi

July 2023

1 Repository

github repository: https://github.com/am-azadi/NLP$_{P}roject/tree/master$

2 Source

Twitter doesn't save user genders. so using api we can not gather data with gender labels. so I used a preprovided dataset and download it form the below link:
https://drive.google.com/uc?id=1rbQ5a95uyXl20TTECn3dS4dl42OTcmM_
$thereisasetoftweetswithgenderspecified.$

3 Data Format

Data is represented in four different directories. The data/raw path includes the raw tweets as a CSV file.
The data/clean path includes a CSV file clean_data that is the result of cleaning our raw data.

4 Preprocessing

After cleaning the data, we have a dataset that only have gender and text. we generate sentence and word tokenizations for each tweet and save it as json file in data/sentencebroken and data/wordbroken directories.

5 Statistics

5.1 Data Count

all tweets	men	female
12894	6194	6700

5.2 Sentence Count

all tweets	men	female
18881	9148	9733

5.3 Word Count

all tweets	men	female
221966	108163	113803

5.4 Unique Word Count

all tweets	men	female
26483	17200	16313

5.5 Common and Uncommon Unique Words

common	uncommon
7030	19453

5.6 Frequent Uncommon words

men	female
"'Million': 15, 'Himself': 15, 'Gospel': 15"	"'FALive': 18, 'makeup': 18, 'giveaway': 16"

5.7 TF-IDF

5.7.1 Male

about	people	think	would	there
1087.8163262354844	876.2169222964534	833.5178781371301	826.3372838942449	763.6570615641643

5.7.2 Female

about	people	there	still	makes
1301.660561307417	1131.5715682228483	871.9772121406415	821.5664200410638	806.1923506904919

5.8 Top Unique Words