Topics in *Wiki*-dataset

This document is part of the supplementary material for the following paper:

> Isar Nejadgholi and Svetlana Kiritchenko (2020) On Cross-Dataset Generalization in Automatic Detection of Online Abuse - WOAH2020 Shared Exploration: Bias and Unfairness in the Detection of Online Abuse

The following 20 topics are extracted from *Wiki*-dataset using LDA algorithm. Each topic is represented as a multinomial combination of 10 top words. The number of utterances that fall under each topic and the percentage of offensive utterances is reported.

Category 1: incoherent and mixture of general topics

Topic #0:
0.025*"know" + 0.019*"like" + 0.019*"thank" + 0.015*"think" + 0.014*"want" + 0.013*"look" + 0.012*"ll" + 0.012*"ve" + 0.012*"hi" + 0.011*"time"

31210 documents - 0.195 of dataset
0.18% labeled as *Personal Attack* and 36% of all *Personal Attack*s

---

Topic #1:
0.012*"time" + 0.010*"like" + 0.009*"peopl" + 0.009*"think" + 0.007*"life" + 0.007*"year" + 0.007*"idiot" + 0.007*"day" + 0.006*"know" + 0.006*"drink"

9598 documents - 0.060 of dataset
0.19% labeled as *Personal Attack* and 12% of all *Personal Attack*s

---

Category 2: coherent and high association with offensive language
Topic #2:
0.020*"suck" + 0.015*"year" + 0.010*"new" + 0.009*"citi" + 0.009*"school" + 0.008*"cock" + 0.008*"old" + 0.008*"dick" + 0.007*"pussi" + 0.007*"women"

6466 documents - 0.040 of all the documnets
0.18% labeled as *Personal Attack* and 8% of all *Personal Attack*s

---

Topic #7:
0.018*"english" + 0.017*"languag" + 0.016*"peopl" + 0.011*"countri" + 0.010*"amer-

ican" + 0.009*"nation" + 0.008*"term" + 0.008*"use" + 0.008*"word" + 0.007*"german"

7706 documents - 0.048 of dataset
0.06% labeled as *Personal Attack* and 3% of all *Personal Attack*s

---

Topic #8:
0.047*"kill" + 0.038*"live" + 0.027*"pro" + 0.024*"die" + 0.023*"eat" + 0.018*"jewish" + 0.017*"anti" + 0.017*"islam" + 0.016*"al" + 0.016*"israel"

997 documents - 0.006 of dataset
0.34% labeled as *Personal Attack* and 2% of all *Personal Attack*s

---

Topic #9:
0.024*"god" + 0.020*"book" + 0.019*"christian" + 0.009*"jesus" + 0.009*"cast" + 0.008*"king" + 0.008*"prime" + 0.008*"william" + 0.008*"presid" + 0.008*"japanes"

1102 documents - 0.007 of dataset
0.12% labeled as *Personal Attack* and 1% of all *Personal Attack*s

---

Topic #12:
0.014*"person" + 0.013*"editor" + 0.013*"admin" + 0.011*"attack" + 0.011*"say" + 0.010*"peopl" + 0.009*"like" + 0.009*"wikipedia" + 0.008*"accus" + 0.008*"user"

13949 documents - 0.087 of dataset
0.13% labeled as *Personal Attack* and 12% of all *Personal Attack*s

---

Topic #14:
0.187*"fuck" + 0.078*"shit" + 0.056*"ass" + 0.051*"stupid" + 0.045*"bastard" + 0.037*"em" + 0.033*"bitch" + 0.032*"moron" + 0.030*"cunt" + 0.027*"hate"

1294 documents - 0.008 of dataset
0.97% labeled as *Personal Attack* and 8% of all *Personal Attack*s

---

Topic #16:
0.040*"team" + 0.031*"footbal" + 0.029*"infobox" + 0.027*"award" + 0.023*"win" + 0.022*"gay" + 0.015*"air" + 0.015*"engin" + 0.015*"match" + 0.014*"station"

475 documents - 0.003 of dataset
0.13% labeled as *Personal Attack* and 0.0% of all *Personal Attack*s

---

Category 3: coherent and low association with offensive language
Topic #3:
0.047*"redirect" + 0.040*"talk" + 0.039*"utc" + 0.036*"categori" + 0.032*"film"

+ 0.017*"episod" + 0.013*"merg" + 0.012*"octob" + 0.012*"decemb" + 0.011*"januari"
3023 documents - 0.019 of dataset
0.04% labeled as *Personal Attack* and 1% of all *Personal Attack*s

---

Topic #4:
0.086*"page" + 0.082*"wikipedia" + 0.059*"edit" + 0.043*"talk" + 0.035*"help" + 0.033*"articl" + 0.028*"thank" + 0.022*"question" + 0.016*"ask" + 0.015*"revert"
10097 documents - 0.063 of dataset
0.04% labeled as *Personal Attack* and 3% of all *Personal Attack*s

---

Topic #5:
0.086*"sourc" + 0.021*"reliabl" + 0.015*"claim" + 0.012*"cite" + 0.012*"refer" + 0.012*"wikipedia" + 0.011*"inform" + 0.011*"fact" + 0.010*"publish" + 0.010*"research"

9439 documents - 0.059 of dataset
0.04% labeled as *Personal Attack* and 3% of all *Personal Attack*s

---

Topic #6:
0.034*"link" + 0.029*"list" + 0.023*"add" + 0.022*"page" + 0.014*"game" + 0.013*"inform" + 0.011*"articl" + 0.010*"date" + 0.010*"chang" + 0.009*"googl"

7807 documents - 0.049 of dataset
0.03% labeled as *Personal Attack* and 1% of all *Personal Attack*s

---

Topic #10:
0.102*"delet" + 0.048*"articl" + 0.045*"imag" + 0.033*"wikipedia" + 0.029*"tag" + 0.028*"copyright" + 0.025*"file" + 0.025*"notabl" + 0.024*"page" + 0.017*"use"

6308 documents - 0.040 of dataset
0.02% labeled as *Personal Attack* and 1% of all *Personal Attack*s

---

Topic #11:
0.017*"univers" + 0.017*"th" + 0.012*"law" + 0.012*"scienc" + 0.011*"theori" + 0.009*"capit" + 0.008*"centuri" + 0.008*"definit" + 0.008*"state" + 0.007*"student"

1875 documents - 0.012 of dataset
0.03% labeled as *Personal Attack* and 0.0% of all *Personal Attack*s

---

Topic #13:
0.028*"page" + 0.025*"discuss" + 0.024*"review" + 0.024*"talk" + 0.023*"thank" + 0.023*"request" + 0.020*"vertic" + 0.019*"comment" + 0.018*"templat" + 0.017*"wp"

7555 documents - 0.047 of dataset
0.02% labeled as *Personal Attack* and 1% of all *Personal Attack*s

---

Topic #15:
0.050*"articl" + 0.011*"think" + 0.011*"section" + 0.009*"wp" + 0.009*"discuss" + 0.008*"refer" + 0.008*"editor" + 0.008*"point" + 0.008*"need" + 0.007*"chang"

30128 documents - 0.189 of dataset
0.02% labeled as *Personal Attack* and 3% of all *Personal Attack*s

---

Topic #17:
0.054*"http" + 0.045*"com" + 0.033*"www" + 0.033*"org" + 0.026*"en" + 0.017*"state" + 0.015*"wiki" + 0.013*"unit" + 0.011*"compani" + 0.008*"html"

2757 documents - 0.017 of dataset
0.04% labeled as *Personal Attack* and 1% of all *Personal Attack*s

---

Topic #18:
0.121*"edit" + 0.097*"block" + 0.038*"vandal" + 0.030*"user" + 0.029*"account" + 0.028*"stop" + 0.027*"ip" + 0.023*"war" + 0.022*"page" + 0.021*"revert"

6559 documents - 0.041 of dataset
0.09% labeled as *Personal Attack* and 4% of all *Personal Attack*s

---

Topic #19:
0.130*"style" + 0.096*"px" + 0.069*"align" + 0.059*"color" + 0.052*"background" + 0.051*"pad" + 0.044*"middl" + 0.038*"border" + 0.033*"solid" + 0.023*"size"

1341 documents - 0.008 of dataset
0.06% labeled as *Personal Attack* and 1% of all *Personal Attack*s