Reyhane Shahrokhian
99521361
r_shahrokhian@comp.iust.ac.ir

Project Phaze1 Report
Natural Language Processing

2023-06-07

1 Gathering data

Due to the fact that tweets older than a week cannot be crawled using the official Twitter APIs, the source of data is actually the Telegram ”Farsi Twitter” channel. So, crawling data is a bit different and it can’t be done by a script or something like that.In fact i use the Export-Chat feature of telegram. After doing that , you will get json files that are all in data/row folder of the project.

2 Preprocessing and Labeling

To preprocess the data(cleaning data, word braking, sentence breaking, sentiment analysing) you should run related cells of project_phase1.ipynb which is in src folder. After running them you will get some files in your drive that are all in data/clean, data/wordbroken and data/sentencebroken folders of the project.

3 Statics

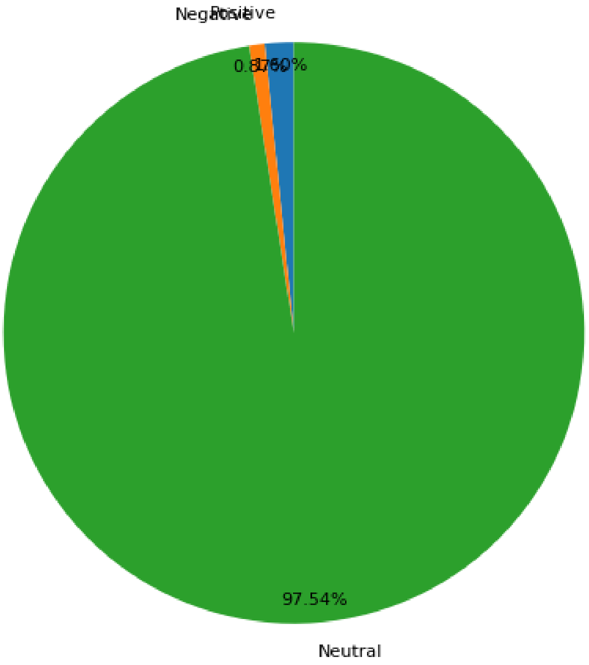
To show the static details(Data count of each label, word and sentence counts of labels, unique word count of labels, common and uncommon counts of words in comparison of labels, top10 words of each label and all words, top10 words of labels according to RNF and TF-IDF) you could run related cells of project_phase1.ipynb(statics section) which is in src folder. After running them you will get tables as csv file and charts as png in your drive that are all in stats folders of the project.

3.1 Statics of current data:

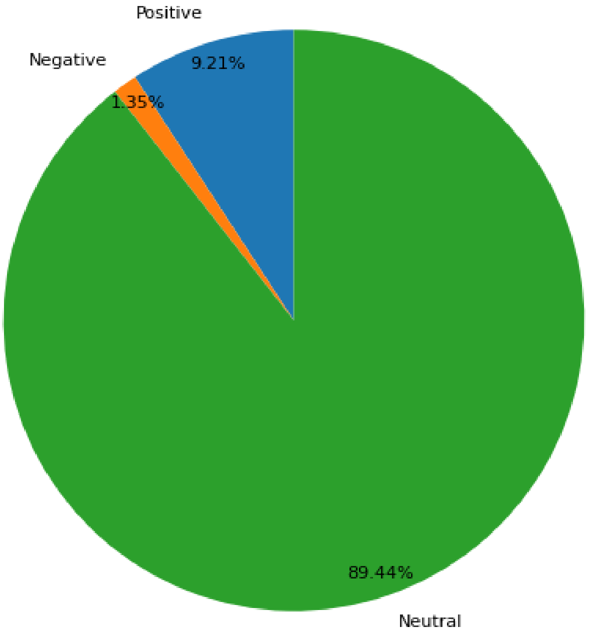
Label	Data count
2017Negative	39
2017Positive	72
2017Neutral	4396
2020Negative	69
2020Positive	469
2020Neutral	4555
2023Negative	67
2023Positive	550
2023Neutral	4995

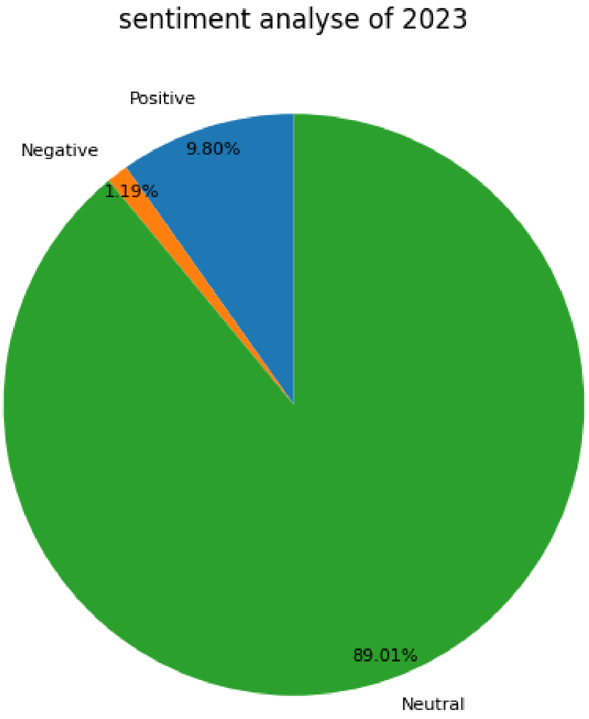
Label	word count	unique word count	sentence count
2017	79058	17212	5203
2020	118308	20444	6274
2023	194177	28396	9842

sentiment analyse of 2017



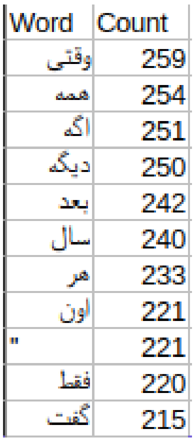
sentiment analyse of 2020



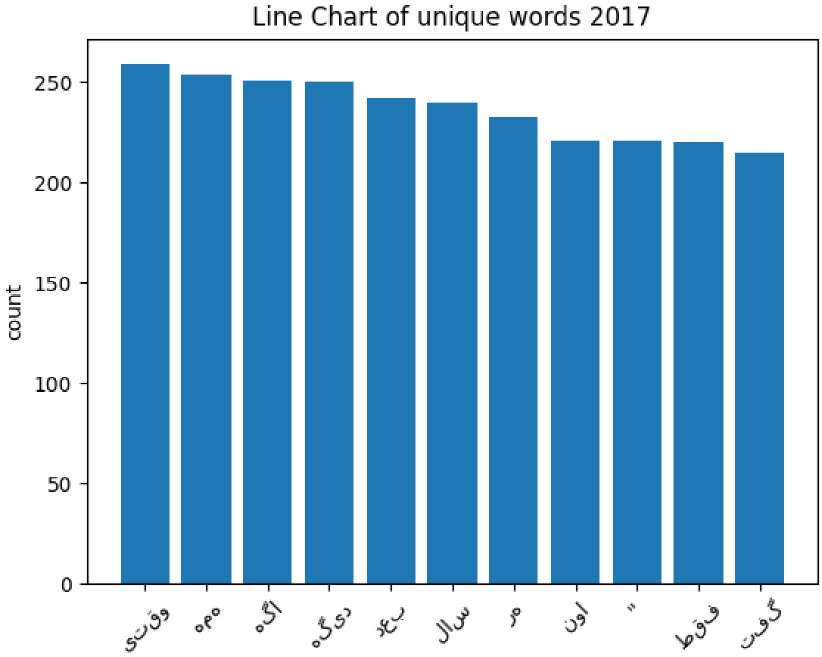

sentiment analyse of 2023
Positive
9.80%
Negative
1.19%
89.01%
Neutral

3.2 Top10 words of each label:

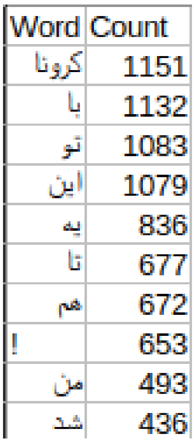
3.2.1 2017:



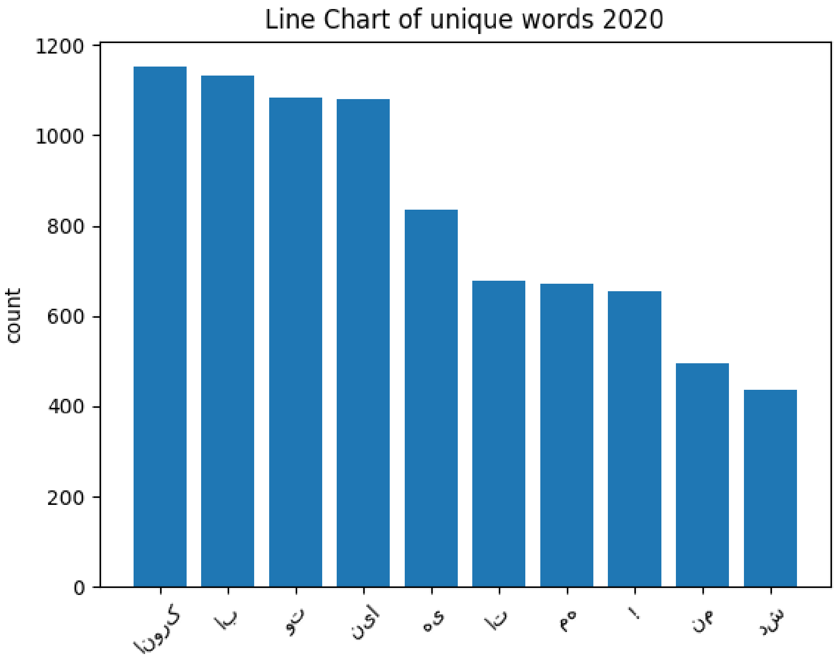
Word	Count
وقتی	259
همه	254
اگه	251
دیگه	250
بعد	242
سال	240
هر	233
اون	221
"	221
فقط	220
گفت	215



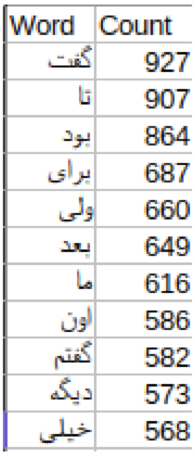
3.2.2 2020:



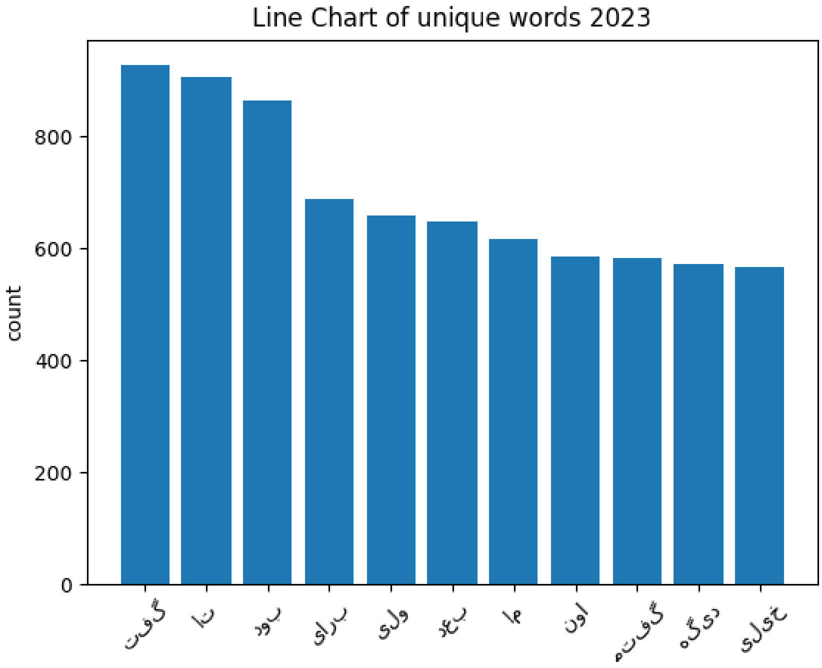
Word	Count
کرونا	1151
با	1132
تو	1083
این	1079
یه	836
تا	677
هم	672
!	653
من	493
شد	436



3.2.3 2023:

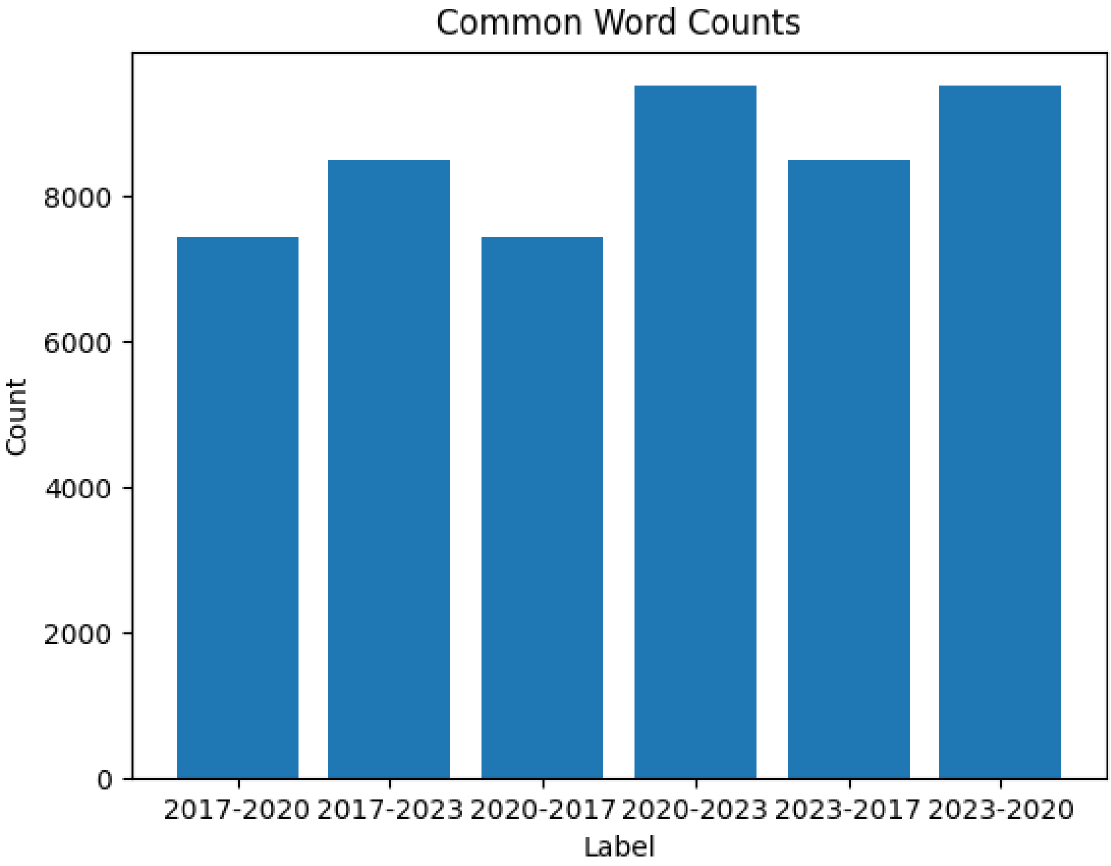


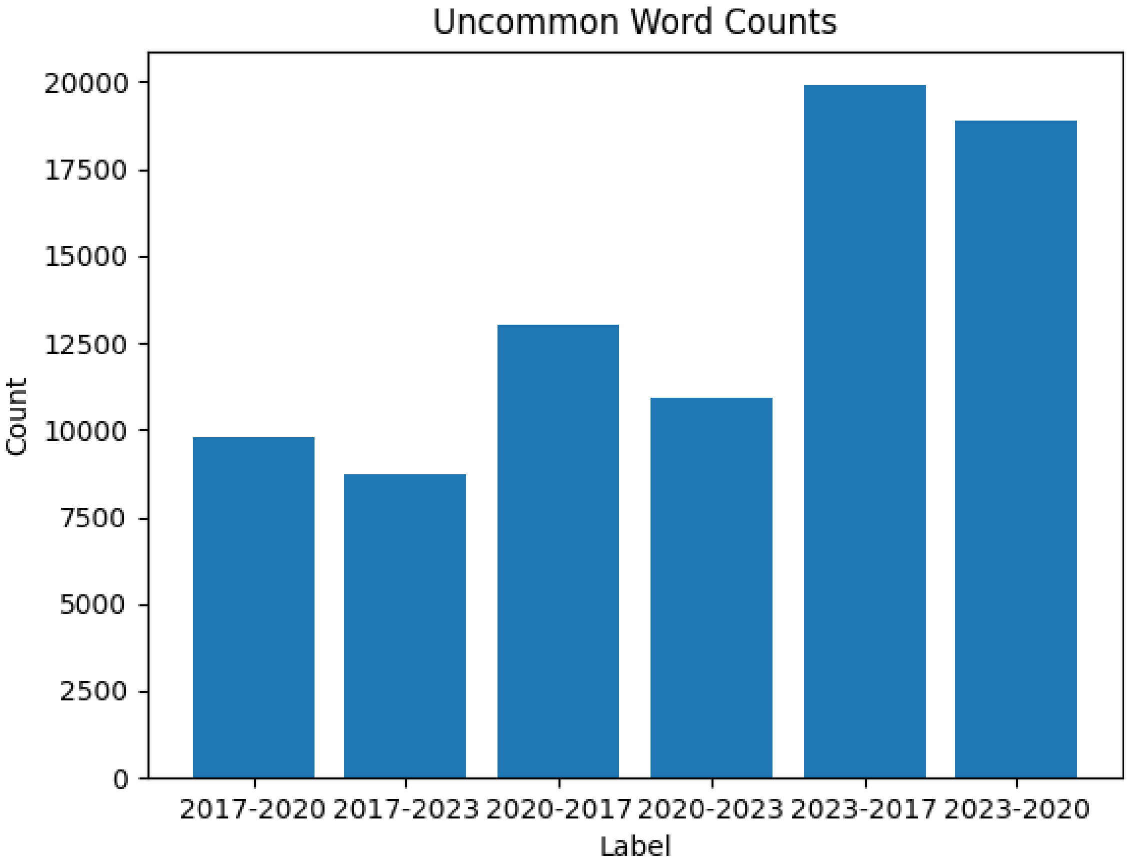
Word	Count
گفت	927
تا	907
بود	864
برای	687
ولی	660
بعد	649
ما	616
اون	586
گفتم	582
دیگه	573
خیلی	568



3.3 Common and Uncommon counts of words:

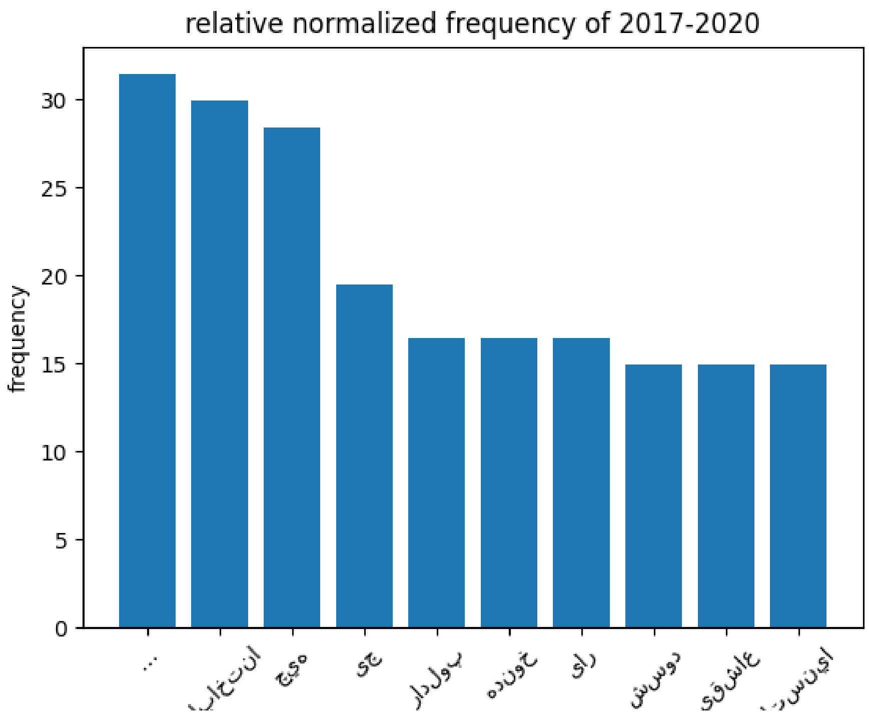
Label	common word	uncommon word
2017-2020	7439	9773
2017-2023	8500	8712
2020-2017	7439	13005
2020-2023	9522	10922
2023-2017	8500	19896
2023-2020	9522	18874


Common Word Counts
Count
0
2000
4000
6000
8000
2017-2020
2017-2023
2020-2017
2020-2023
2023-2017
2023-2020
Label

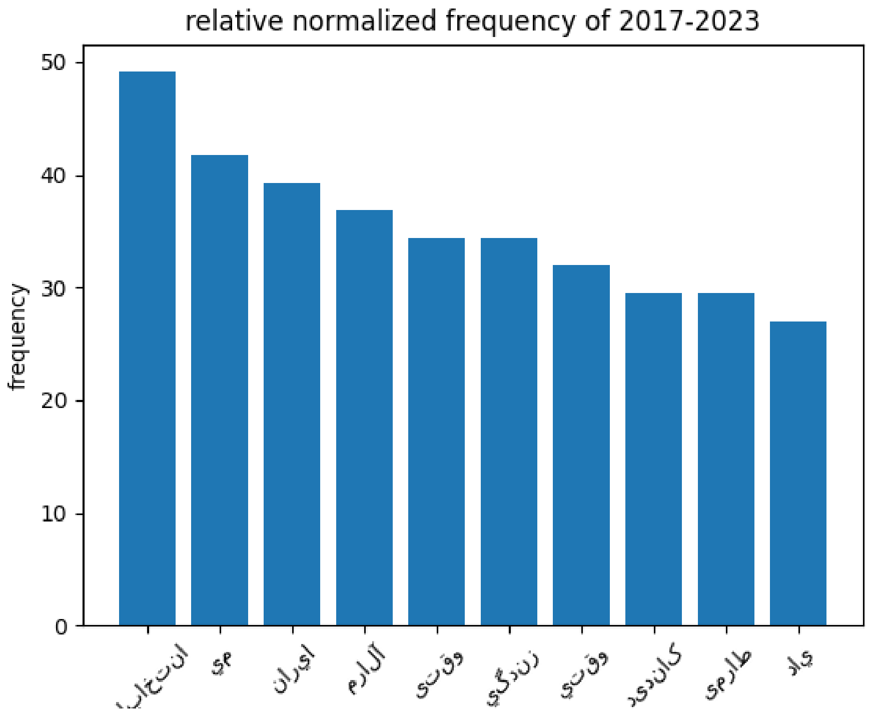

Uncommon Word Counts
Count
0
2500
5000
7500
10000
12500
15000
17500
20000
2017-2020
2017-2023
2020-2017
2020-2023
2023-2017
2023-2020
Label

3.4 Relative Normalized Frequency(RNF):

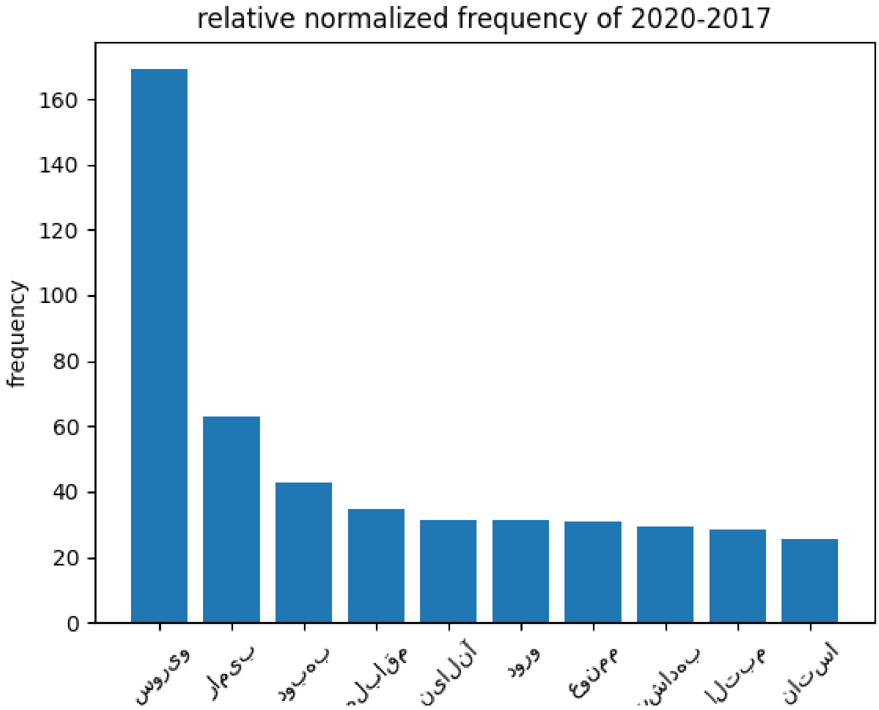
3.4.1 2017-2020 Common words by RNF



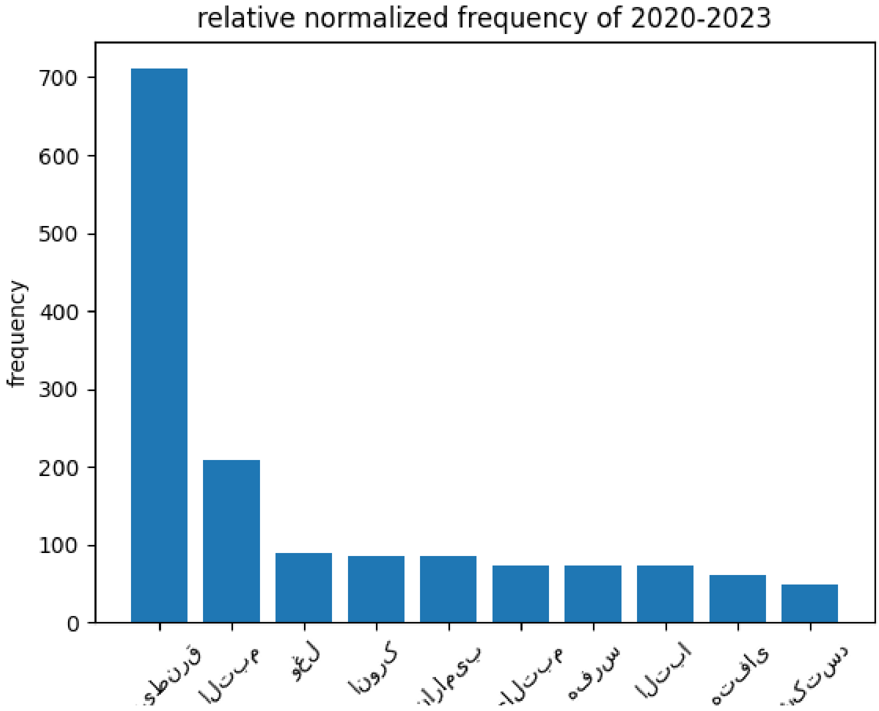
3.4.2 2017-2023 Common words by RNF



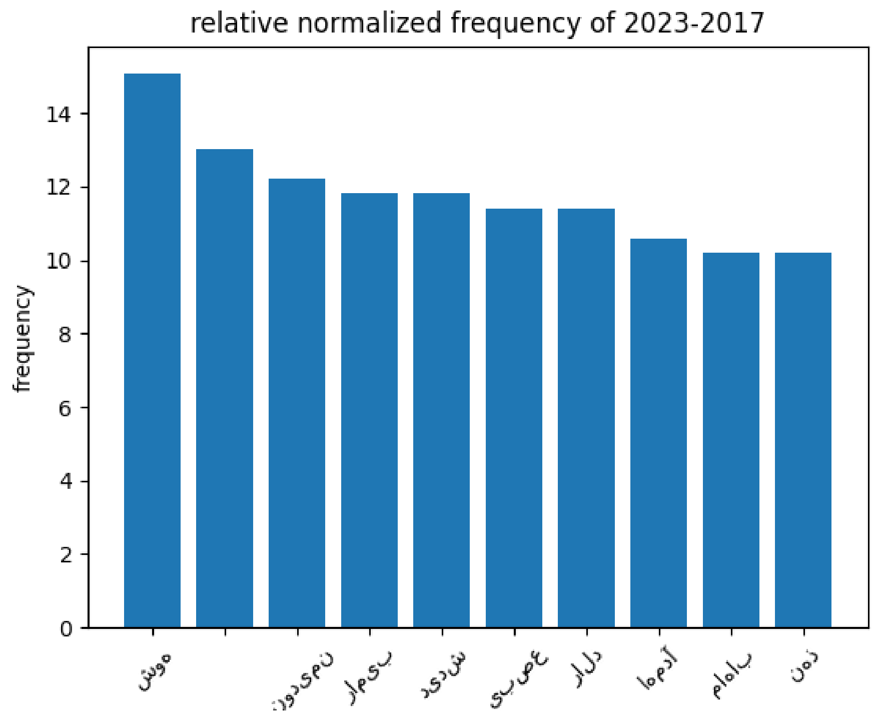
3.4.3 2020-2017 Common words by RNF



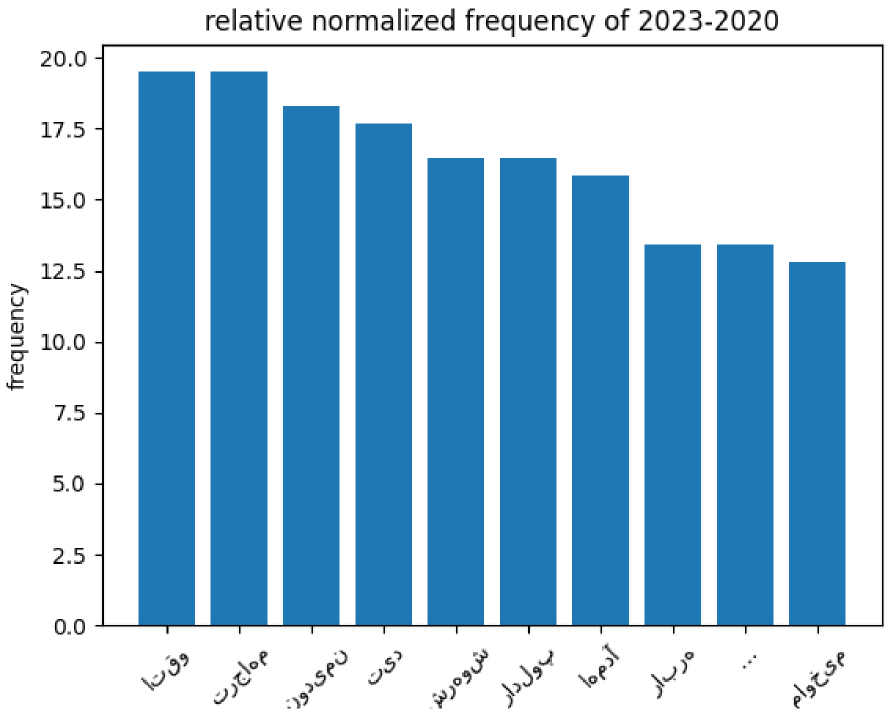
3.4.4 2020-2023 Common words by RNF



3.4.5 2023-2017 Common words by RNF

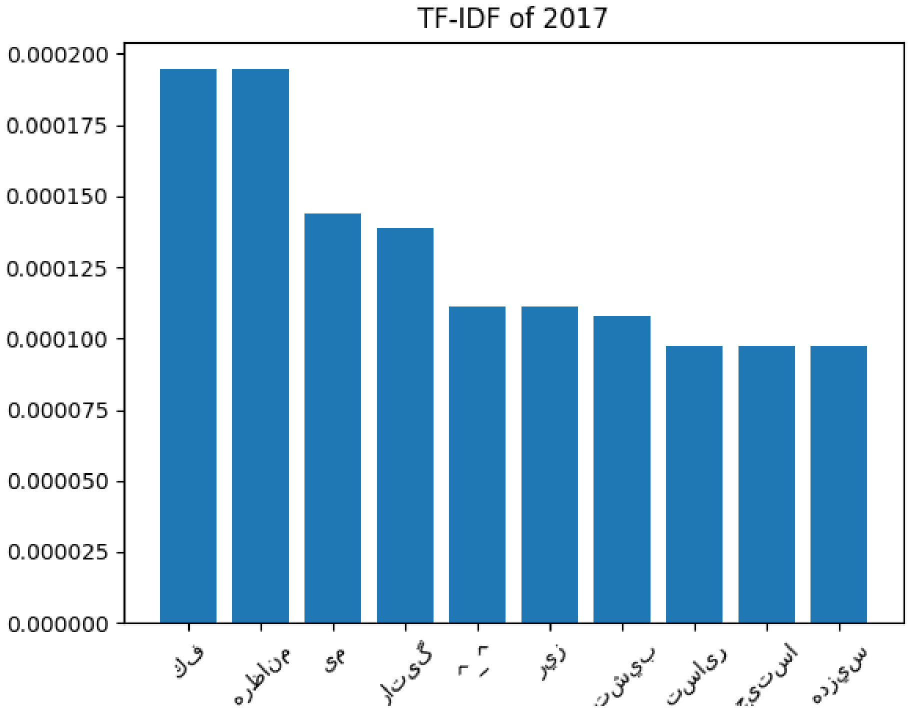


3.4.6 2023-2020 Common words by RNF

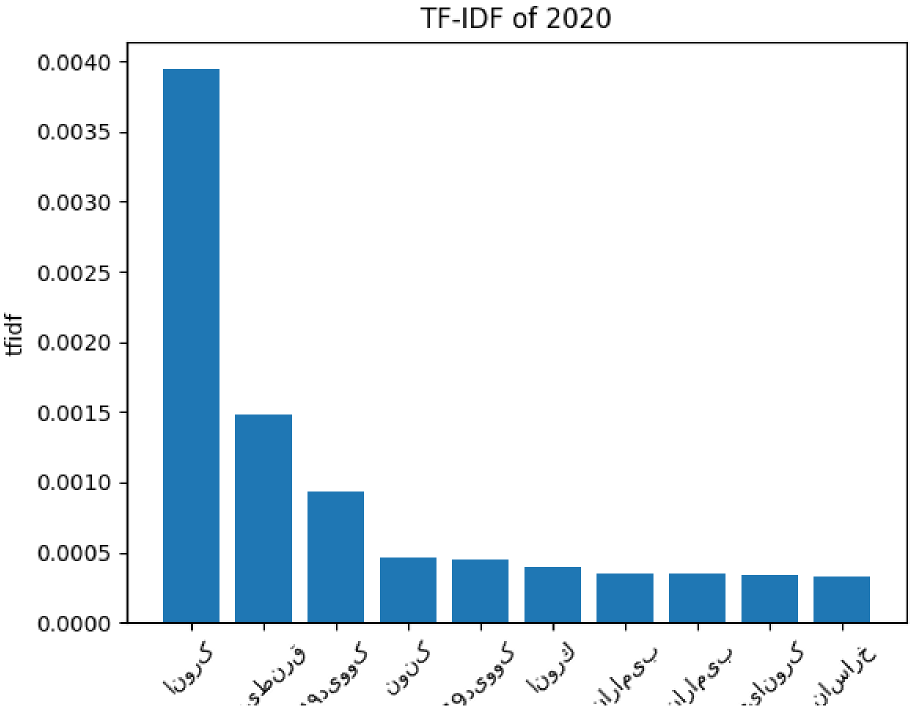


3.5 TF-IDF:

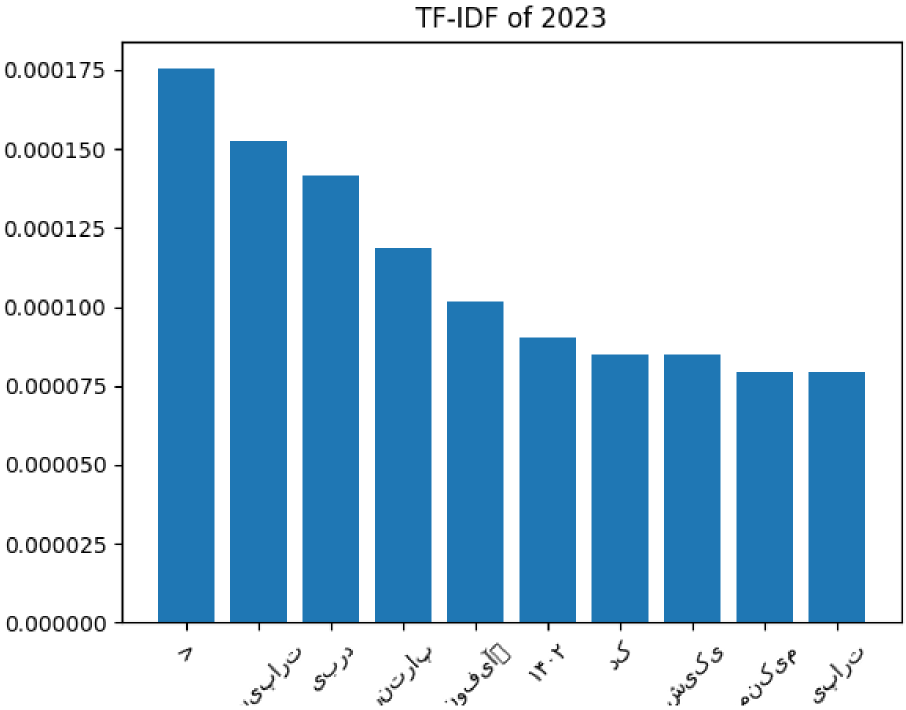
3.5.1 2017 tf-idf most Common words



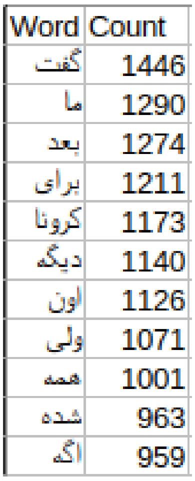
3.5.2 2020 tf-idf Common words



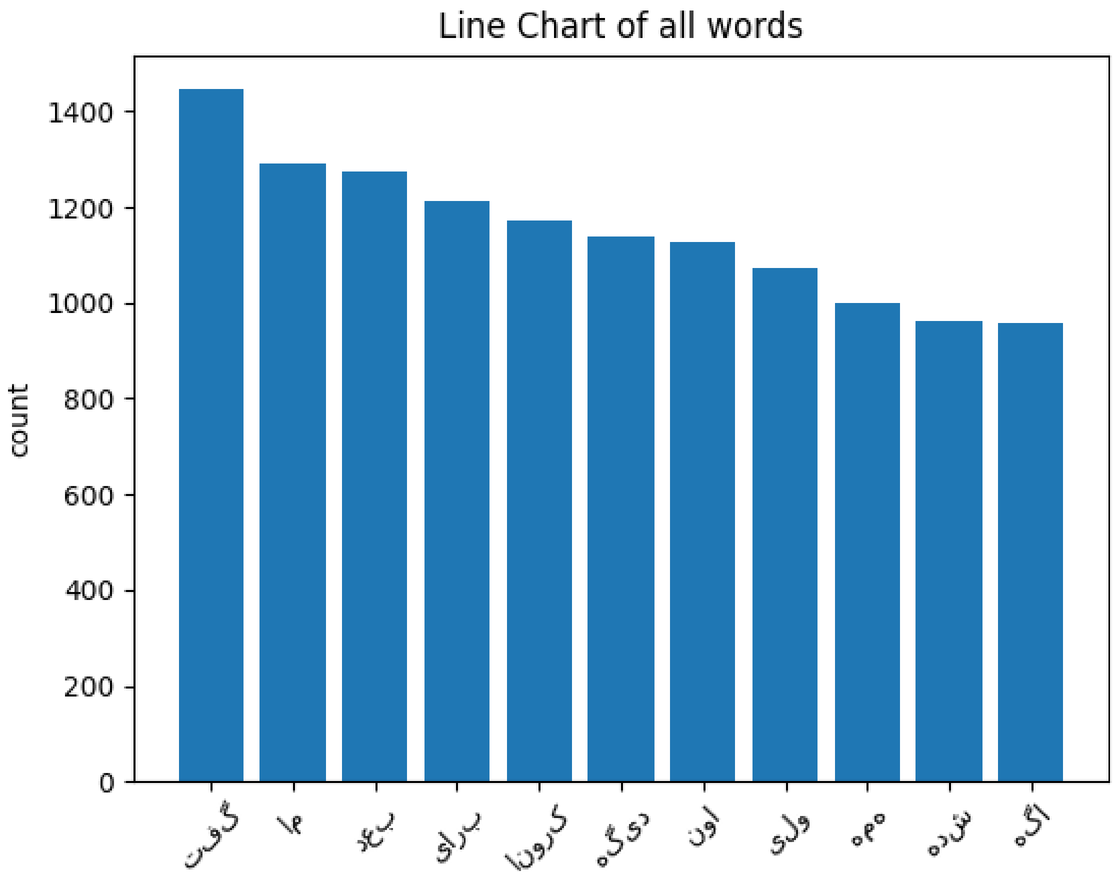
3.5.3 2023 tf-idf most Common words



3.6 histogram of all word:



Word	Count
گفت	1446
ما	1290
بعد	1274
برای	1211
کرونا	1173
دیگه	1140
اون	1126
ولی	1071
همه	1001
شده	963
اگه	959


Line Chart of all words
count
گفت
ما
بعد
برای
کرونا
دیگه
اون
ولی
همه
شده
اگه