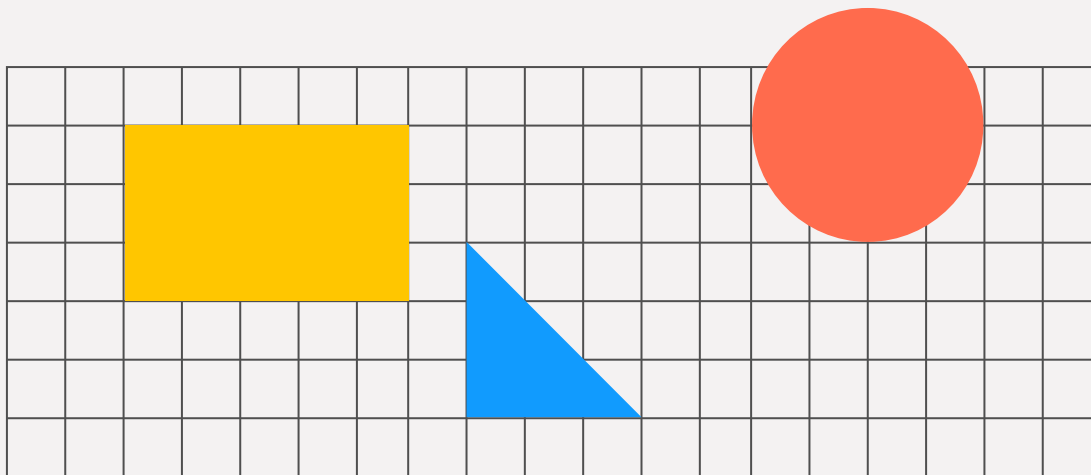
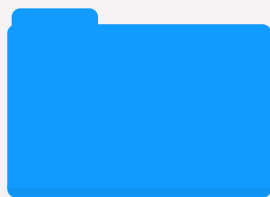


Course Final Report

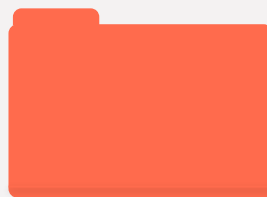




Dataset
Collection



Data
Exploration



Data
Exploration
Improvements







Additional
Contributions



Dataset
Collection

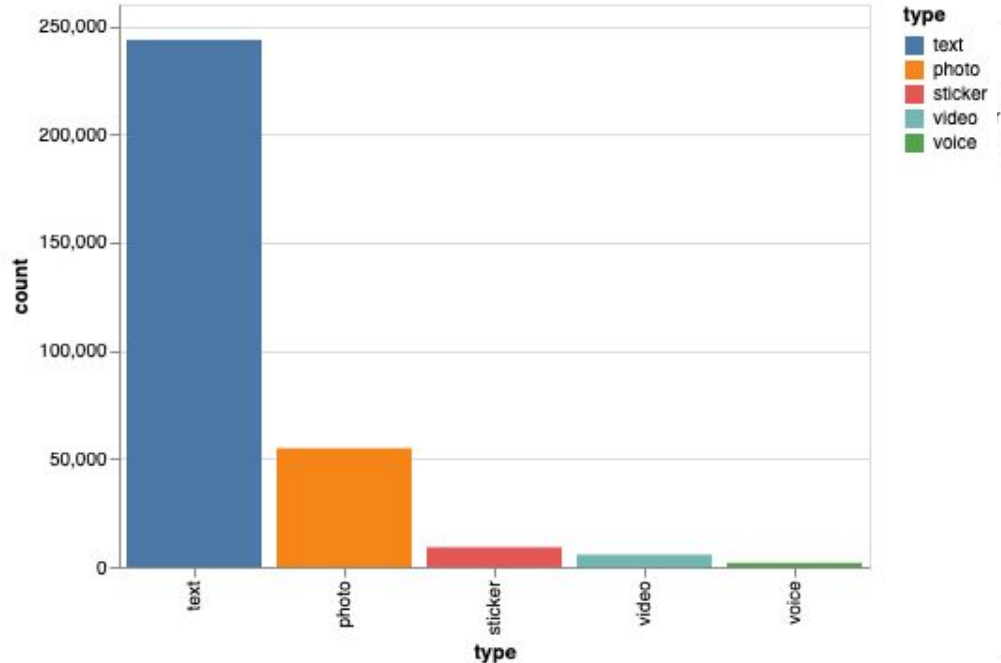
Data Collection

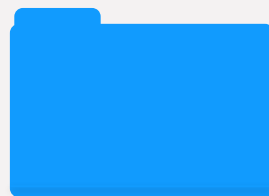
- Approach:
 - Telegram data were collected using Andrew's telegram collection tool. 
 - We also downloaded V1 of propaganda dataset by Kate Burovova.
- Time Spent: Around 4h 
- Problems:
 - Telegram API Keys 
 - Data Collection Takes Forever 

Data Collection (Statistics)

We collected 315271 messages of which:

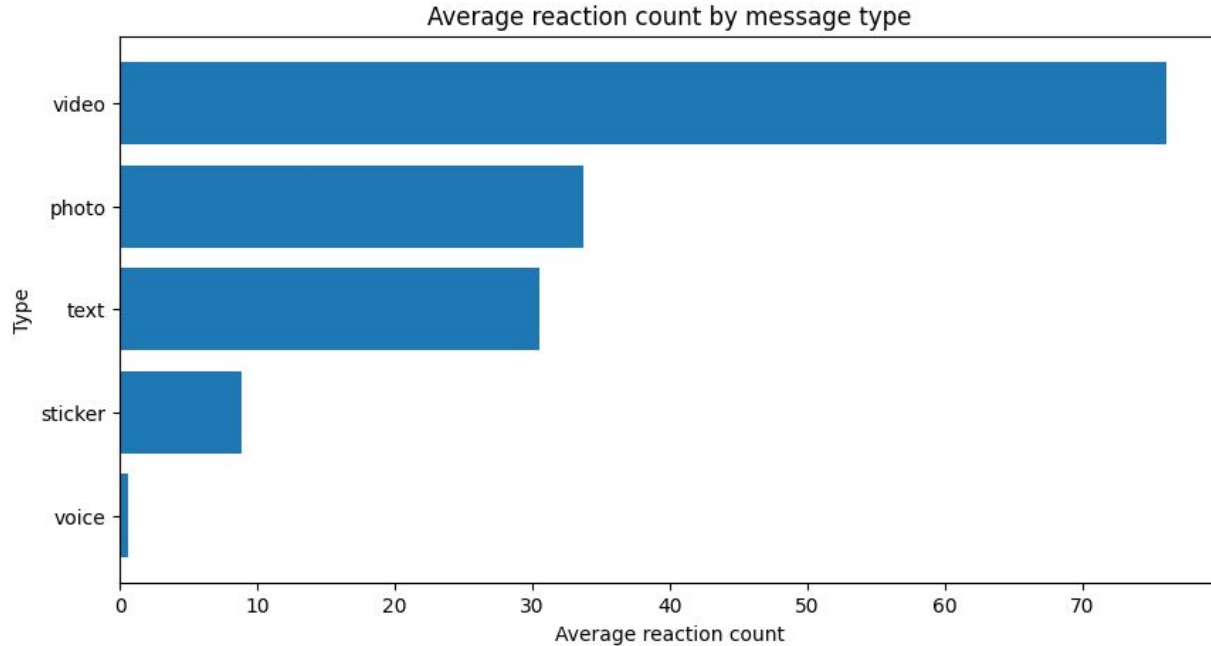
- 243603 - text 📝
- 54889 - photo 📷
- 9152 - stickers 🏷️
- 5716 - video 🎬
- 1911 - voice 🗣️



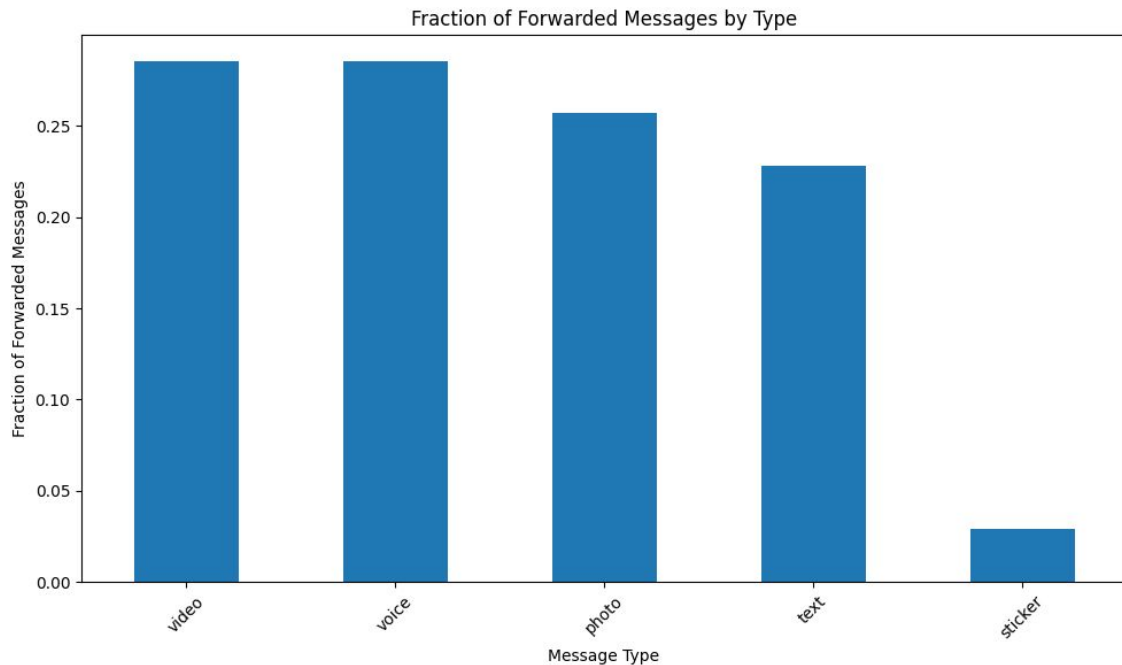


Data
Exploration

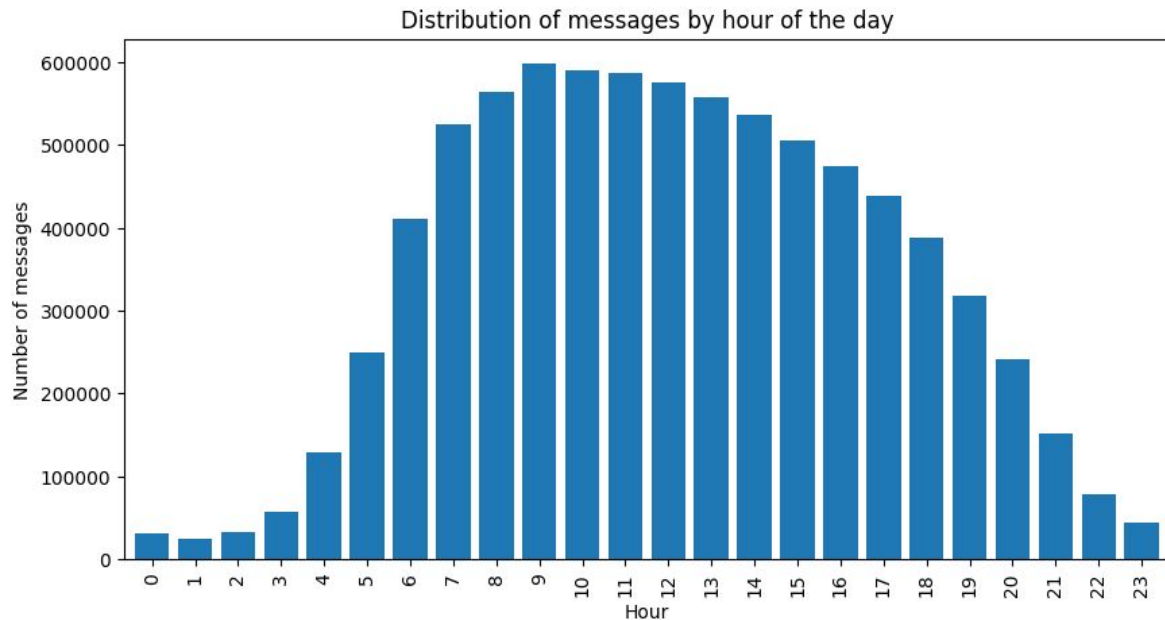
Data Exploration



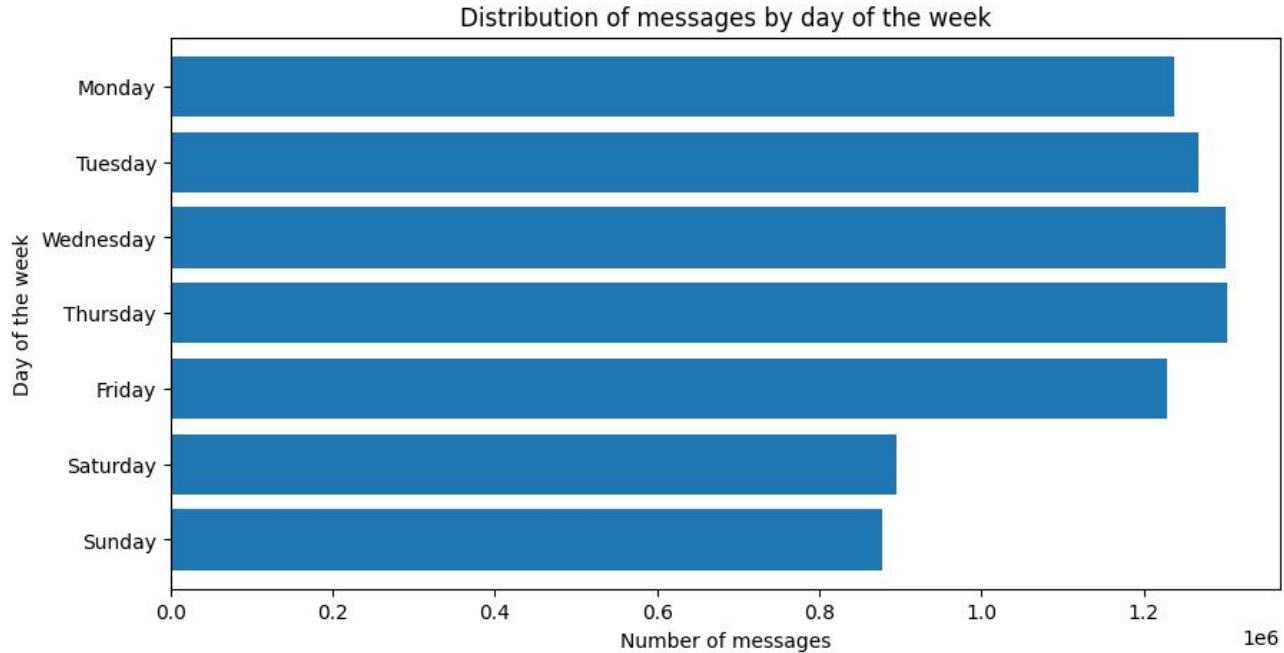
Data Exploration



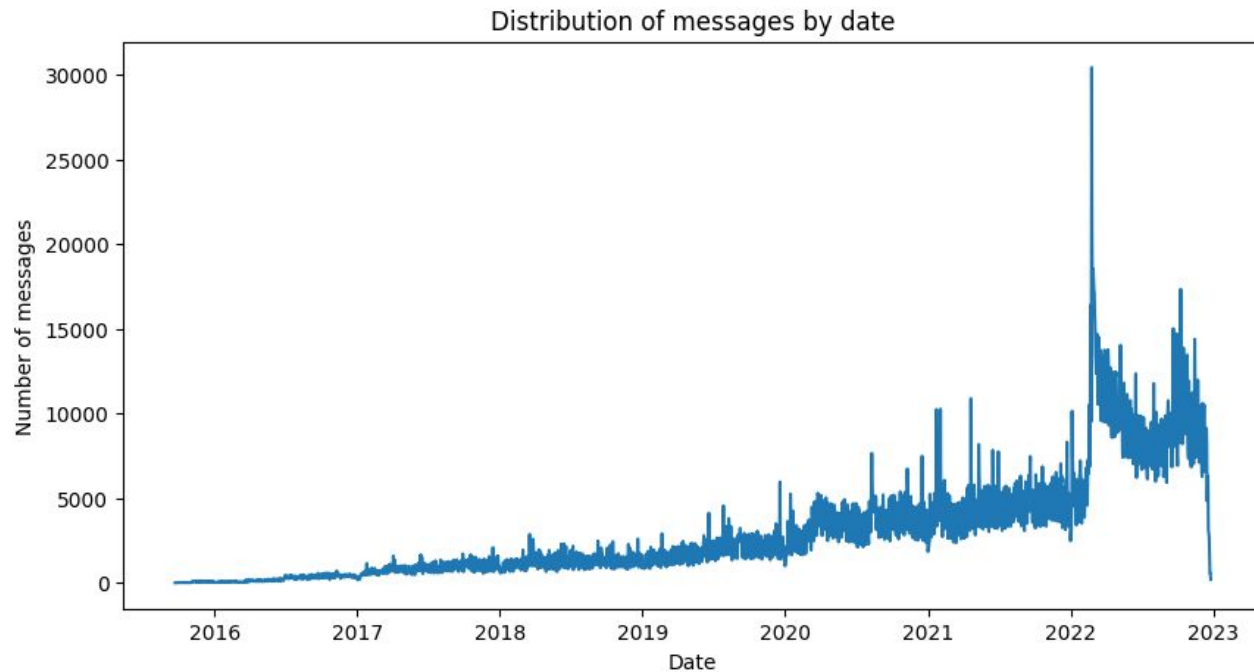
Data Exploration



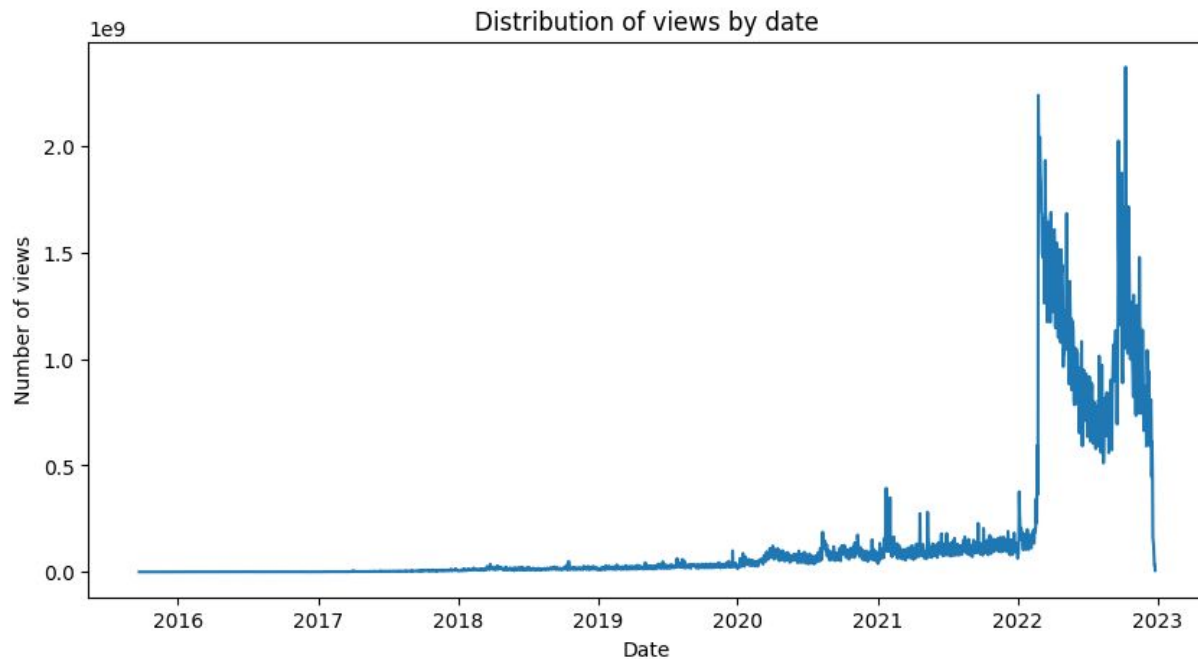
Data Exploration



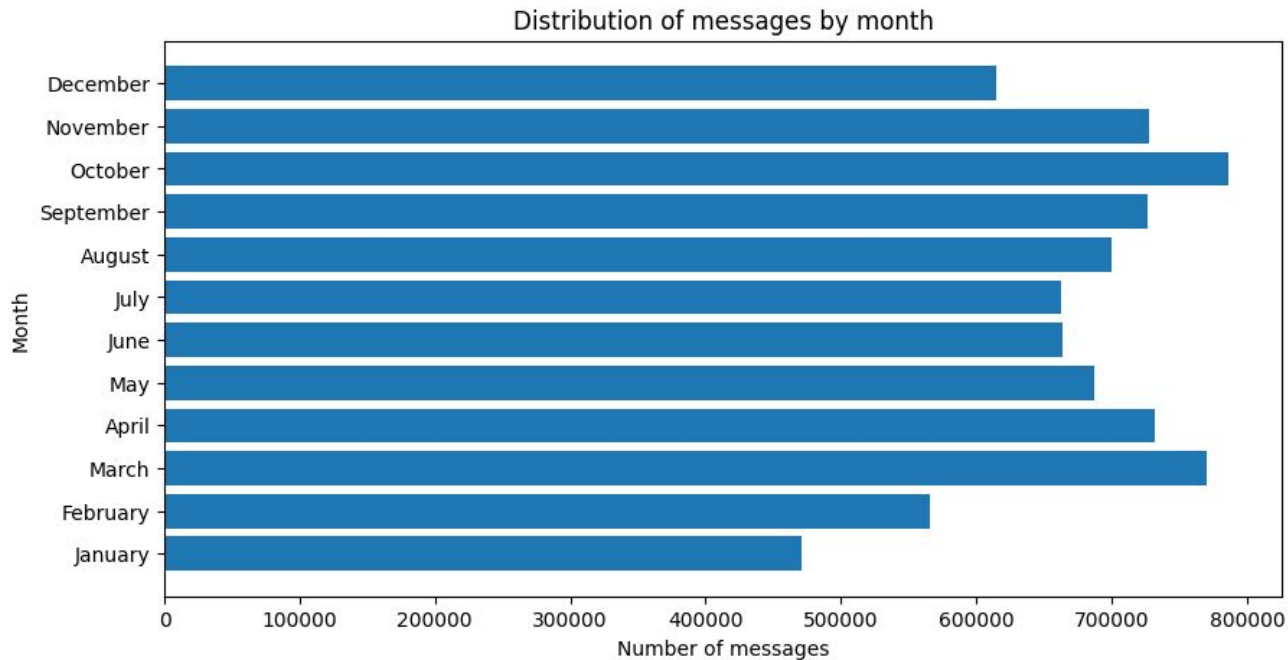
Data Exploration



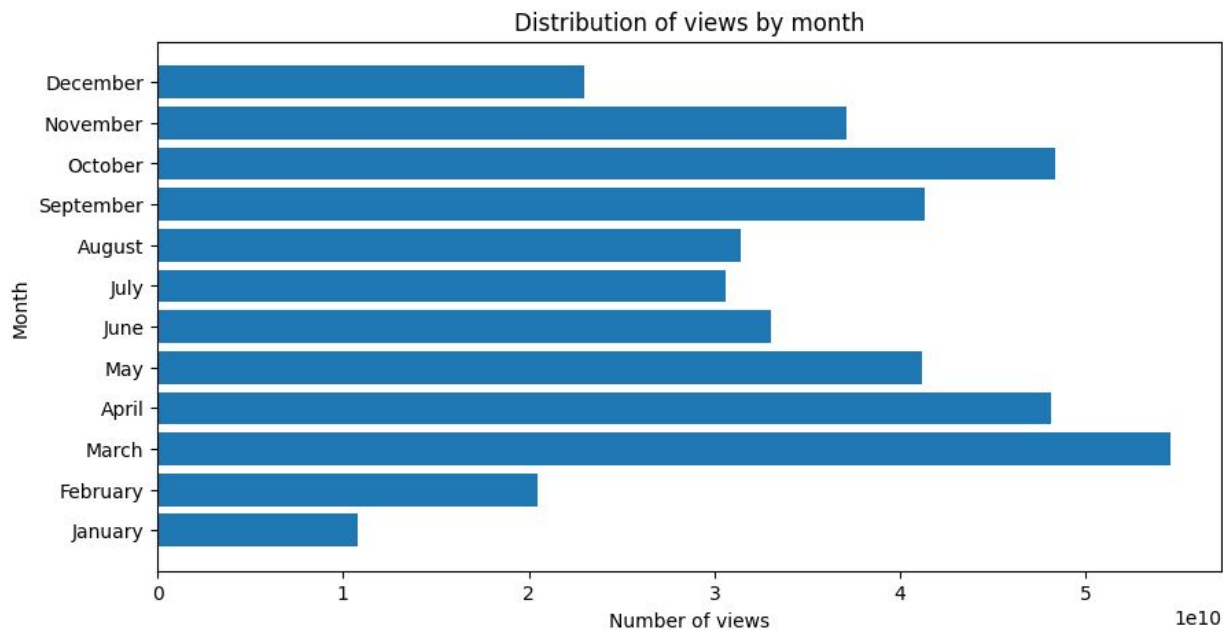
Data Exploration



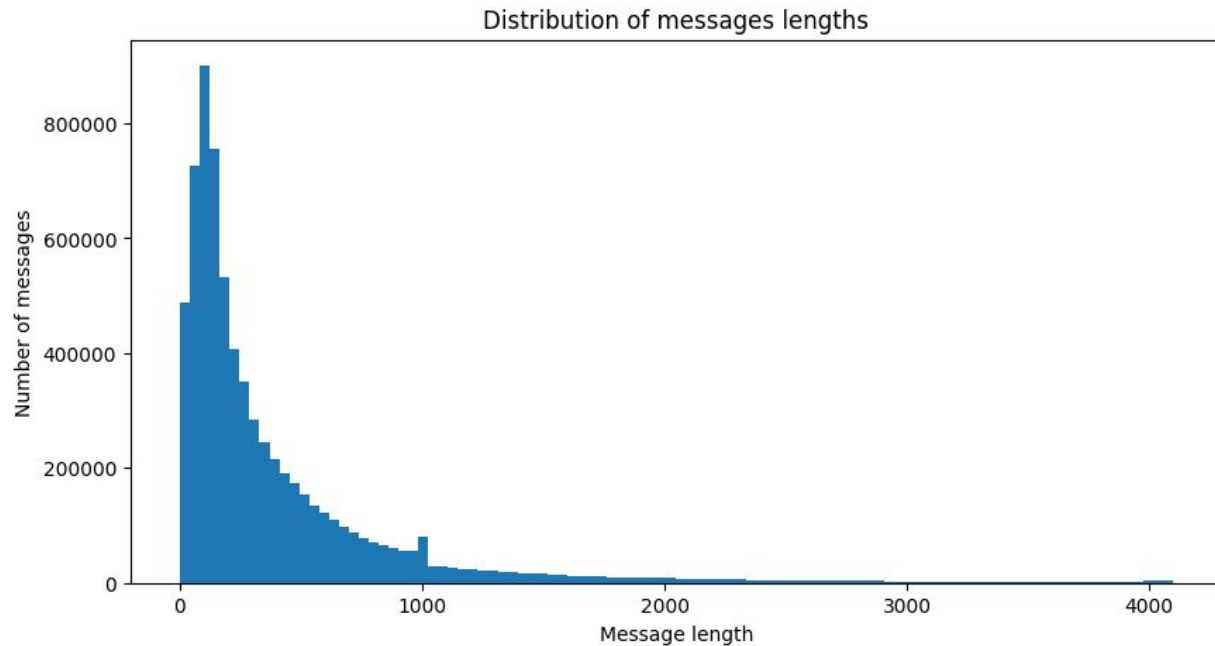
Data Exploration



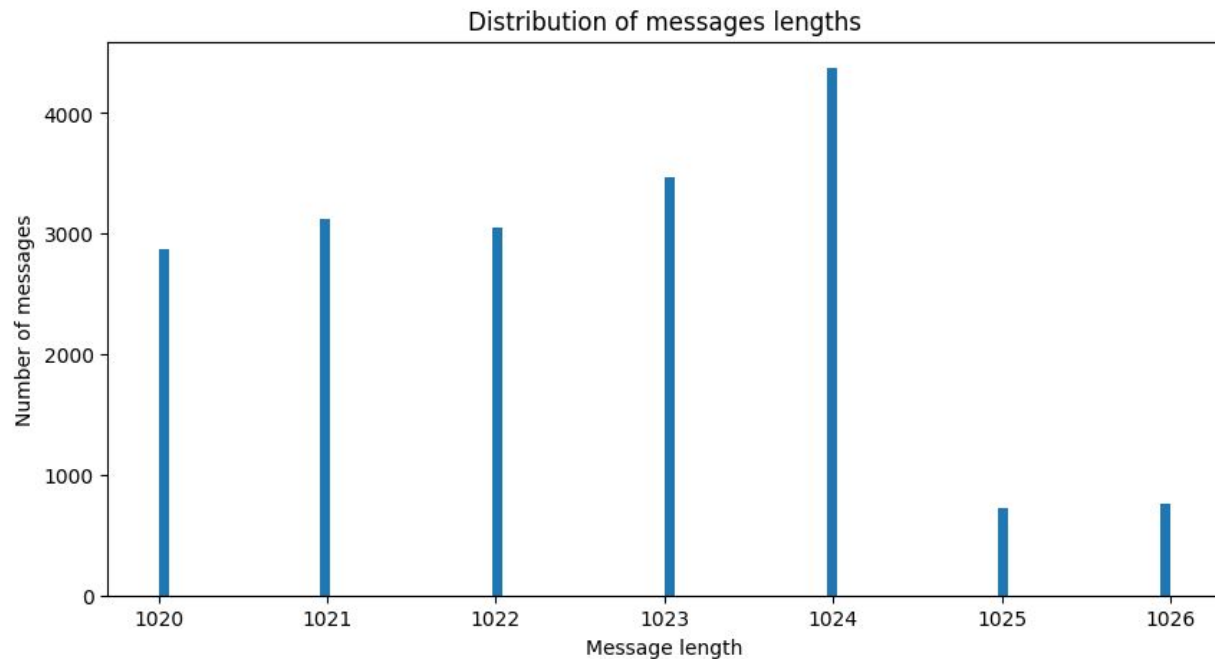
Data Exploration



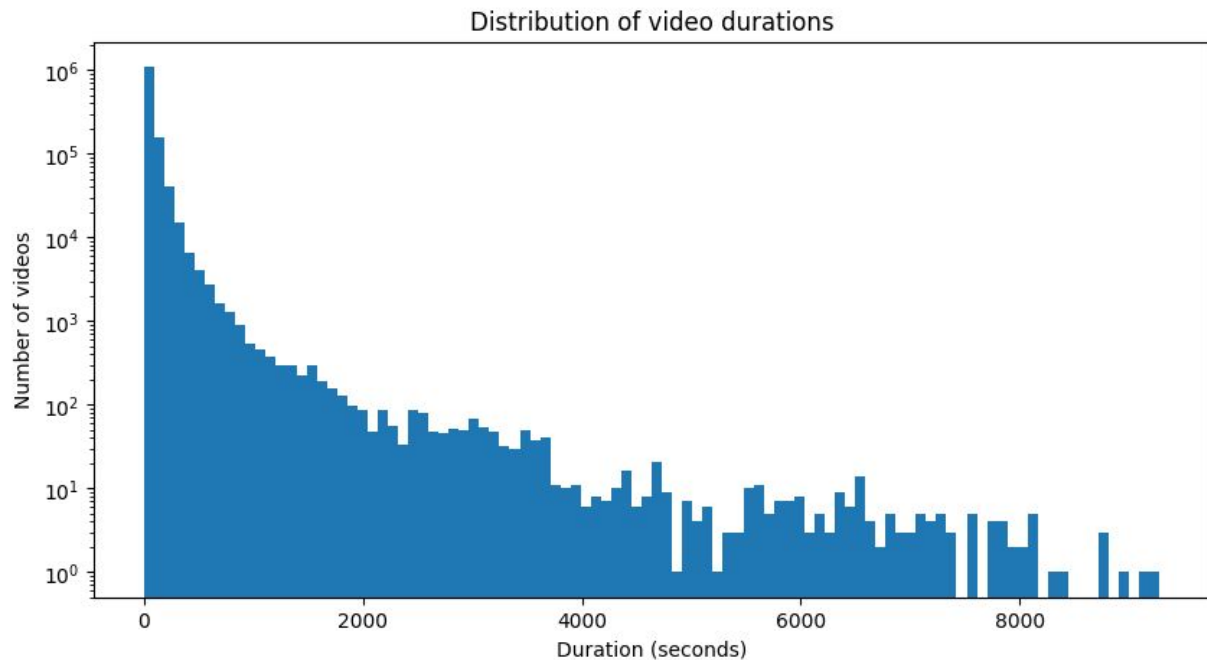
Data Exploration



Data Exploration



Data Exploration



Data Exploration

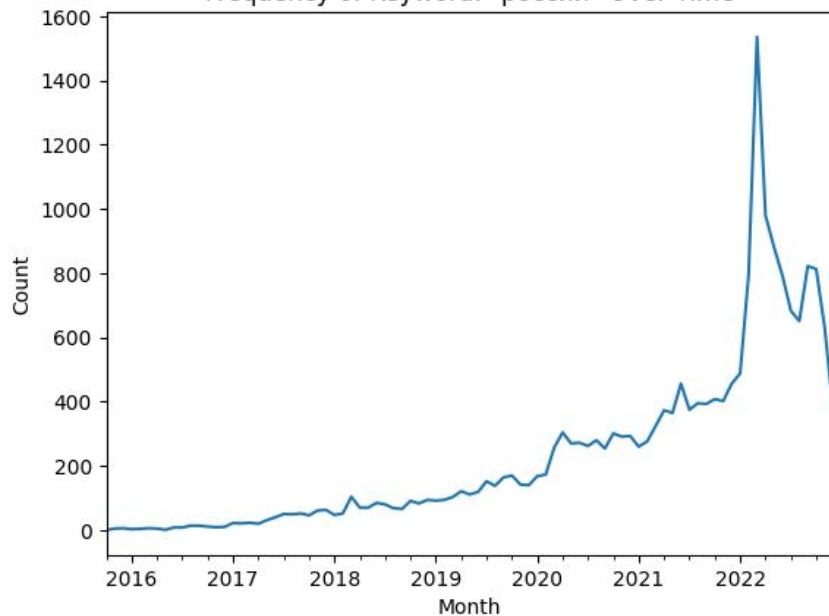
- Top 10 most common words:

- это: 20523
- россии: 13023
- рф: 6284
- года: 6135
- сша: 6114
- которые: 5542
- также: 5506
- время: 4925
- который: 4568
- области: 4135
- украины: 4133
- лет: 4068
- человек: 4031
- сегодня: 3972
- заявил: 3822
- очень: 3723
- году: 3698
- россия: 3567
- будут: 3509
- изза: 3442

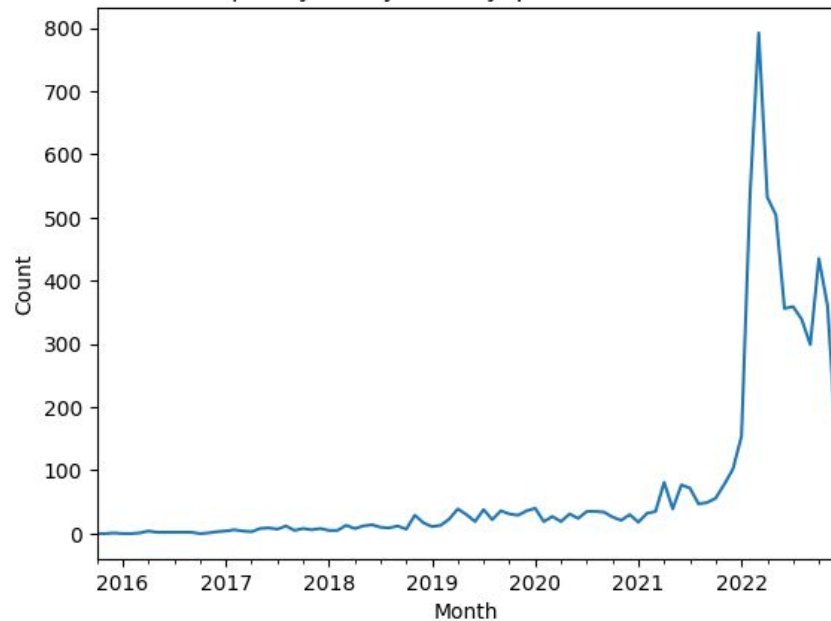
Data Exploration



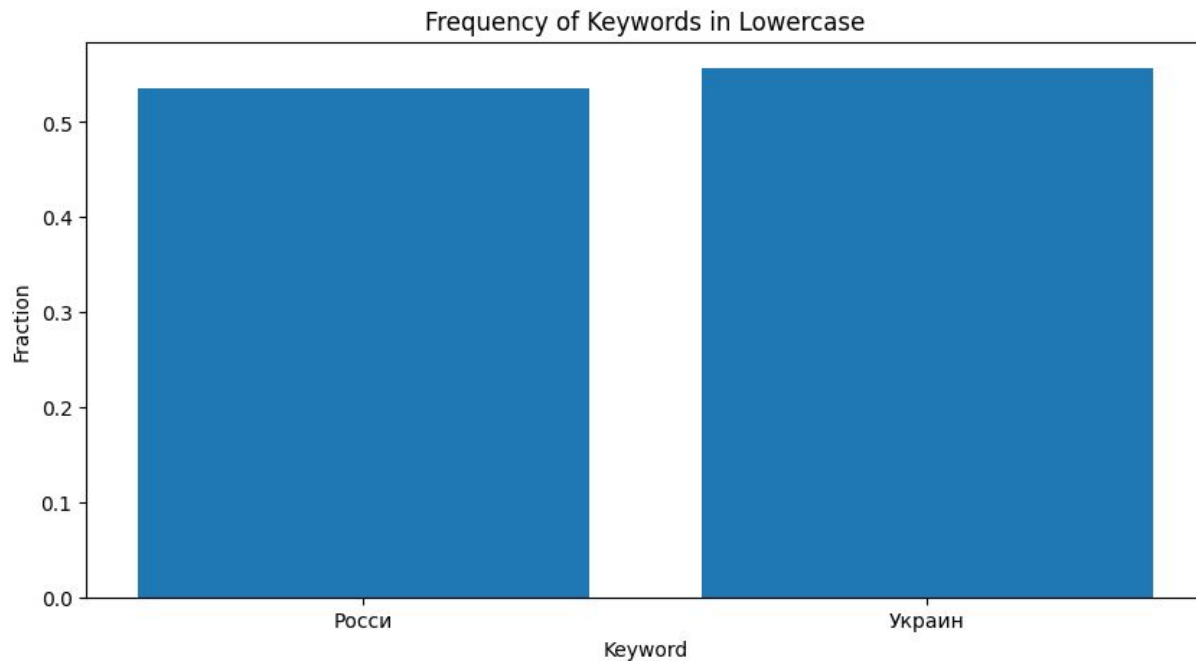
Frequency of Keyword: "россии" Over Time



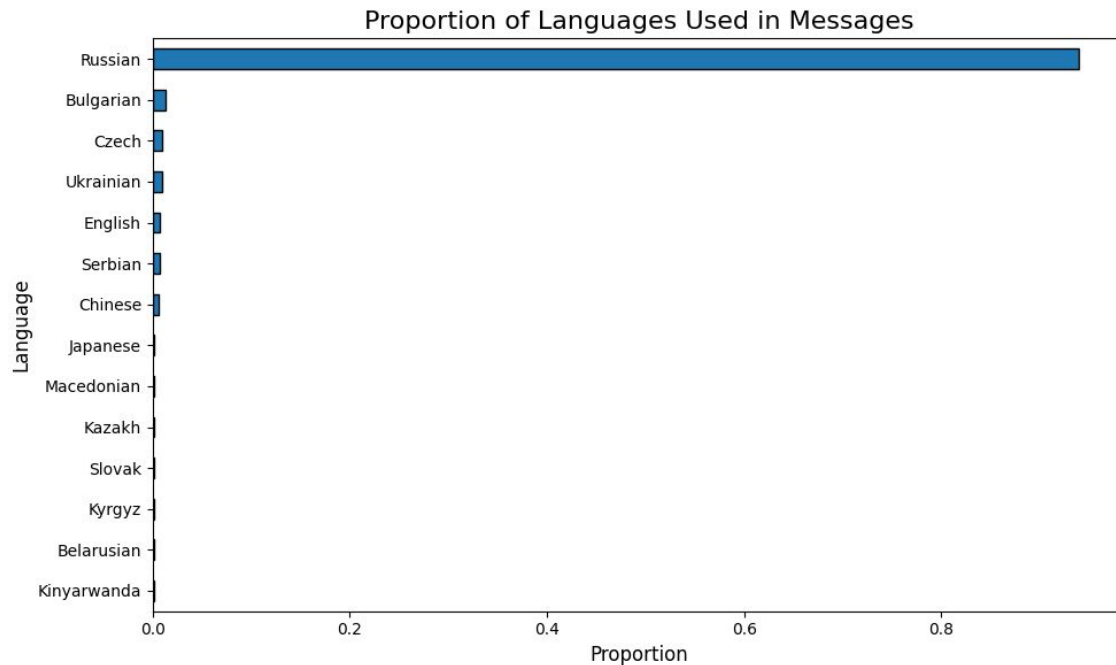
Frequency of Keyword: "украины" Over Time

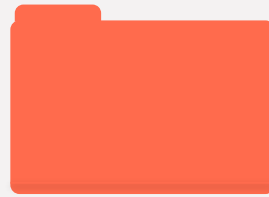


Data Exploration



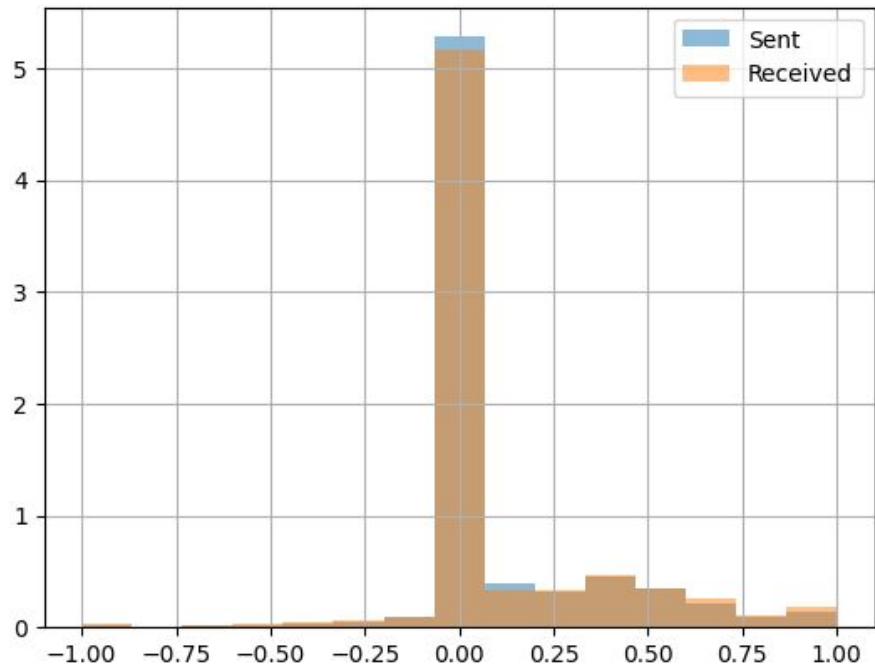
Data Exploration



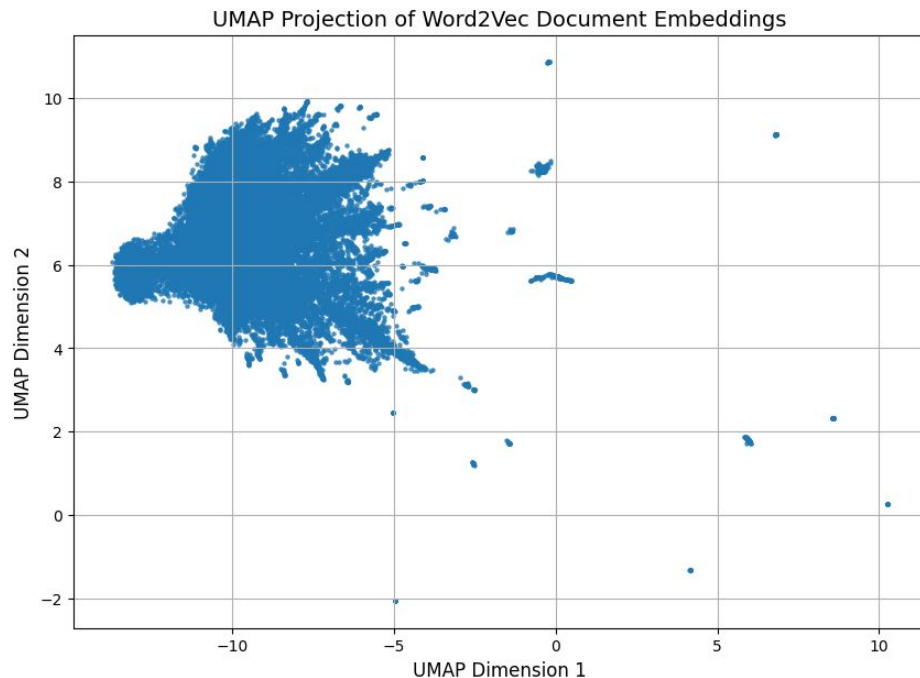


Data
Exploration
Improvements

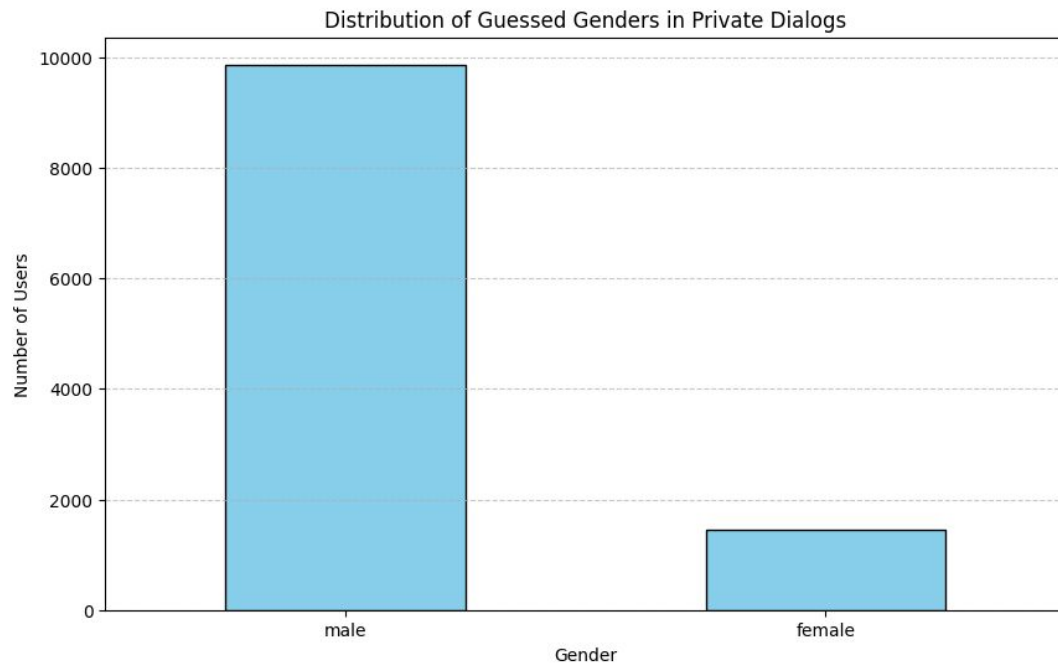
Data Exploration (+)



Data Exploration (+)



Data Exploration (+)



Data Exploration (+)

Accuracy: 0.5622837370242214

Classification Report:

	precision	recall	f1-score	support
female	0.57	0.49	0.53	289
male	0.55	0.63	0.59	289
accuracy			0.56	578
macro avg	0.56	0.56	0.56	578
weighted avg	0.56	0.56	0.56	578



Additional
Contributions

Additional Contributions

The screenshot shows the GitHub repository page for 'SanGree / telegram-data-collection'. The repository is public and has 2 branches, 0 tags, and 126 commits. The repository is currently on the 'master' branch. The repository is a collection of tools to download Telegram data. The repository is currently on the 'master' branch. The repository is a collection of tools to download Telegram data.

Repository: telegram-data-collection (Public)

Watch 2 Fork 9 Star 3

master 2 Branches 0 Tags

Go to file Add file Code

SanGree Merge pull request #21 from Roman4466/master c9932bd · last week 126 Commits

File	Commit Message	Time
config	moved data folder in the parent folder	4 years ago
utils	run black to reformat code	last month
.gitignore	fix code	2 years ago
0_download_dialogs_list.py	run black to reformat code	last month
1_download_dialogs_data.py	refactored	last month
README.md	feat: Added --skip_progress, --all_at_once flags to script ...	last month
requirements.txt	refactoring	2 years ago

README

Telegram data collector v0.01

Combination of tools to download your telegram data.

About

No description, website, or topics provided.

Readme

Activity

3 stars

2 watching

9 forks

Report repository

Releases

No releases published






Packages

No packages published

Contributors 8

8 contributors

Additional Contributions

-  Code Refactoring
-  Logging
-  Asynchronous Execution
-  Status Bars
-  Execution Speedup
(not sure, need to be confirmed by measurements)

Additional Contributions

```

~/De/telegram-data-collection-v2 initial_contrib > python download_dialogs_data.py --concurrency 10
# Downloading all dialogs data... - 2% 0:00:16 0:12:58
# Processing: Markiian Mandzak 0% 0:00:16 -:--:--
# Processing: Data Science 2023 0% 0:00:16 -:--:--
  Downloading: MyUCU 100% 0:00:13 0:00:00
  Downloading: MyUA 100% 0:00:13 0:00:00
  Completed: MyCertificates 100% 0:00:13 0:00:00
# Processing: Артур 0% 0:00:16 -:--:--
# Processing: Едік 0% 0:00:16 -:--:--
# Processing: GlobalLogic Ukraine 0% 0:00:16 -:--:--
# Processing: Український Фонд Стартапів 0% 0:00:16 -:--:--
# Processing: Тарас 0% 0:00:16 -:--:--
# Processing: Computer Vision 0% 0:00:00 -:--:--

```

Q/A



GitHub Course Repository

I have too many
answers, so ask me your
✨ Question ✨



GitHub Fork of Data Collection