Final report for telegram data analysis

Computational Social Science

Student: Valentyna Dermenzhy
Mentor: Andrew Kurochkin

Repository: telegram-dialogs-analysis-v2

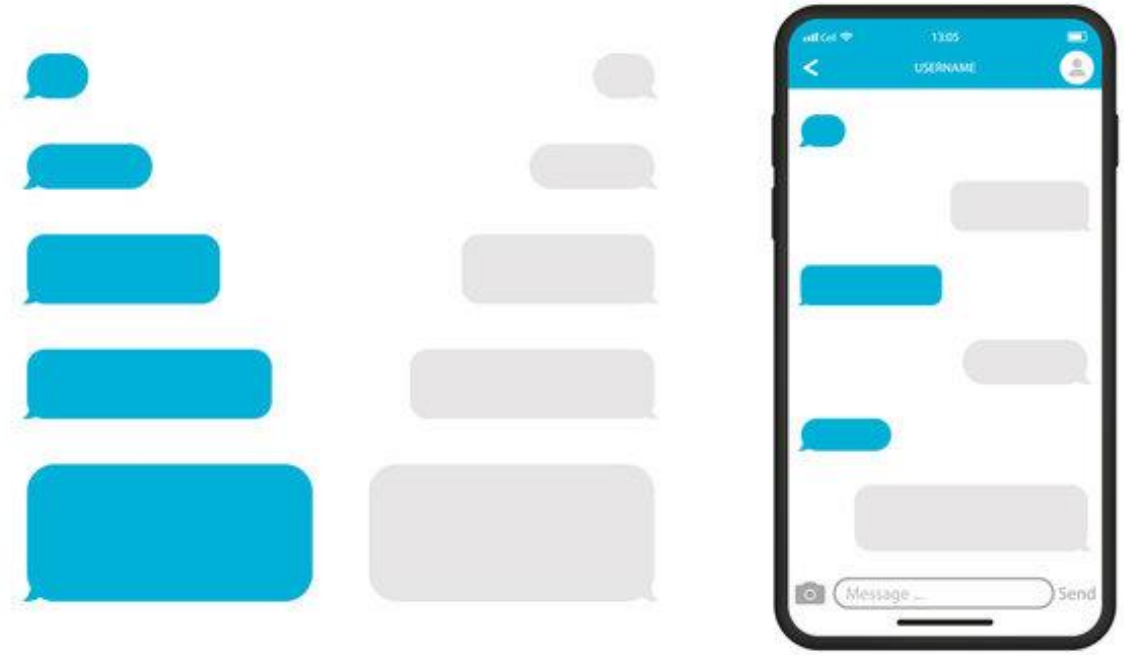
29.11.2024

Plan

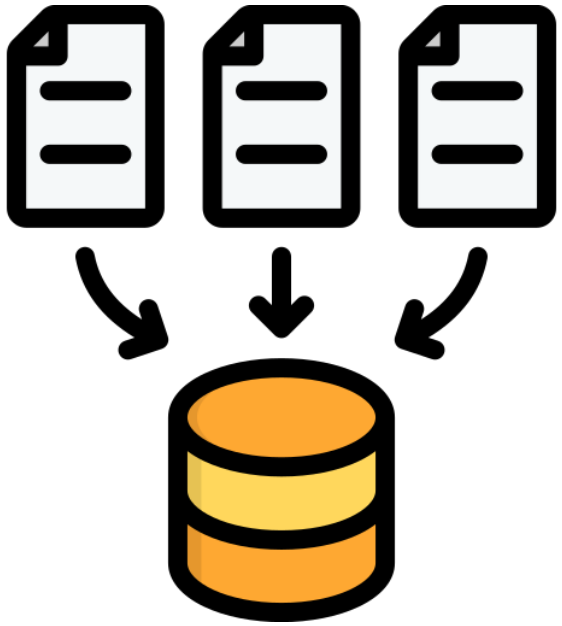
1. Introduction
2. Data collection
3. Exploratory Data Analysis
   - Activity analysis
   - Text and context analysis
   - Sentimental anylisis
4. Results
5. Further work

Introduction

Exploring personal behavioral patterns through Telegram activity, combining data analysis and social science to uncover meaningful insights.



Data Collection



Data statistics:

- dialog_data_all.csv (data about texts) — 300MB
  dialogs_users_all.csv (data about dialog users) — 1.6MB
- Sent messages: 209 464
  Received messages: 1 484 577

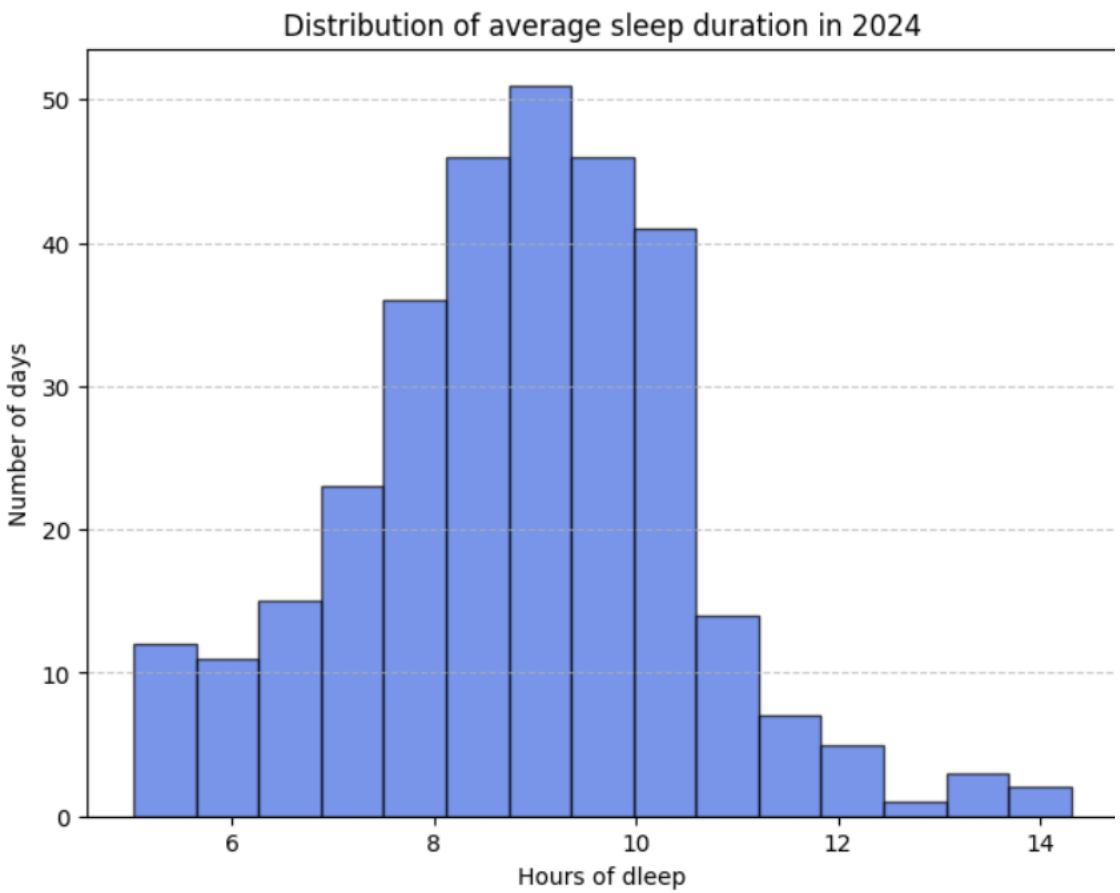
Total time spent: 1600 min

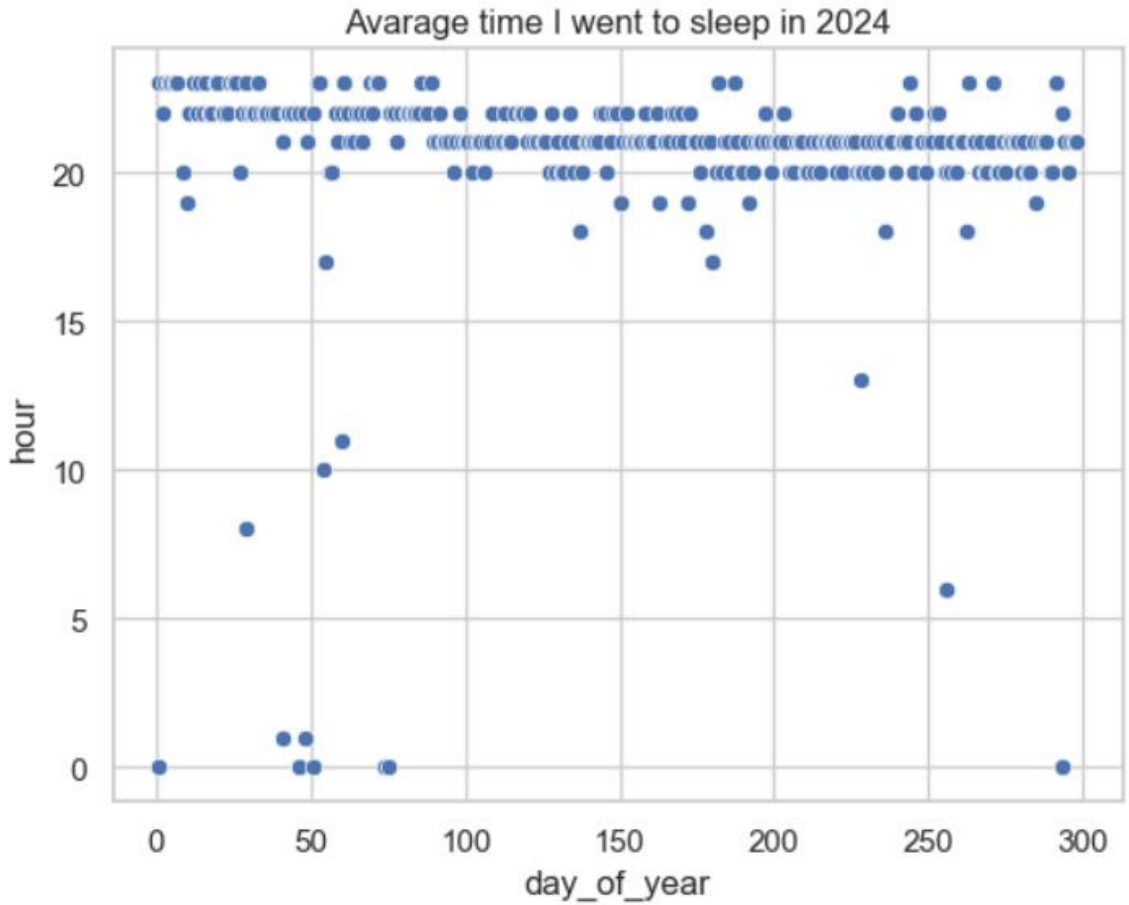
Challenges:

- Dependency Compatibility Issues
- Storage Limitations

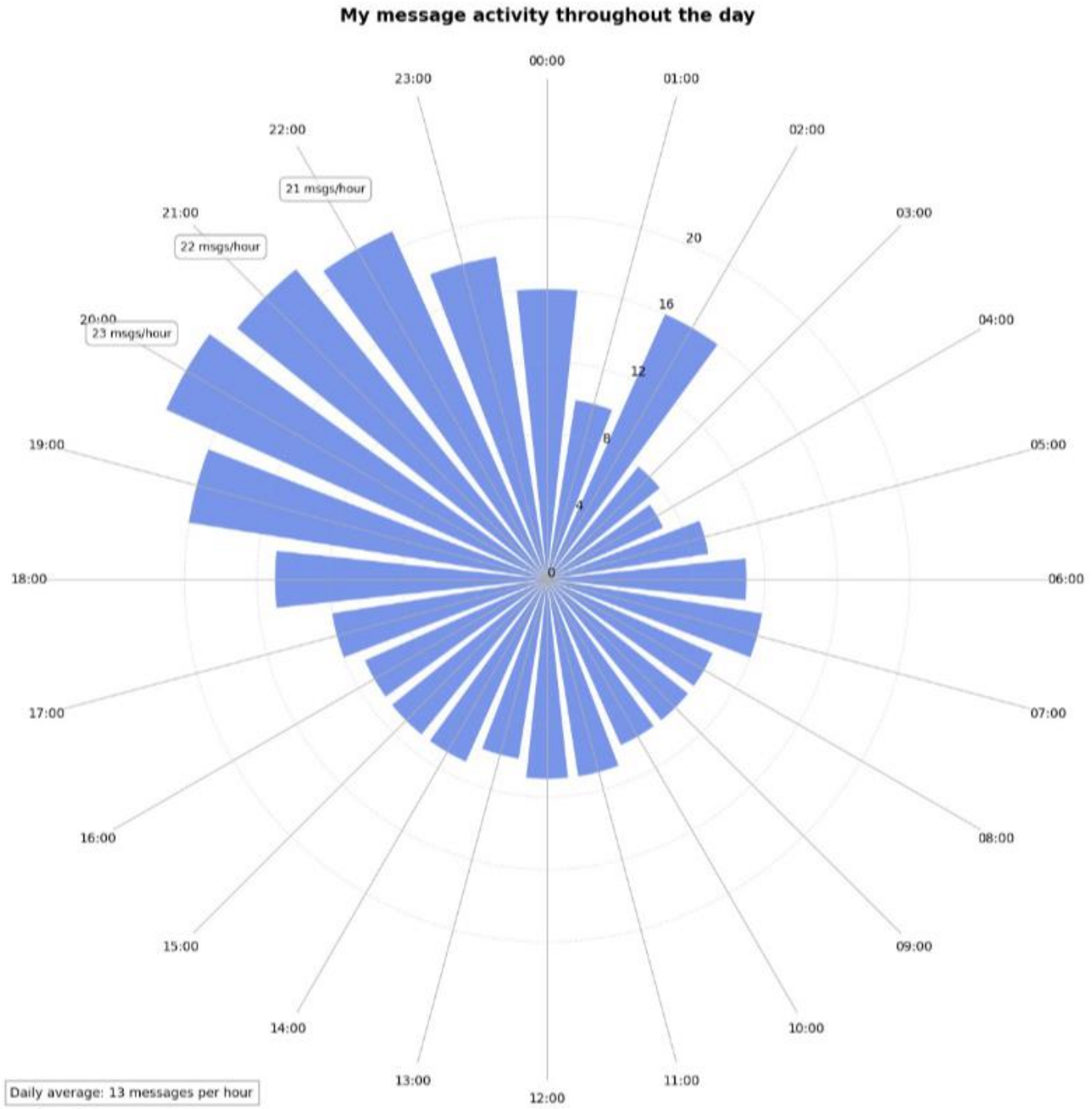
Exploratory Data Analysis

Sleep patterns





The distribution shows that the average sleep duration in 2024 typically ranges between 8 and 10 hours. Sleep times throughout the year show a generally consistent pattern with occasional deviations.

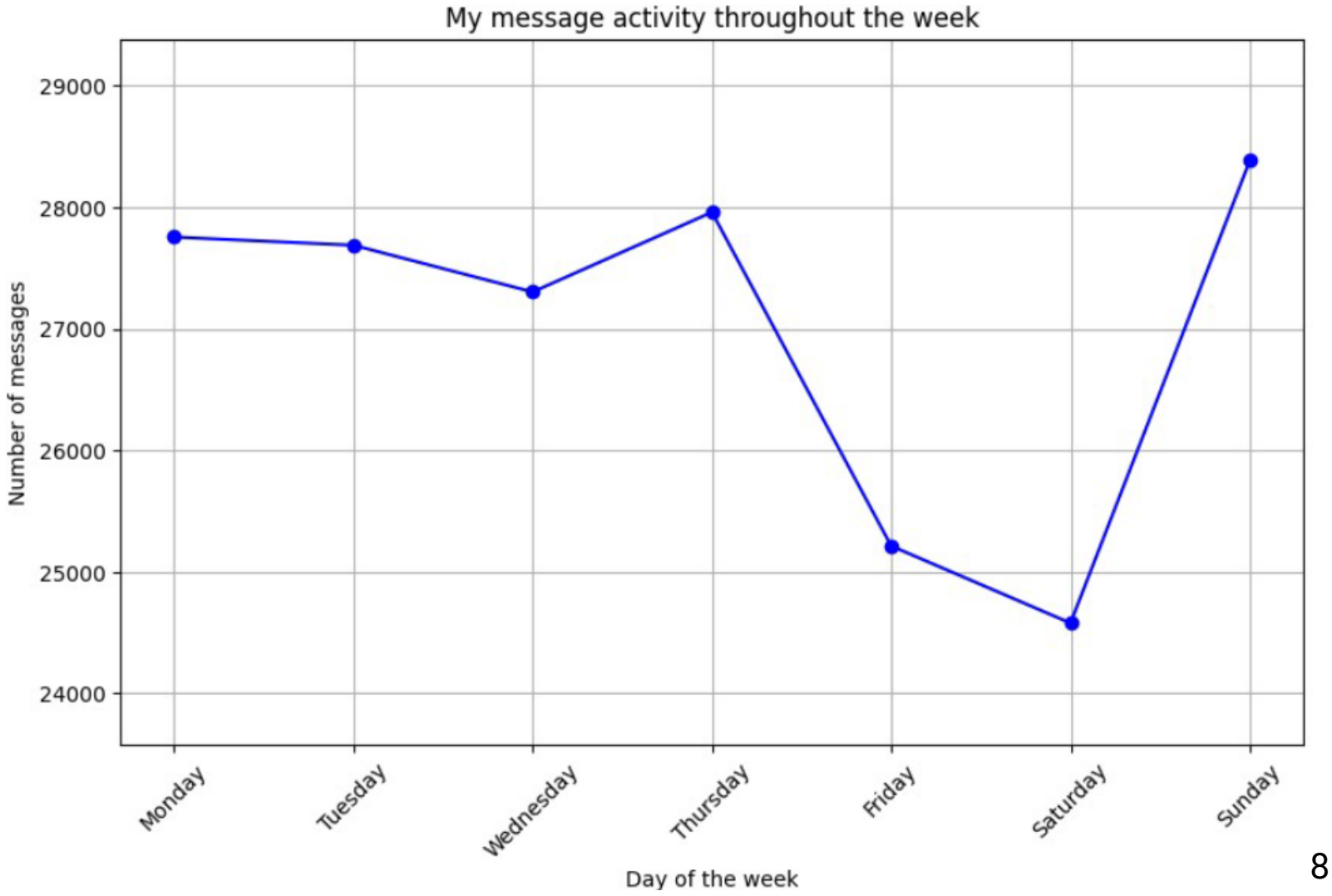


Activity patterns

- The analysis reveals distinct peaks in messaging activity during **8 – 10PM**, which align with typical social and leisure times for my age group. This trend indicates that these hours are the most engaging for interactions, suggesting that communication is most effective during these windows.
- Such patterns can provide valuable insights for timing important messages or posts to maximize visibility and response rates.

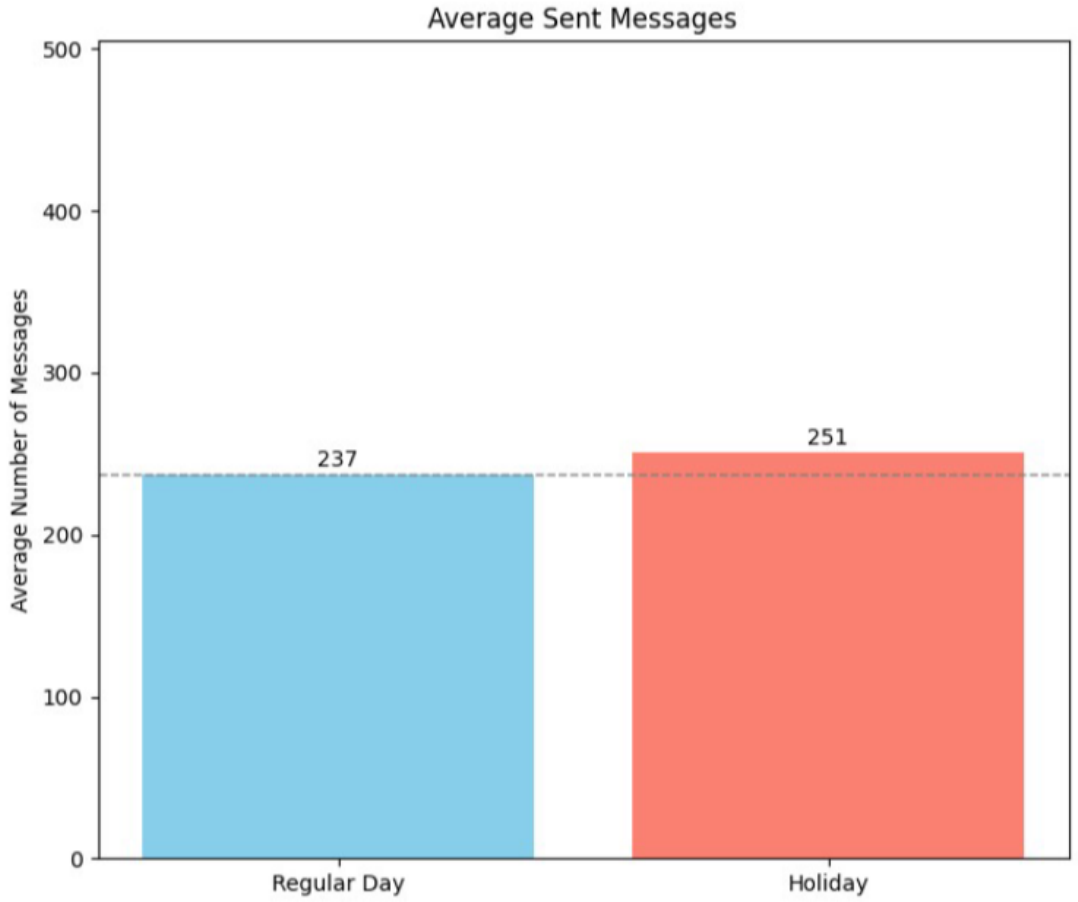
Activity patterns

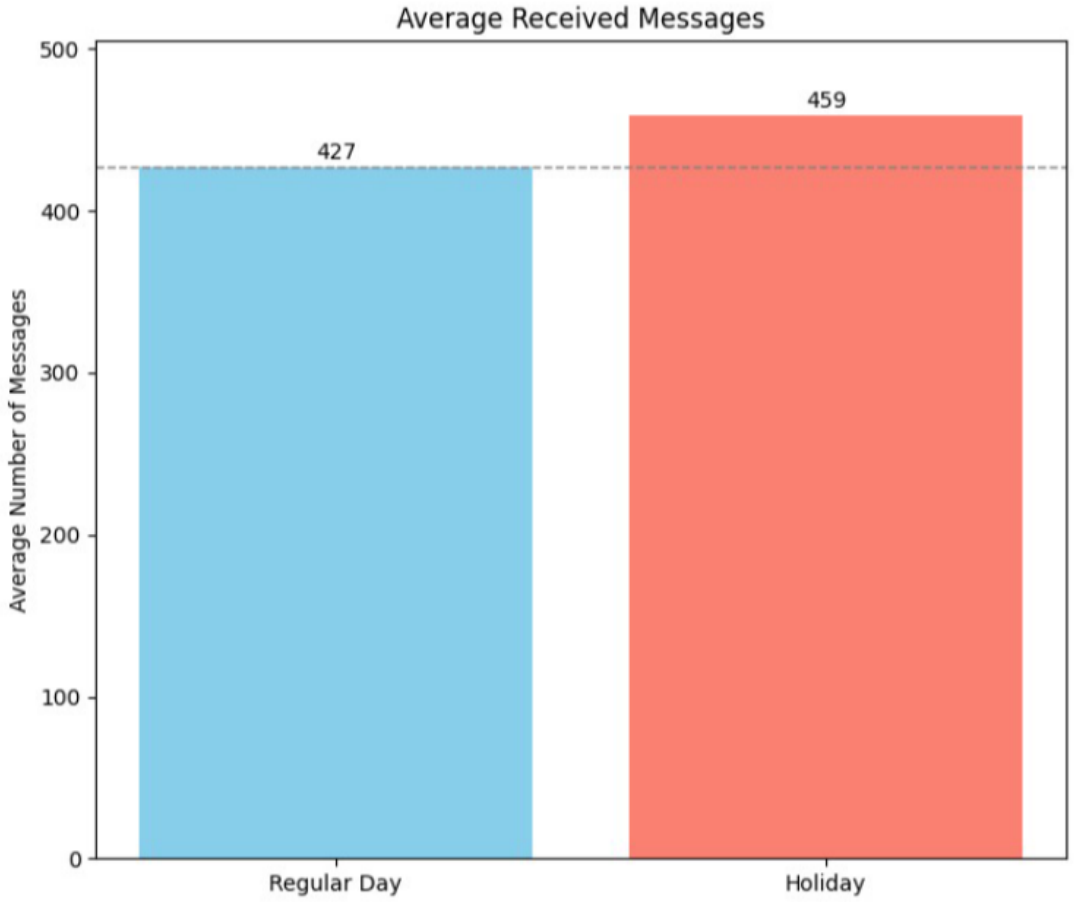
The most active day in terms of my activity is Sunday, while the day I am least active is Saturday(probably resting after long week).



Difference in messaging activity during holidays

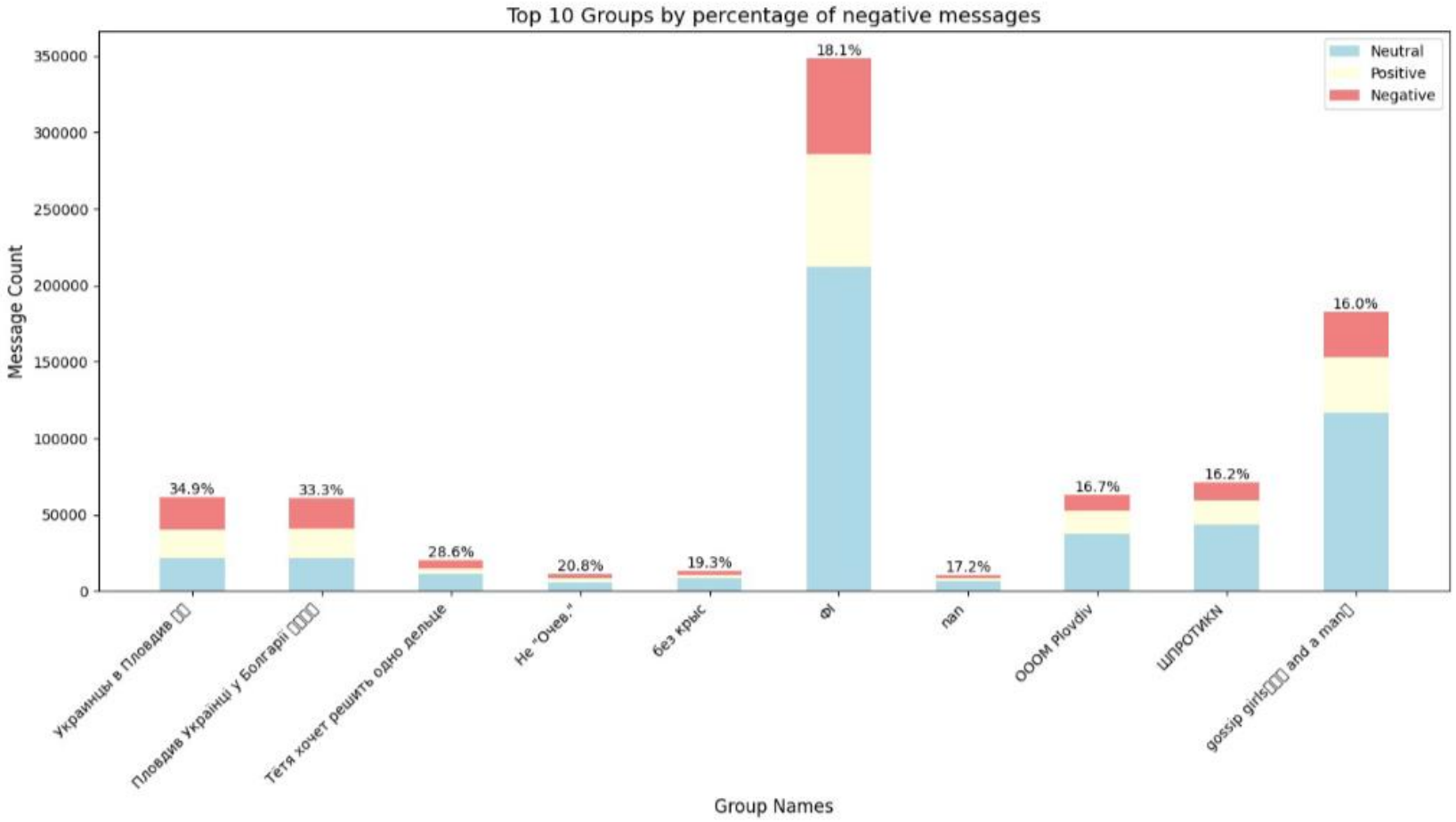
Messaging activity increases slightly during holidays, with both sent and received messages showing higher averages compared to regular days.

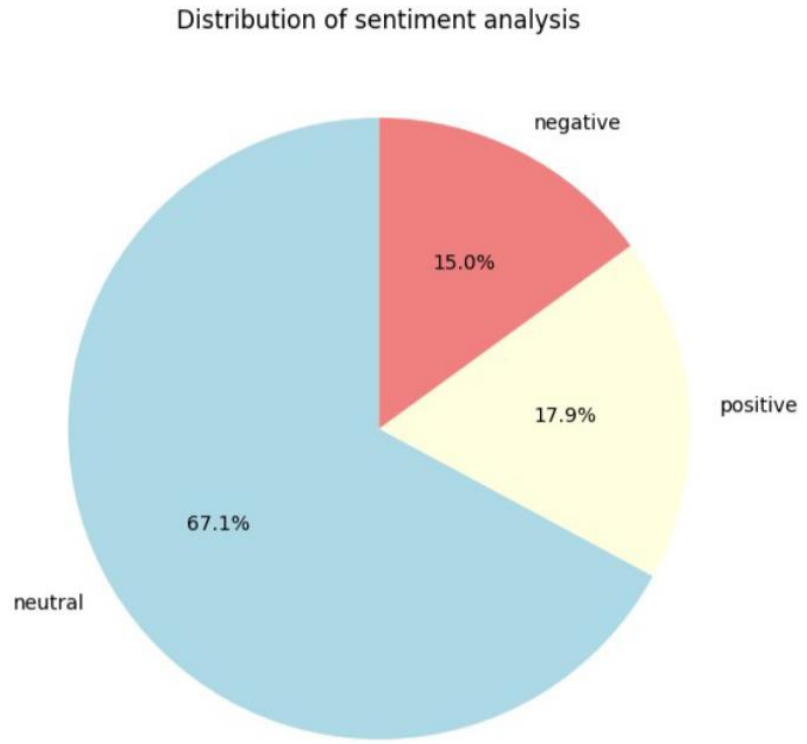




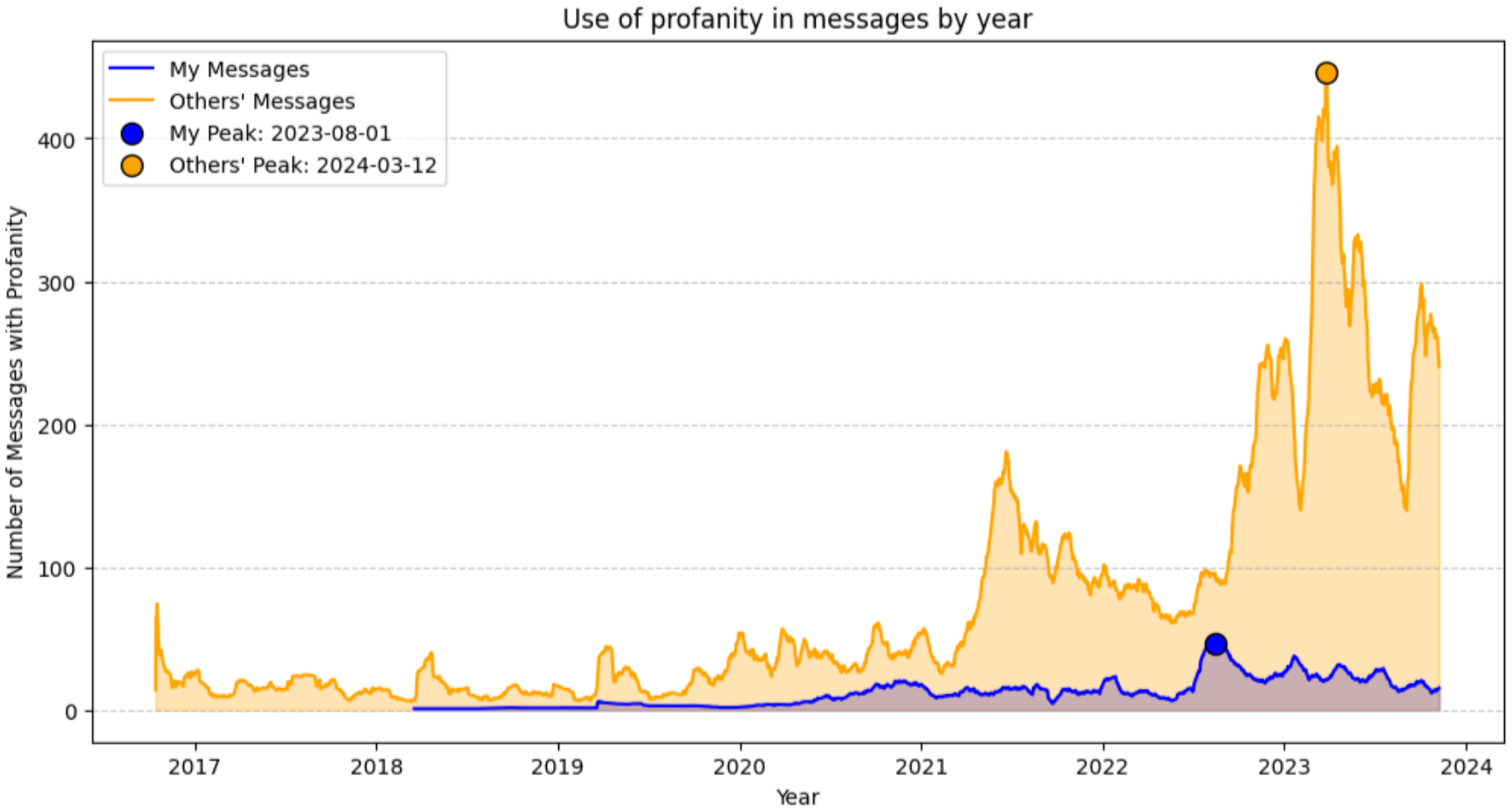
Sentimental analysis

The sentiment analysis reveals a clear distribution of message tones, with the majority being neutral, while a smaller portion shows positive and negative sentiments, with the bar chart highlighting the top 10 chats exhibiting the most negative messages.





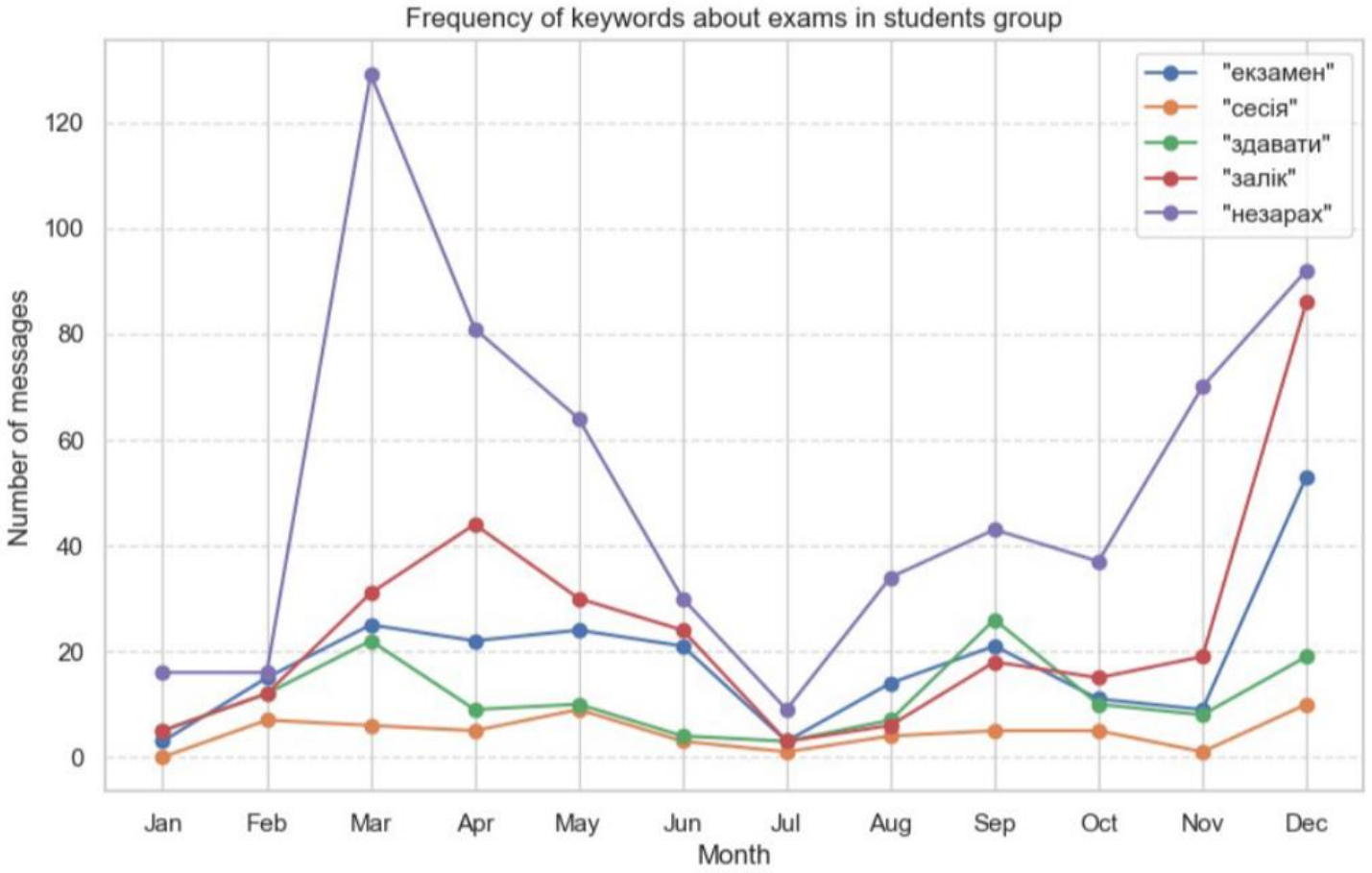
Use of profanity over time



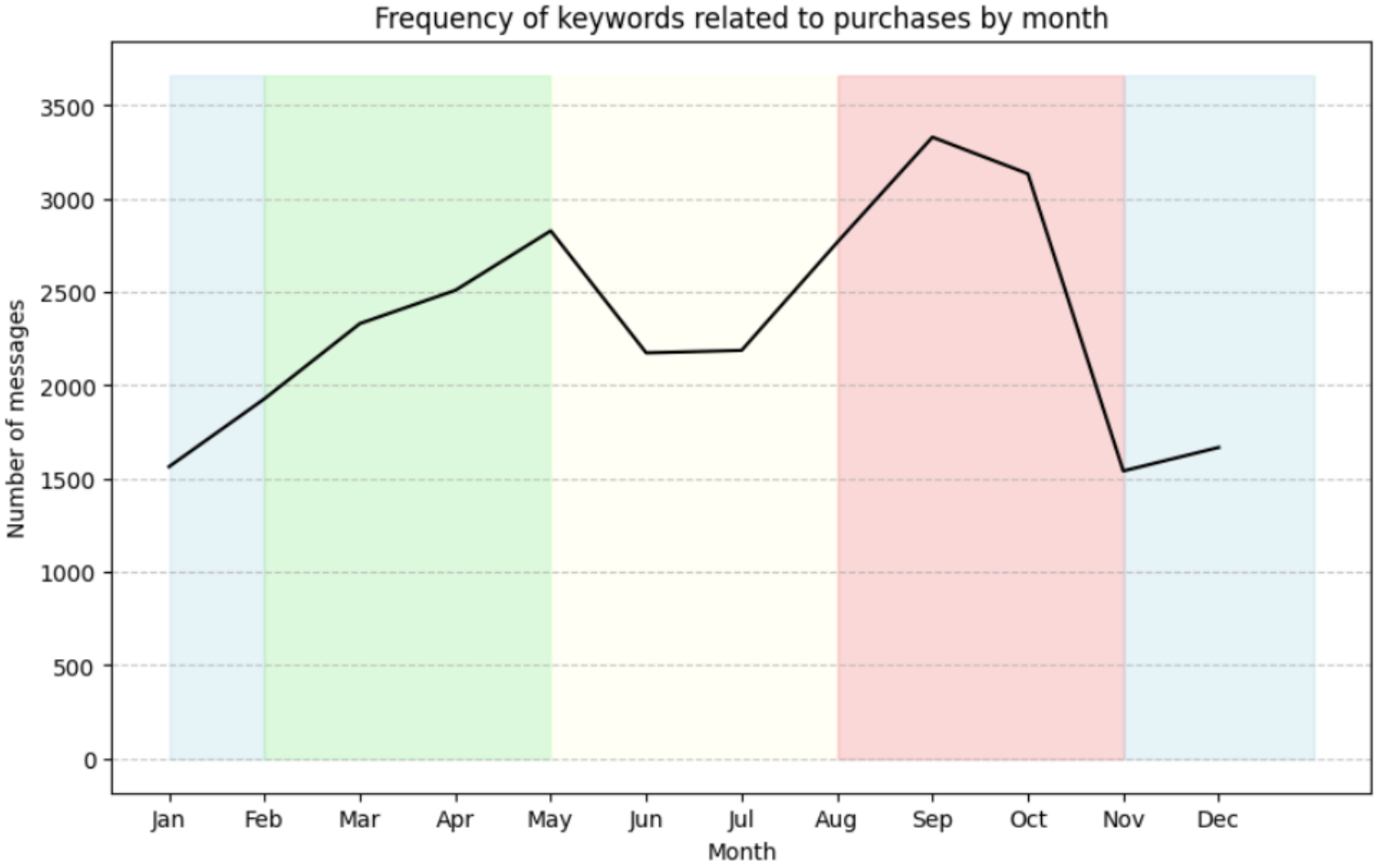
Profanity in others' people messages peaked on **March 2, 2024**, coinciding with a massive attack on Ukraine, including the destruction of an apartment block in Odesa. These trends highlight how external events influence emotional expression in communication.

Exam-related activity trends

- Activity in student groups significantly increases before exam periods, with a noticeable spike in mentions of "незарах" (dismissal) just before exams. This indicates heightened anxiety and pressure as deadlines approach. The trend suggests that students often delay preparation, leading to last-minute discussions and stress near exam sessions.



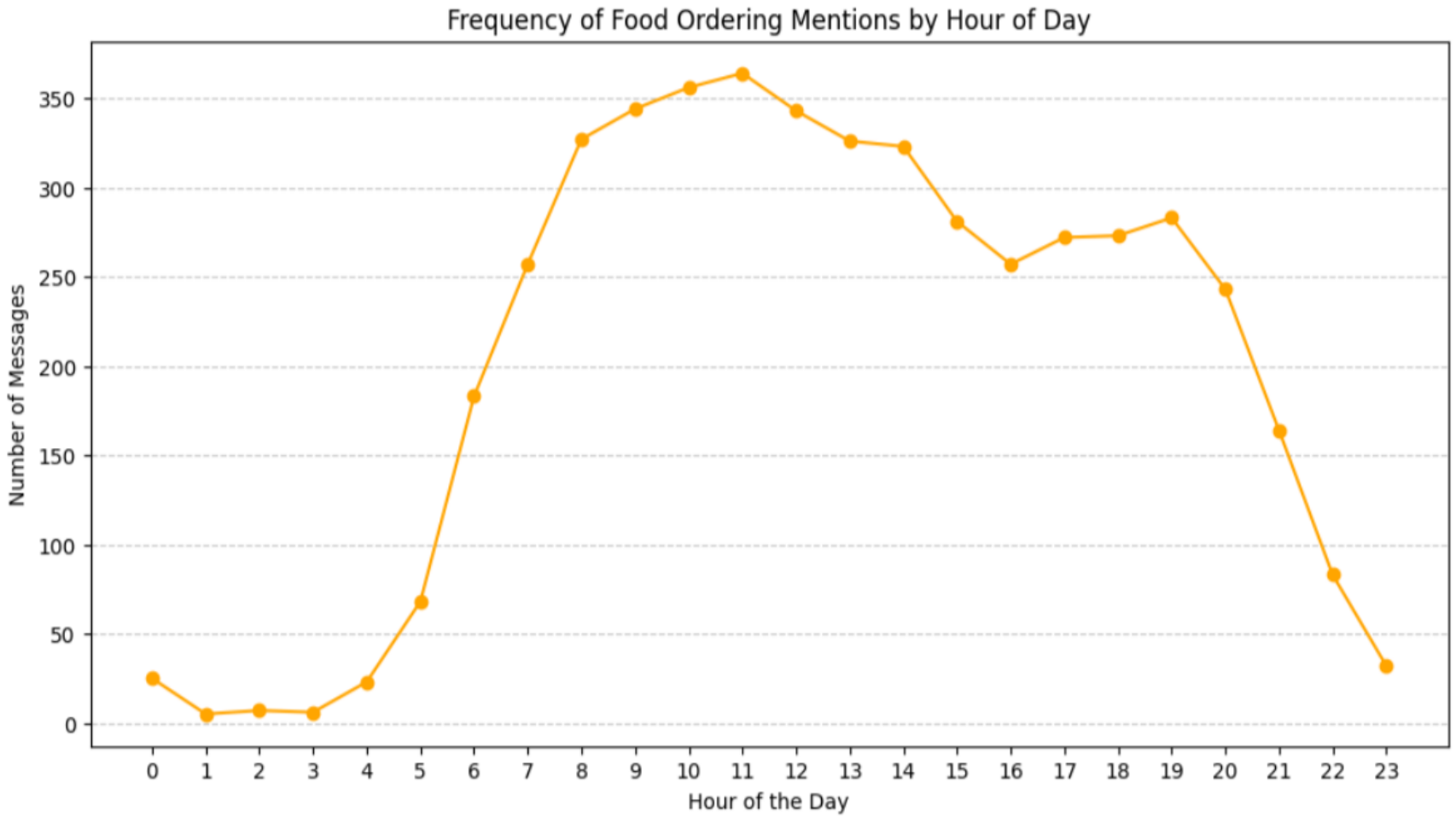
Frequency of keywords related to purchases by month



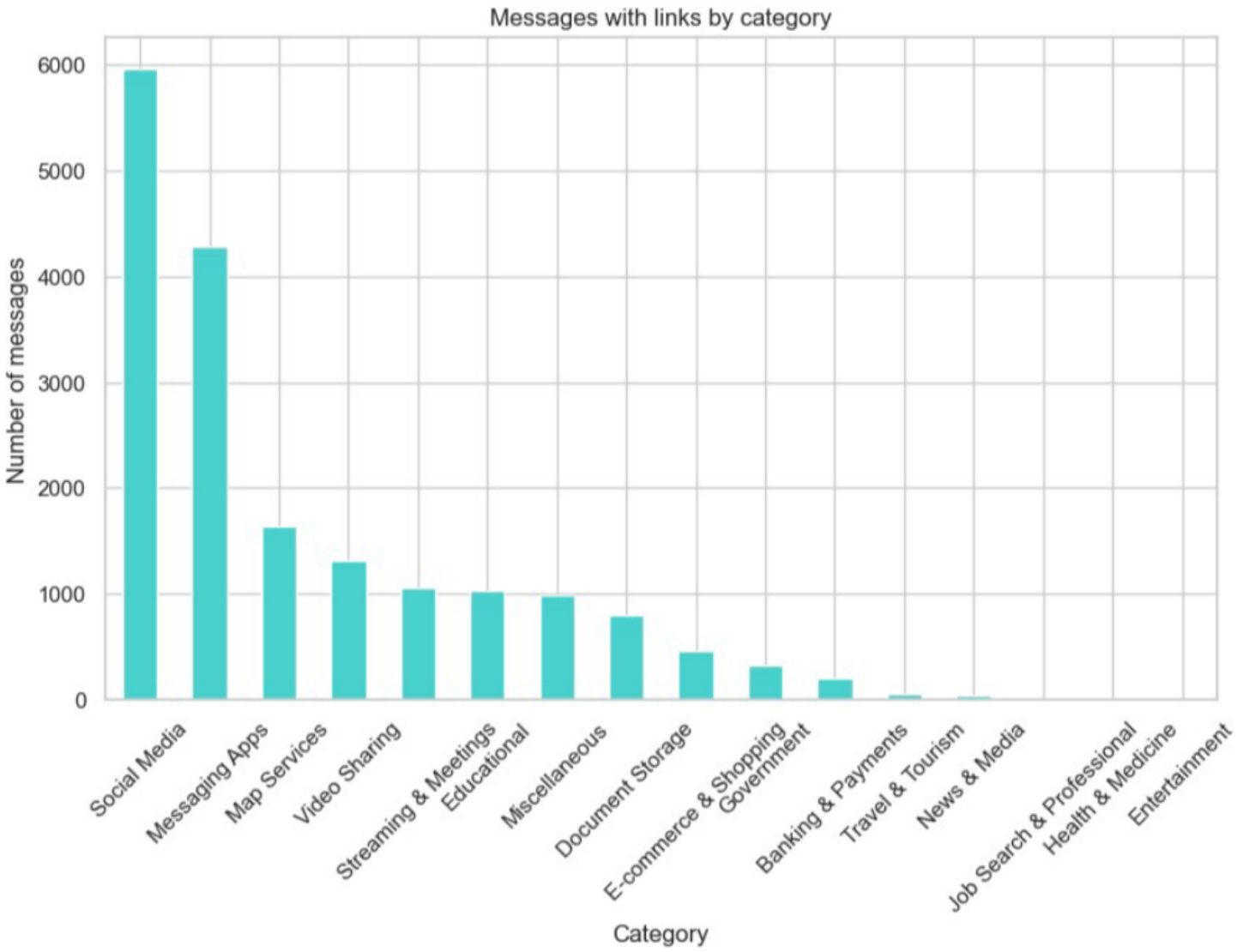
Purchase-related keywords peak in **May** and **September**, coinciding with seasonal changes when people are more likely to shop for clothing and other essentials. This pattern reflects consumer behavior tied to preparation for summer and fall transitions.

Frequency of food ordering mentions by hour

The data shows that mentions of food ordering increase just before meal times, likely as people start feeling hungry. More detailed research could help food delivery companies optimize their services and promotions around these peak hours.



Commonly Shared links



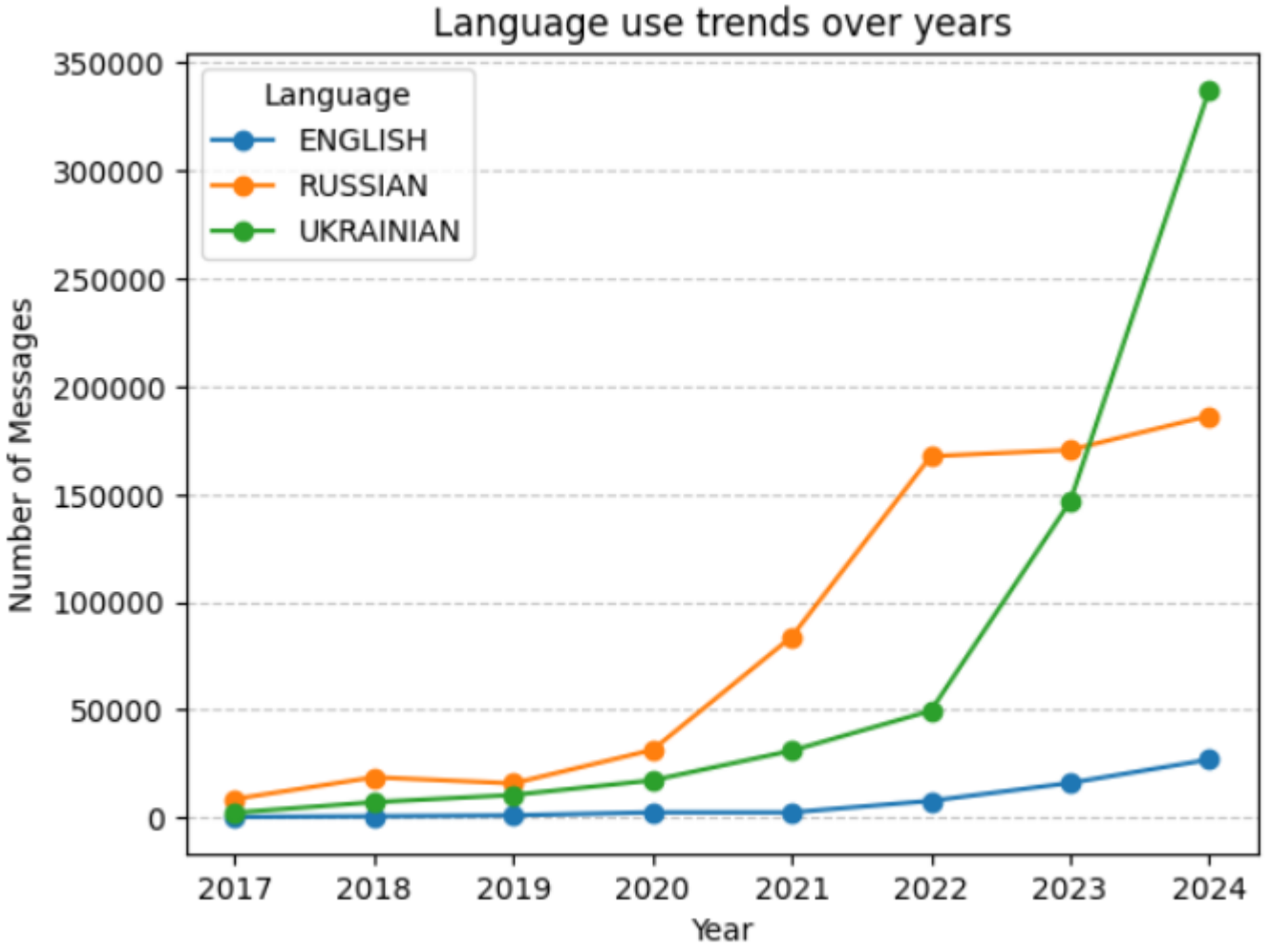
The largest number of messages with links belong to the "Social Media" category, indicating that users of the platform actively share content from various social networks and services. Other significant categories include "Messaging Apps", "Video Hosting", and "Streaming & Entertainment". This data provides insight into the types of content that users commonly share and engage with on the platform.

Key Findings

Language use trends over years

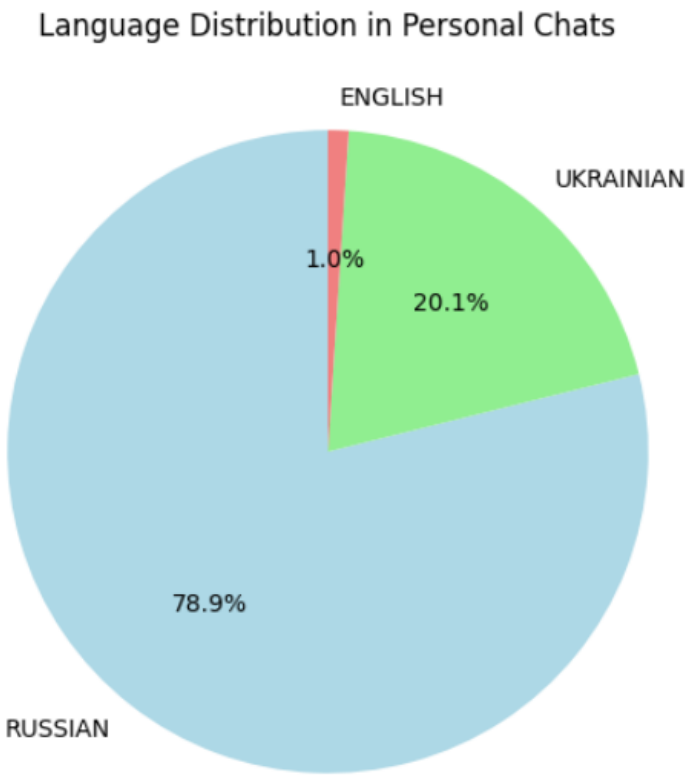
There is sharp increase in the use of Ukrainian starting in 2021, driven by the ongoing war and a strengthened sense of national identity.

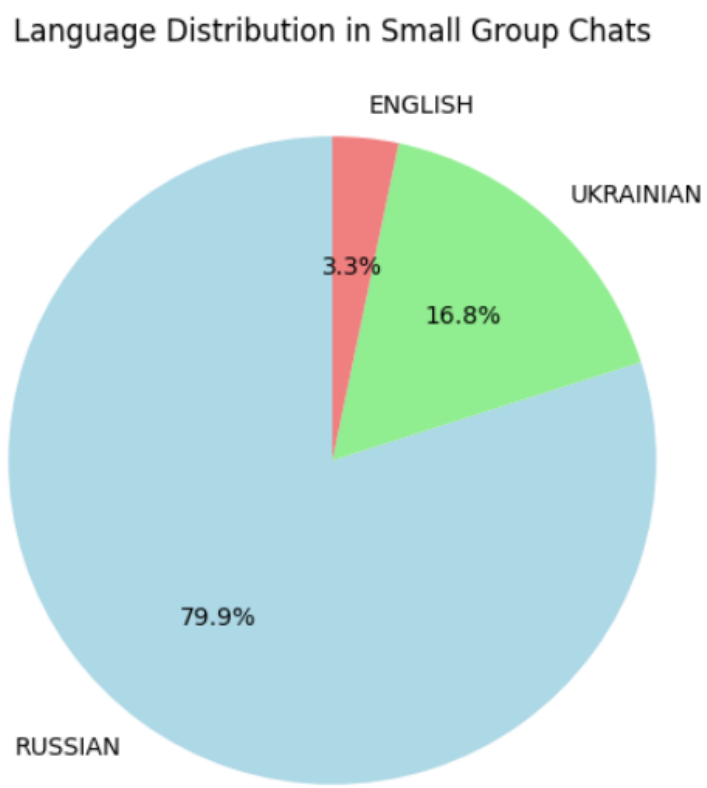
Russian usage has declined in growth, while English remains the least used but shows a steady, gradual rise. This reflects a cultural shift favoring Ukrainian over Russian in daily communication.

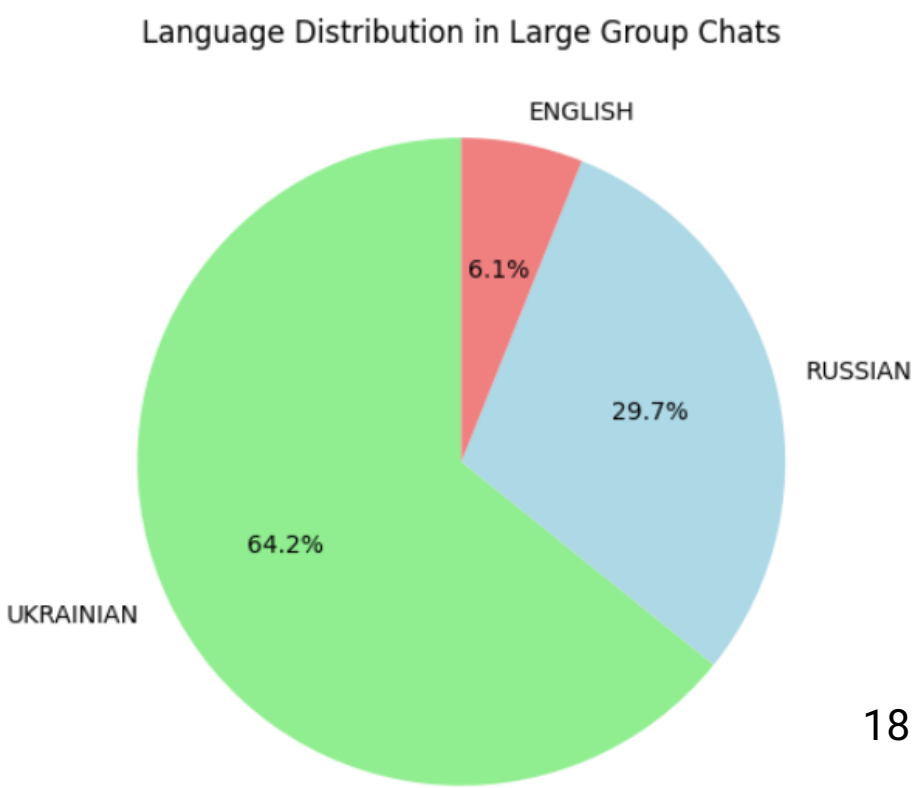


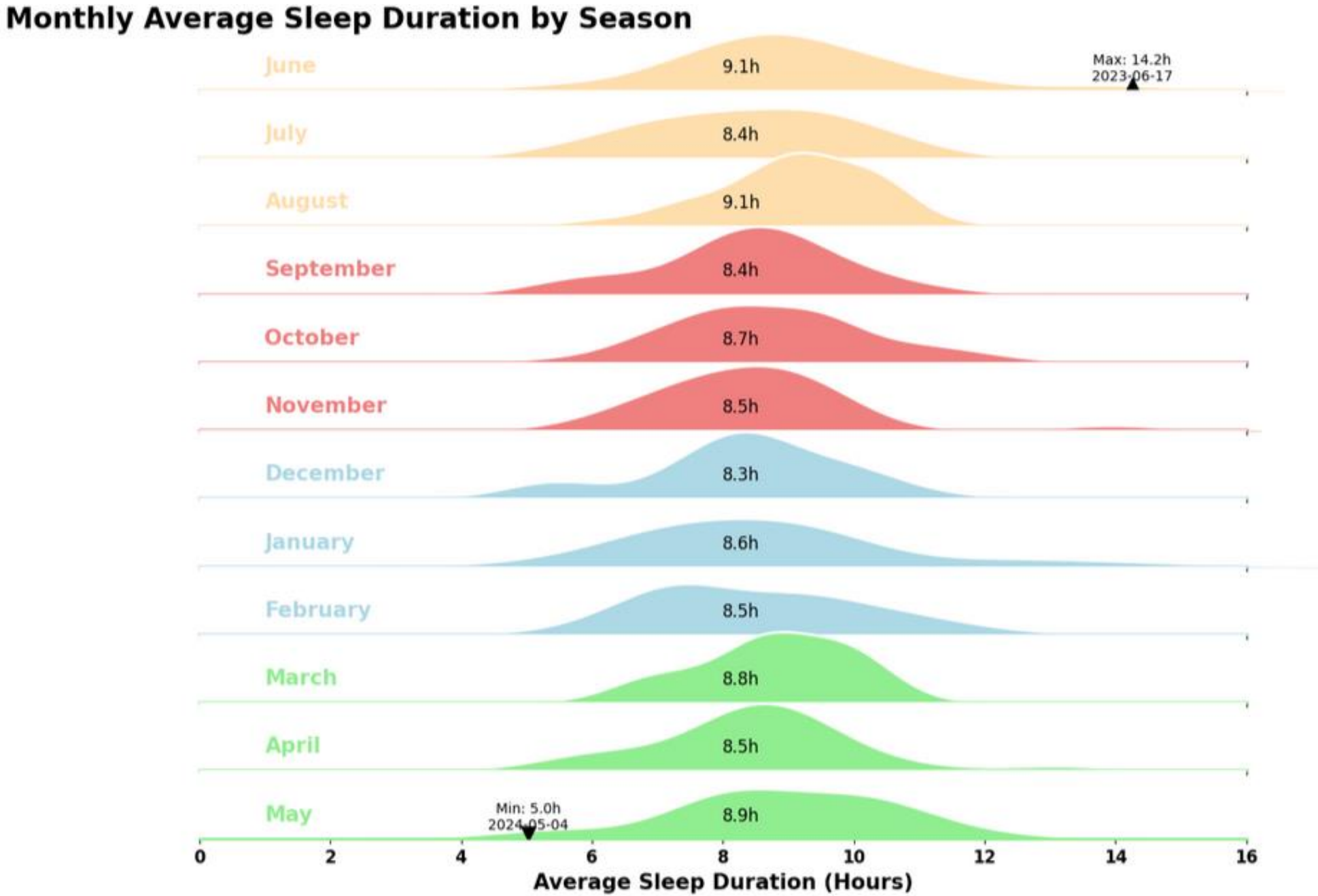
Language distribution across different chat types

In large group chats, Ukrainian accounts for 64.2%, reflecting its dominance in collective communication. In contrast, Russian remains prevalent in personal (78.9%) and small group chats (79.9%) due to the linguistic habits of close contacts. Ukrainian is used more in public spaces, while Russian continues to dominate private interactions.







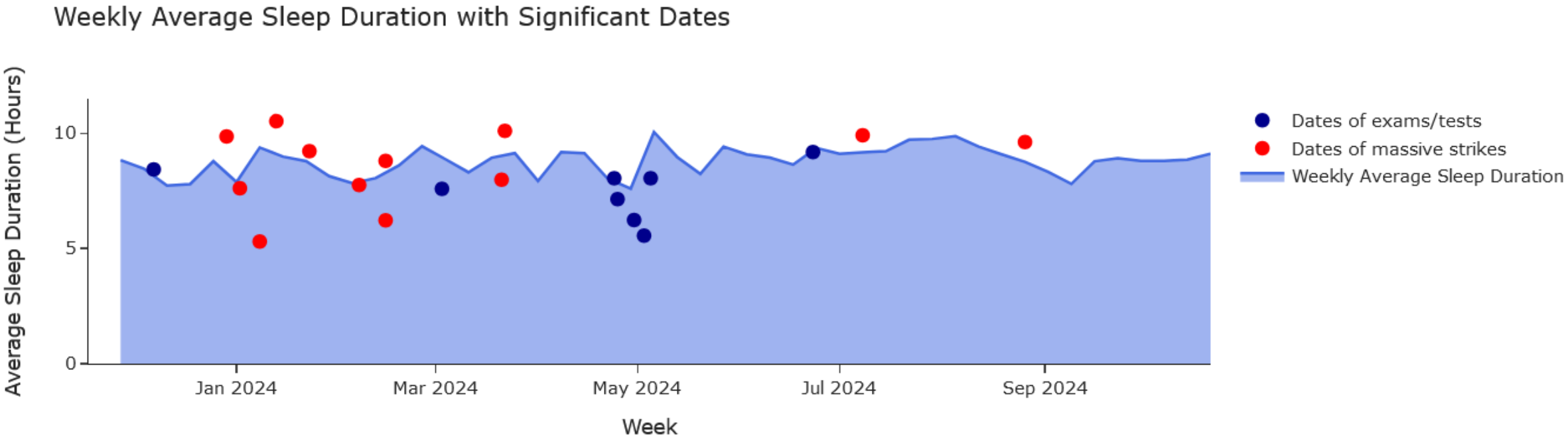

Monthly Average Sleep Duration by Season
June
9.1h
Max: 14.2h
2023-06-17
July
8.4h
August
9.1h
September
8.4h
October
8.7h
November
8.5h
December
8.3h
January
8.6h
February
8.5h
March
8.8h
April
8.5h
May
8.9h
Min: 5.0h
2024-05-04
0
2
4
6
8
10
12
14
16
Average Sleep Duration (Hours)

Seasonal trends and sleep patterns:

- **The longest sleep duration** of 14.2 hours was recorded on **June 17, 2023**, shortly after completing a significant NMT test for university admissions.
  The shortest sleep duration of 5.0 hours occurred in **May 2024**, during a period of intensive coding work on an assembler project with tight deadlines.

- **Winter:** Longer sleep durations, reflecting reduced academic pressure during January break.
  Autumn: Shorter and more irregular sleep patterns due to the start of a new academic semester.
  Summer: Moderate sleep durations, likely influenced by a more relaxed schedule compared to the structured demands of the academic year.

Impact of exams and strikes on sleep



Impact of exams and strikes on sleep

- Exam and test days are associated with shorter sleep durations, reflecting the stress and preparation required during these periods. Conversely, massive strikes also caused disruptions, as they often involved waking up during the night, reducing overall rest.

- Despite these fluctuations, a **general recovery trend** in sleep duration can be observed in weeks without major external stressors(for example, period from July to September), illustrating the ability to stabilize rest patterns after intense events.

Future Work

- Expanding **sentiment analysis** to capture more nuanced emotional responses.
- Exploring **deeper patterns in emoji** and sticker usage for richer communication insights. Identify instances where the sentiment of the emoji/sticker does not align with the sentiment of the message text (sarcasm, passive-aggression, or other forms of indirect communication)
- Examine how**, informal language** including slang and idioms, varies across user demographics and contexts to uncover linguistic patterns and evolution within messaging.
- Investigating **deeper connections** between messages, emotional tone, and external factors.

References

- Source Code: telegram-dialogs-analysis
- Dataset: (private data)exported Telegram messages using scripts SanGreel/telegram-data-collection