Final report for russian propaganda dataset analysis

Computational Social Science

Author: `Oleksandr Korniienko¶`

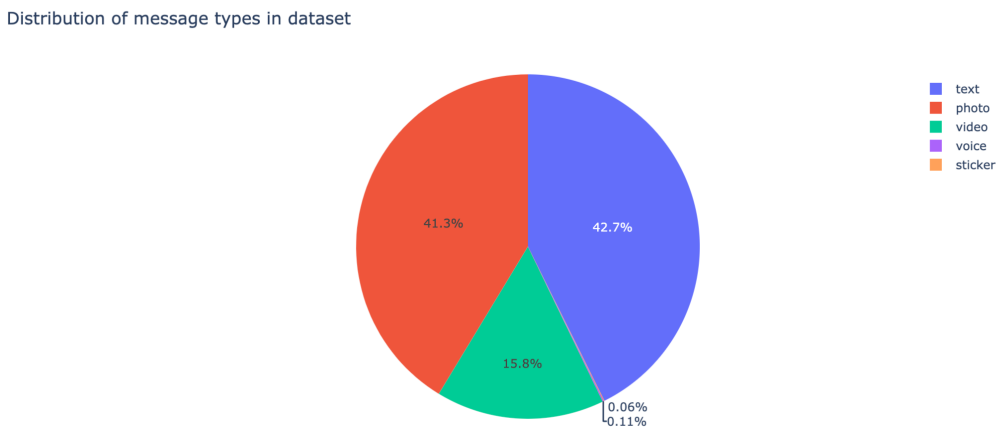
Source of data: private link
Source code: https://github.com/ironiksk/css-final

Data consists from 8M messages from 299 telegram channels and contains the following number of columns:

- Message - text message
- Views - number of views
- Reactions - lisy of reactions with count
- To_id - id of the channel
- Channel - name of channel
- From_id - id of channel the message forwarded from
- Type - type of the message (photo/video/reaction/text)
- Date - data of post

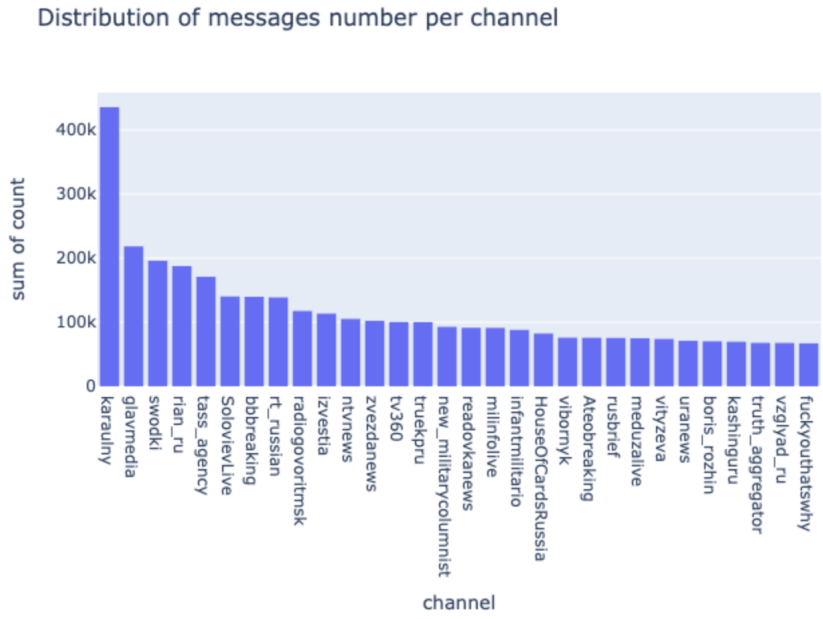
Data analysis

Message type analysis

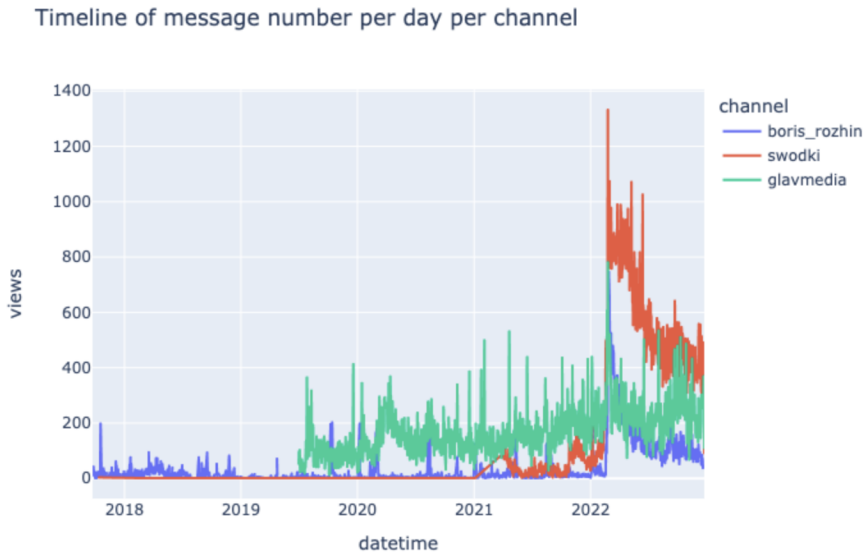


84.1% of all messages contains text or photos. ~15.7% are mostly videos, other are stikers or voice messages.

Distribution of message number per channel



The most active channel is karaulny; glavmedia and other news channels have mostly the same number of messages for the investigated period from 2015-09-22 to 2022-12-26.

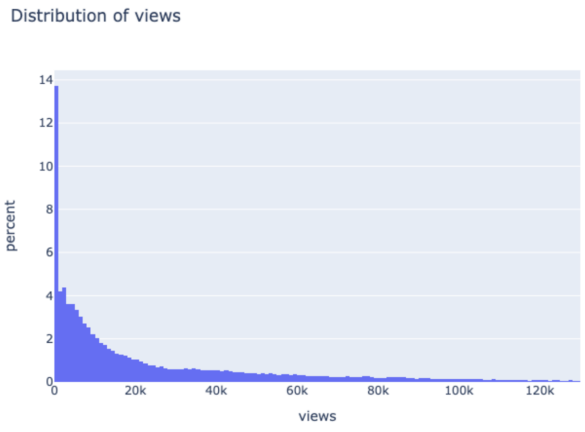


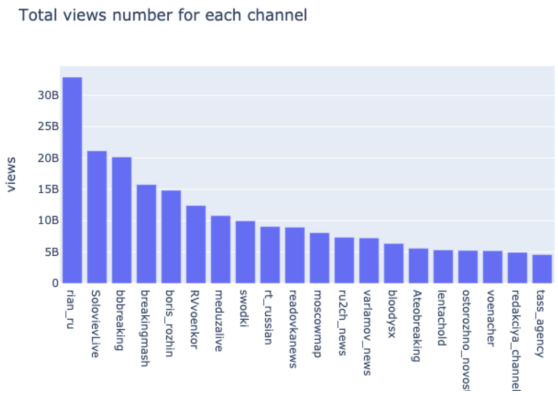
There are a few abnormal number of messages per channel for 24 Feb 2022 from:

- swodki
- bosis rozhin
- glavmedia

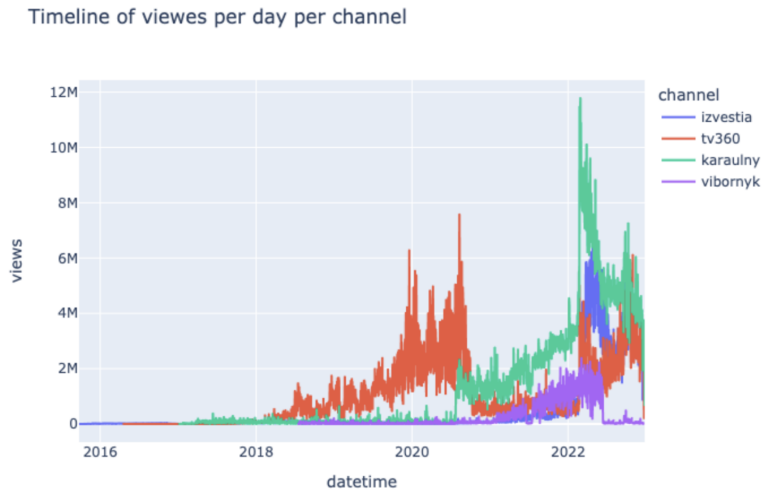
These channels were considered the sources of news about the war during the first days.
Also, this metric shows an abnormal message count per channel (peaks with a number of 500! messages per day).
Message numbers per channel also have a clearly visible seasonality trend with a duration of one week.

Distribution of message views in data

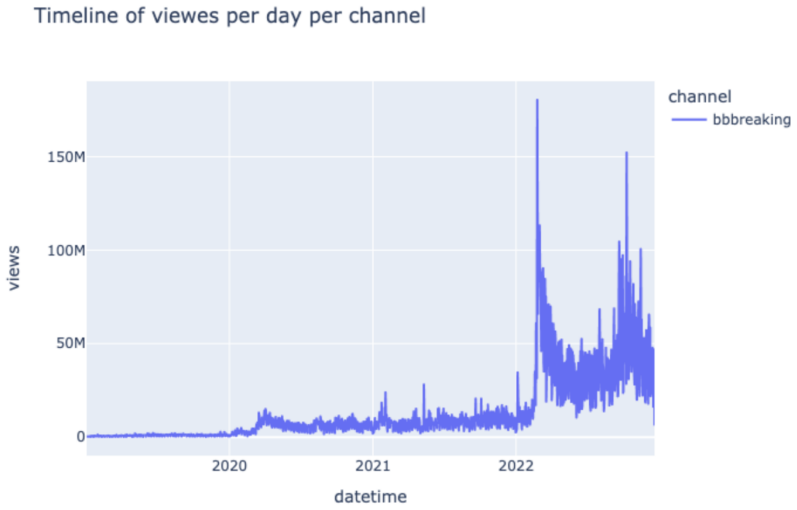




The views distribution seems to be a log-normal distribution but with long tail. The most viewed channel is rian_ru, soloviev and bbbreaking.

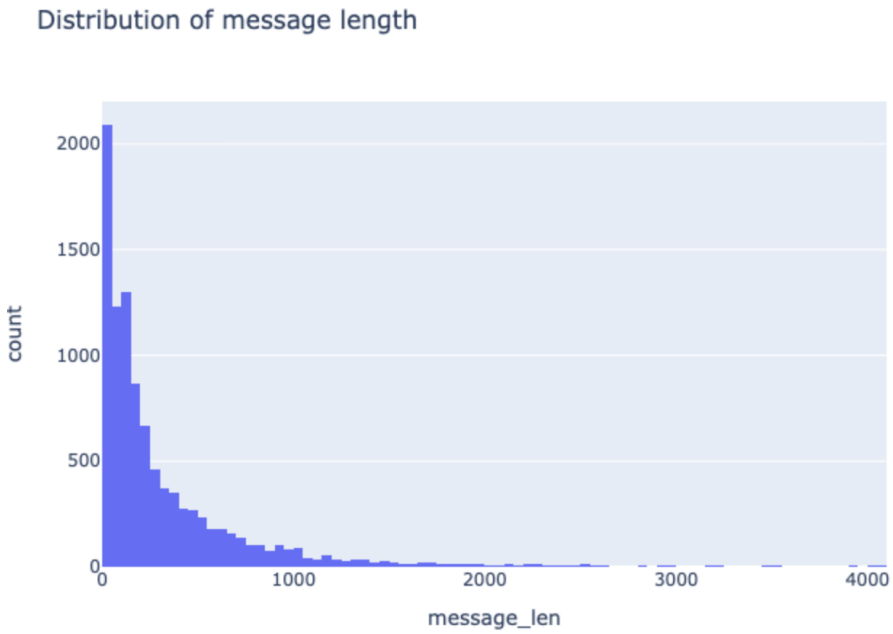


Views number per day per channel allows us to identify the general popularity of the channels and how it was boosted in time acocrdint to SMM techniques. Outliers bring us the most important dates like full-scale invasions (all channels react to this) and other political and war events.
One example of channel boosting is the tv360 channel, which took a bump at the beginning of 2020 and stopped in the summer of 2020.



Another interesting channel is bbbreaking with “trampoline” at April 2020.

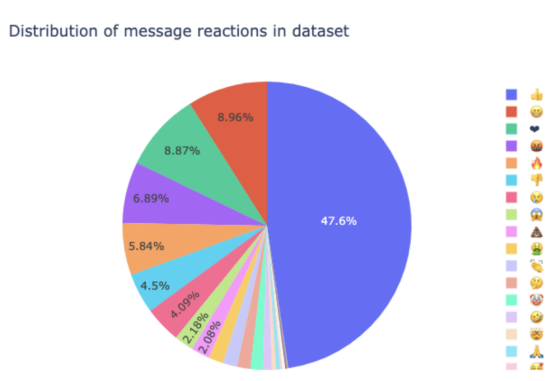
Distribution of message length

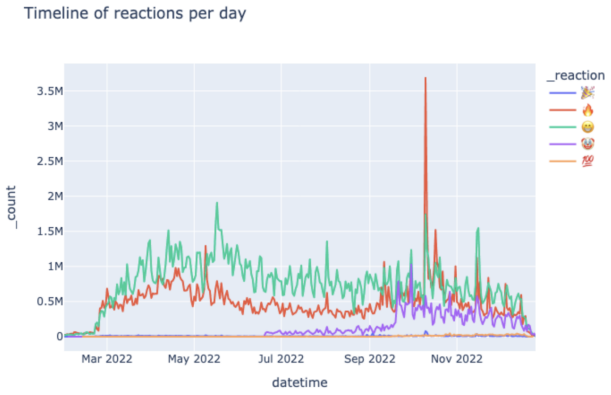


The mode for message length is ~110 characters.
Also, we see a lot of empty messages that contains only photos (see notebook for examples).
A very long message came from different persons, who, I believe, are bloggers and post stories.

Analyze reaction distribution in dataset

76% of all messages do not contain any reaction, but others could be a nice feature for analysis of bot networks and general society reactions on different events.



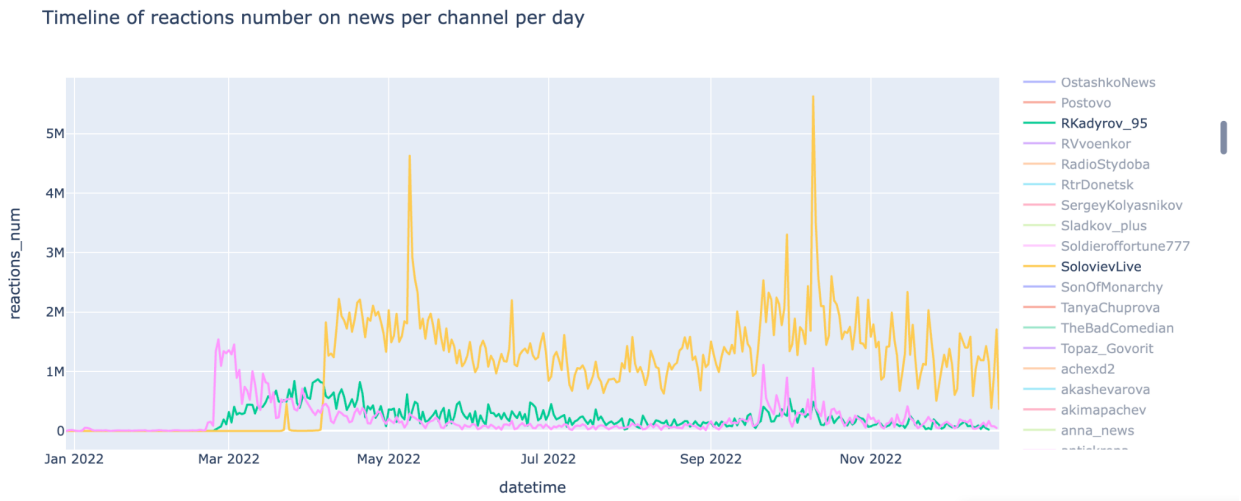


Brief analysis of of plots above shows:

- The most used emoji is a thumbs-up (47%). The next common emojis are lol and heart (~8%)

- Fire 🔥 emoji has two peaks: Sep 30 and Oct 10 2022
- The fire 🔥 emoji also has a significant bump Oct 10
- The clown 🤡 emoji has a couple of bumps: The most significant Sep 30, Sep 21, Oct 27, and Nov 15. The trend significantly increased starting from Sep 21 and has a decreased trend
- angry emoji was used mostly Oct 8
- emoji 💯 seems to have a linear trend starting Sep 11, 2022, and continues growing...
- 🎉 tada emoji has several bumps that matches with clown

According to the number of the reaction in time dependency, we see a set of bumps (only 3 out of all channels showed: varlamov - pink, rkadyrov - green, and soloviev - yellow).



Users (or bots) have different reaction numbers during the time; they all activate on 24 Feb 2022, and starting from May 2022, the reaction number slowly decreases. The next period of high engagement in reactions is September 2022. Let's check the dates of the peaks:

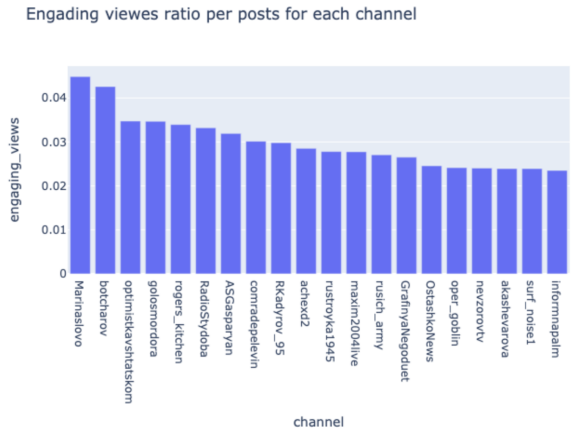
- 5 Jan 2022: Russia released plans for a new wave of mobilization: https://www.pravda.com.ua/eng/news/2022/12/30/7383028/
- 21-28 Feb 2022: Full-scale invasion
- 21 Sep 2022: Russia declared a partial mobilization of military reservists https://en.wikipedia.org/wiki/2022_Russian_mobilization
- 10 Oct 2022: Massive missile attack on Ukraine infrastructure supported with high level of propaganda and disinformation https://www.wilsoncenter.org/blog-post/ukraine-quarterly-digest-october-december-2022
- 15 Nov 2022: Massive missile attack on Ukraine infrastructure

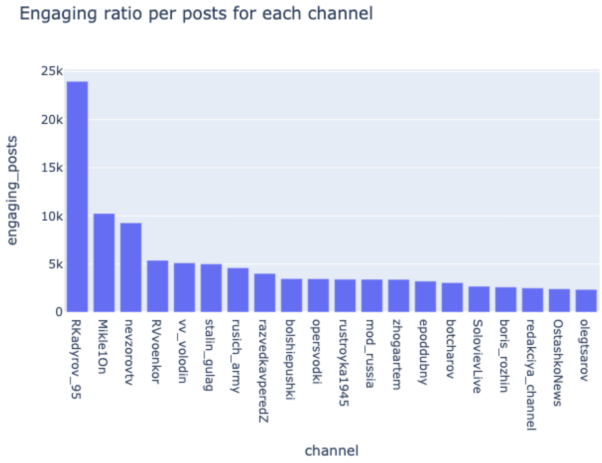
Engagement ratios

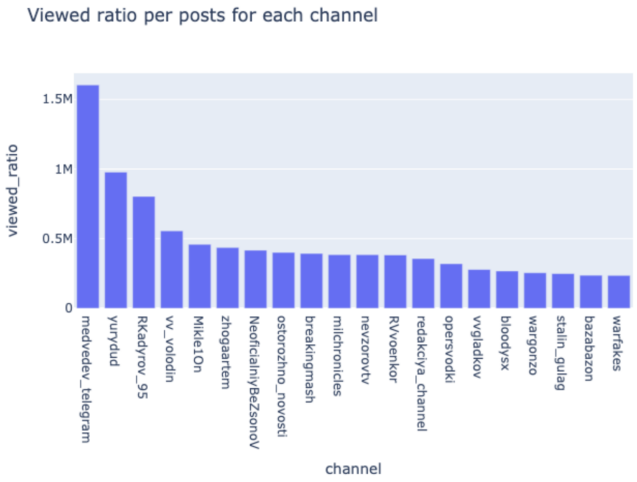
To get a better understanding of user's behavior let's introduce engagement ratios:

- engaging_views: ratio of reactions num per number of views.
- engaging_posts: ratio of reactions number per number of posts for each group
- viewed_ratio: ratio of views number per number of posts

Let's check separately a number of views for specific emojis as well.



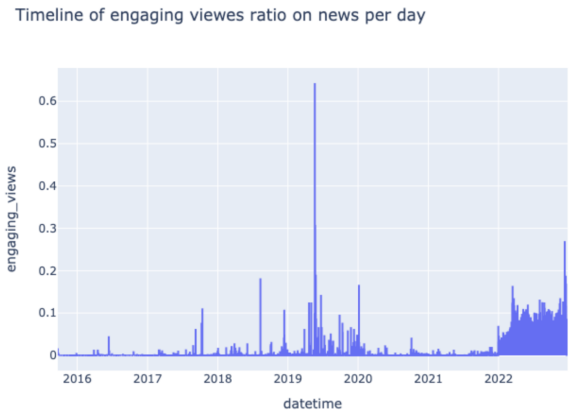


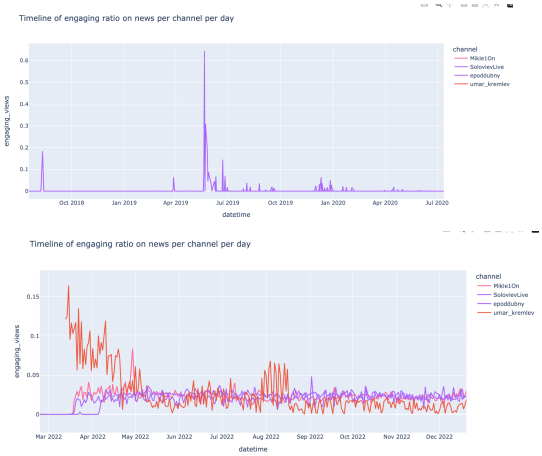


We see the most engaged channels in terms of a number of reactions per number of viewes related to private (nonnews channels), like:
- marinaslovo (4%)
- botcharov (4%)

In this case rkadyrov also is in top 10 engaged channels. Glavmedia has emgagement ratio of 0.5%.
The most engaged channel regarding the number of reaction numbers per post is Kadyrov, then mikle1on.
The most viewed channel is medvedev, yurydud and rkadyrov.

In all 3 groups, rkadyrov is in the top engaged channels, so a well-deserved name is "TikTok Troop".

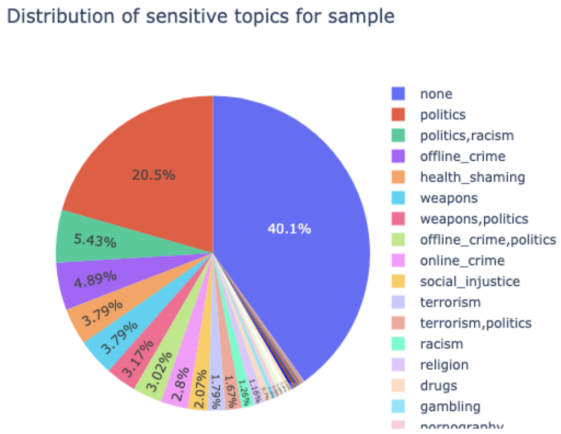




We see significant engagement bumps are for epodduny at Aug 2018, May 2019.
We can see some interesting trends:
- umar_kremlev has inreased engamement at the beginning of channel existence
- kadyrov and mikle1on has similar and mostly constant engamenet starting from March 2022

General distribution of messages between sensitive topics

Paper with description of topics presented here.
https://arxiv.org/pdf/2103.05345.pdf

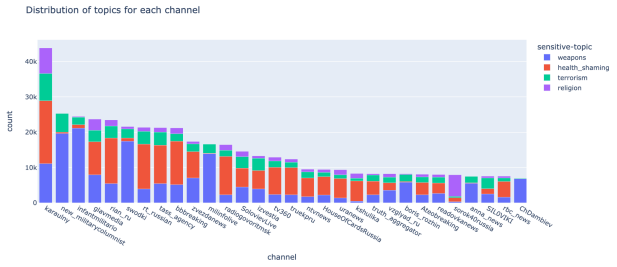


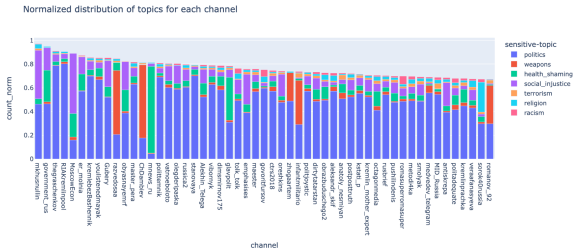
The most common topic is politics.

Next general most common topics are: weapons, online/offline crime, health_shaming, rasims, social injustice.

Unexpected? no!

Let's check the main directions for each channel (remove none and politics as most commonly used topics)

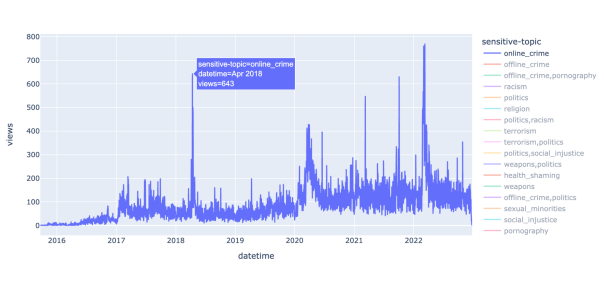


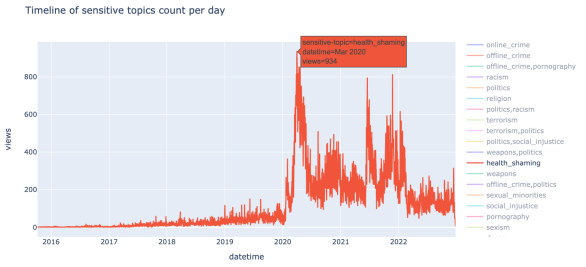


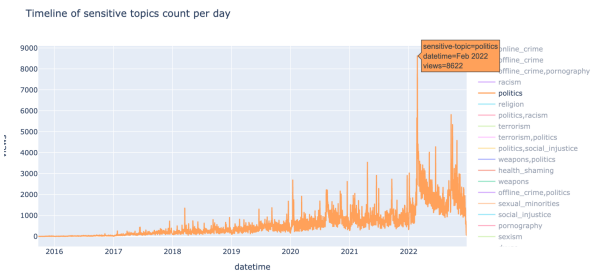
Sensitive topics classification allows to classify channels direction, for example

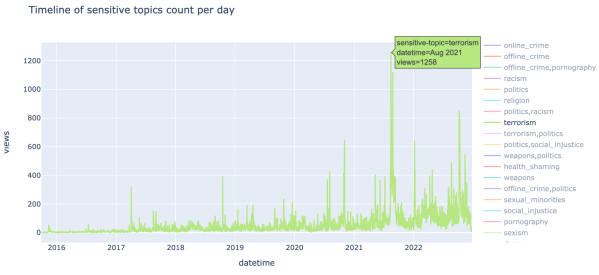
- milinfolive and infantomilitario the channel is about weapons
- mnews_ru 73% of content has about health shaming
- criminalru and ikakprosto are about online crime for 47% and 21% about offline crimes
- umar kremlev has 25% of religion content? the next in top religion channel is sorok40russia
- chdambiev and razvedosaa posts mostly about weapons (50 and 60% respectively)
- the most rasism channel is mnogoznai
- top gamling channel is n_zackhaim, tikandelaki
- maxim2004live posted about drugs the most

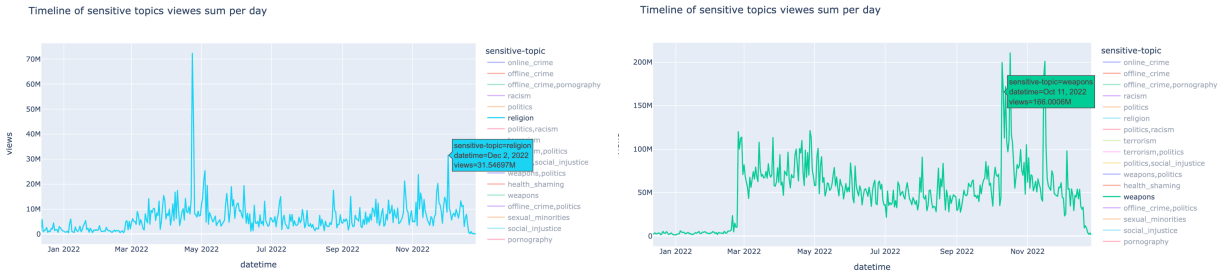
Let's check how general topics distriuted in time.











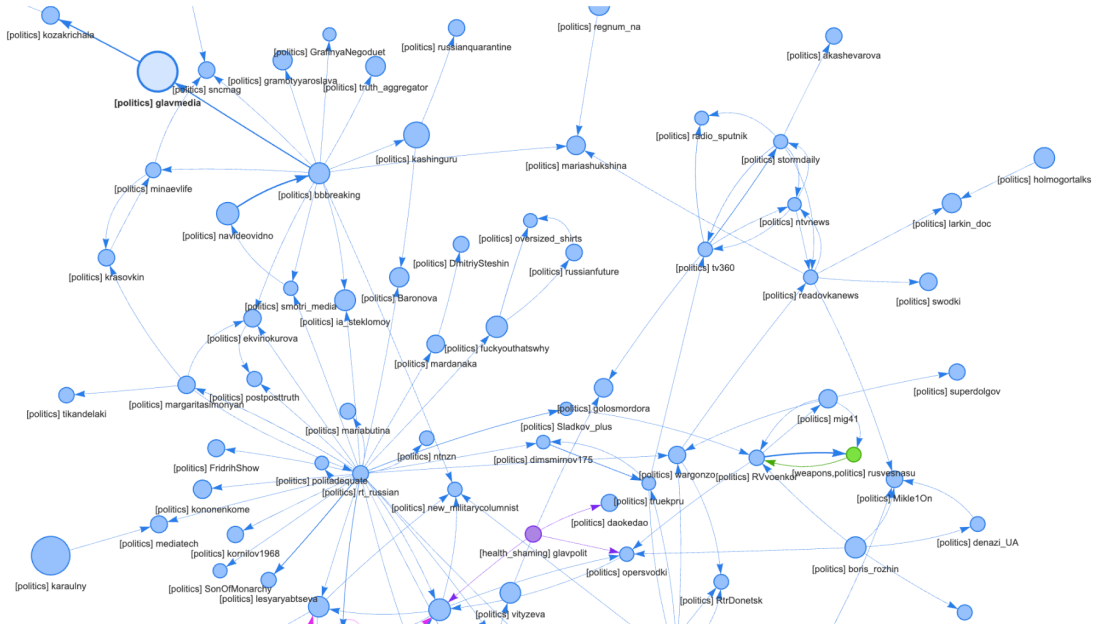
Observations

- April 2018 - Russian intervention in the Syrian civil war
- health shaming news broadcasted April 2020, this is mostly because of Convid-19 quarantine, so people started working from home and do not move lot.
- the bump of number of political news happend 24 Feb 2022. Also there are some other bumps:
    - May 9 2022
    - June 17 2022, Russia will use nuclear weapons if its sovereignty is threatened - Putin (https://www.pravda.com.ua/eng/news/2022/06/17/7353142/)
    - Sep 21 2022, Russia annonsed plans for mobilisation
- Number of political news has seasonality for 7 days with peak on Thursday
- The bump for number of religion news was 23-24 April - ortodox Easter, and 2 Dec 2022 - nationalisation of Kyiv Pechersk Lavra
- The most viewed news about weapons was Oct 8 - Oct 11 2022, Crimea bridge destuction, haha. We also see posts about terrorism started earlier and after continue with bump for weapons topic.
- The second most viewed weapon posts was at Nov 13 2022, nothing special in news but still a bump...
- An interesting fact - reactions started collecting from Feb 24 2022

Graph analysis

The following section analyze the importance of the reposts network for given data and topics analysis. Importance is calculated with PageRank alogithm.

To reduce number of arcs and nodes and selects only extreme values we use graph pruning and nodes filtering according to the importance and number of arcs that connects channels by repost news.



The extremely 1% most important nodes forms the network about politics.

PageRank algorithm selects as most important nodes:
- karaulny
- glavmedia

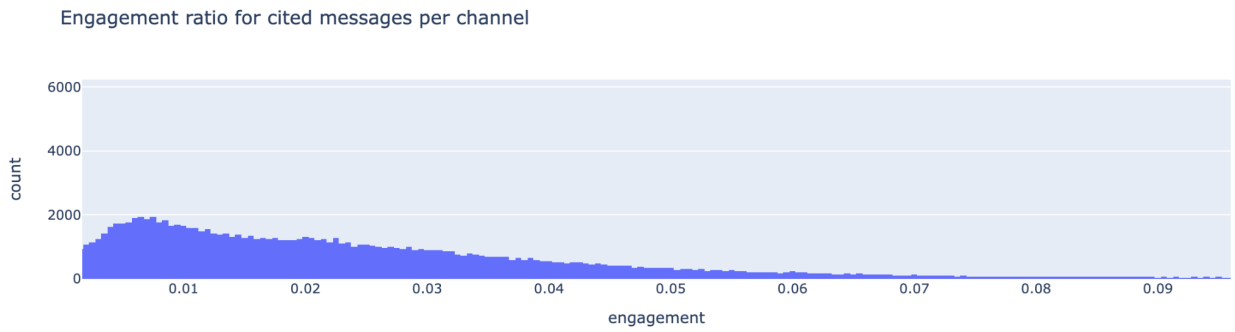
We also see separated clusters of news groups with same topics:
- offline_cime criminalru and chhkogpu
- politics: kstati_p, otsika_bld ctrs2019 and rastriga
- rasstrelny and rlz_the_kraken
- skvir (about drugs), bogemasranaya (politics) gori_spb (online crime) thynk (body shaming) impnotbozhena (politics)

rt_russian, bbbreaking, sashakots seems used mostly as a sources of news for different other nodes, but is not considered as important.

Analyze engamenent for the graph

Let's check the distribution of the engagement ratio for this research and find outliers.

Then let's plot connection in graph for abnormally reacted messages.

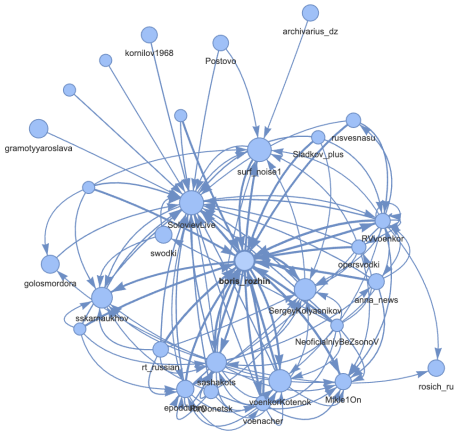


It seems the engagement ratio has log-normal distribution but with small bump at 0.02. Let's find channels with distribution that is not log-normal by using Kolmogorov-Smirnov test.

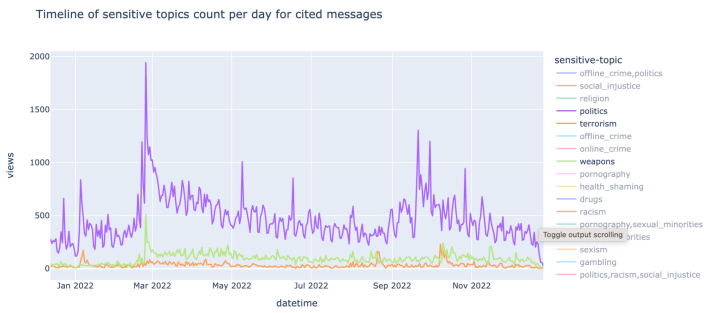
The list of channels that are not pass the statistical test on log-normal distribution are:

'Hinshtein' 'Mikle1On' 'NeoficialniyBeZsonoV' 'Postovo' 'RVvoenkor' 'RtrDonetsk' 'SergeyKolyasnikov' 'Sladkov_plus' 'SolovievLive' 'Topaz_Govorit' 'anna_news' 'archivarius_dz' 'boris_rozhin' 'dimsmirnov175' 'ebobo_rus' 'epoddubny' 'glavmedia' 'golosmordora' 'gramotyyaroslava' 'kornilov1968' 'ksbchk' 'kshulika' 'kstati_p' 'minaevlife' 'mozhemobyasnit' 'navideovidno' 'norin_ea' 'obyasnayemrf' 'opersvodki' 'otsuka_bld' 'parfentiev_club' 'rasstrelny' 'razvedosaa' 'redakciya_channel' 'rosich_ru' 'rt_russian' 'rusbrief' 'russ_orientalist' 'rusvesnasu' 'sashakon' 'sashakots' 'sskarnaukhov' 'strelkovii' 'surf_noise1' 'swodki' 'tikandelaki' 'varlamov' 'varlamov_news' 'voenacher' 'voenkorKotenok' 'warfakes' 'wargonzo'

In this list we see rt_russian, rkadyrov and others channels with high level of viewes or engagement.
Let's build this network and investigate connections between channels.

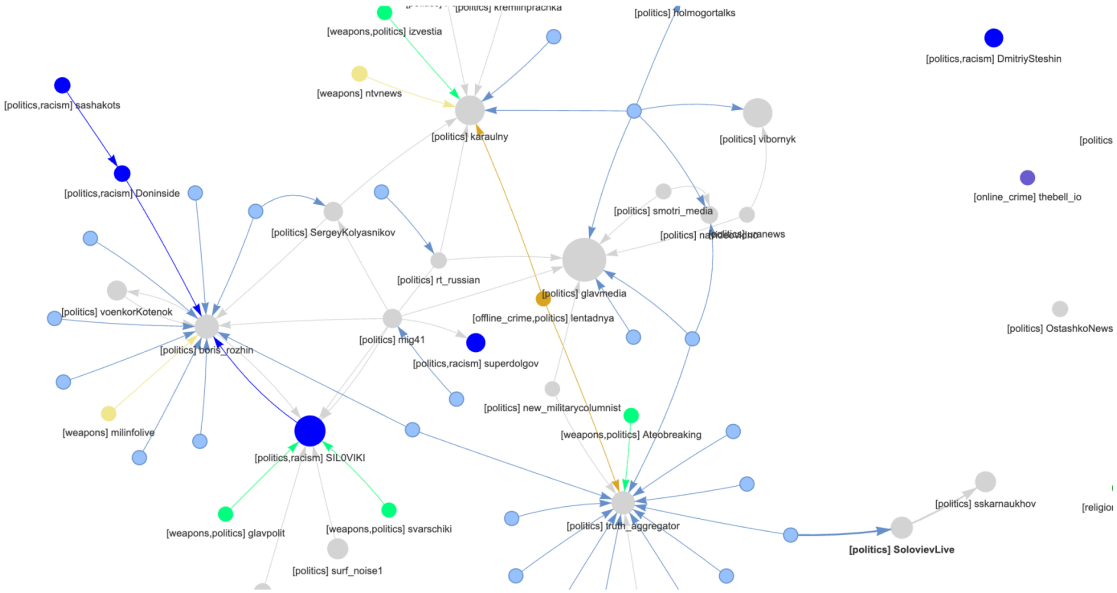


We see, soloviev boris rozhin, rvonekor, voenkomketenor are highly connected between each other.
Let's check distribution of cited messages only through the time



We see the significant increase fo messages about politics and weapons stating from 24 Feb 2022. Let's check the connection of channels at 24 Feb 2022.

Also let's recalculate node importance because it could be different from the general.



We see the most importrant according to PageRank node keeps the same - glavmedia. But in this distribution appeared other nodes like siloviki and truth aggregation with the main topic about politics.

Conclusions

It was investigated russian propaganda dataset based on telegram channels.

- The most active channel is karaulny; glavmedia and other news channels have mostly the same number of messages for the investigated period from 2015-09-22 to 2022-12-26.
- The most viewed channel is rian_ru, the next is solovievlive and bbbreaking.

- The iteresting observation here is that number of views for mostly all channels increased starting from 24 Feb.
- Also it was discovered that some channels boosted popularity with number of viewes for some period of time without correlation on other channels. This could be a reason of artificial viewes boosting. One of the example is bbbreaking channel with bump on April 2020.

Engagement and views increased significantly for set of important dates like 23 Jan 2021 (medusalive, bbbreaking): Protests in Russia in support of the opposition leader Alexei Navalny, victory days in 2nd WW 9 May, 11 May 2021 a school shooting occurred in Kazan, 21-28 Feb 2022 Full scale invasion, and others.

This reaseach uses sensitive-topics classifier to identify the most common topics broadcasted in the dataset through the time. According to this, the most common topic is politics, then weapons, online/offline crime, health_shaming, rasims, social injustice.
According to the importance analysis of news reposting graph it was found that the most importrant according to PageRank node is glavmedia.

Lesson learned

Using external classifier for topic modeling brigns more information about how and what messages distributed through the time.
Berttopic is promising method for topic analysis, but it s hard to specify the semantic of cluster topic by using only keywords. The second challenging point is to process 8M messages.