A Corpus of Russian-Language Film Criticism on Telegram: Structural Topic Modeling and Sentiment Analysis

Aleksandr Shornikov
National Research University Higher School of Economics, Saint Petersburg
Saint Petersburg, Russia
avshornikov@edu.hse.ru

***Abstract*—In recent years, Telegram has become a major platform for media discourse in Russia. One notable topic shared by many Russian-language Telegram channels is content relating to films, the systematic analysis of which has not yet been undertaken. The study presents a corpus of Russian-language film criticism on Telegram: 128,613 publications (9,059,785 tokens, 143,833 unique lemmas) from 40 critics' channels published between 2017 and 2026. We analyze the corpus with three complementary methods. First, we describe posting dynamics and show that the annual activity peak occurs on the day following the Academy Awards ceremony in every year from 2019 to 2025. Second, we compare a lexicon-based approach (RuSentiLex) with a RuBERT classifier against 516 manually annotated posts: the lexicon over-detects positive tone (52.5% vs. 38%), the classifier over-detects neutrality (62.5% vs. 53%), and both over-detect negativity, with comparable post-level accuracy against the human annotation. Third, we apply Structural Topic Modeling with channel as a covariate, obtaining 15 topics that we merge into 10 macro-themes. Merging topic modeling results with those of sentiment analysis, we point out that lengthy analytical posts (158 words on average) and texts on Russian cinema and culture are the most negative, while publications sharing film-related materials are the most positive. Also, since 2021 the topic distribution appears more even. The developed corpus and code are publicly available.**

Keywords—film criticism, corpus linguistics, sentiment analysis, structural topic model, Telegram, Russian language

I. Introduction

Over the past few years, Telegram has become a prominent platform for Russian-language film criticism, hosting thousands of channels. In 2025, the platform's total reach exceeded 100 million users, while the number of channels surpassed 12 million [1]. Unlike traditional periodicals, Telegram generally gives authors greater freedom in choosing the topic, length, and style of their writing. This paper presents a corpus of film critics' Telegram channels and its examination using quantitative methods, sentiment analysis, and Structural Topic Modeling (STM).

Automatic sentiment analysis entails classifying texts by their polarity, thereby enabling the identification of an author's attitude toward entities mentioned in the text [2]. Sentiment analysis continues to develop actively [3], with major application areas including social and political research [4], psychology [5], fake-news detection [6], the study of consumer preferences based on product and service reviews [7, 8], and monitoring of economic trends [9]. One of the most common text genres for sentiment analysis is 'user reviews of products and services' [2]. In such texts, a single author typically expresses an opinion about a specific entity.

Many posts from the selected Telegram channels can be classified as film reviews; in this respect, our work partially aligns with numerous studies on the sentiment of movie reviews. For instance, a classic article on sentiment analysis of film reviews proposes focusing on the subjective parts of a text where opinion is explicitly expressed [10]. According to the authors, this increases the accuracy of the results, particularly when working with long reviews that mention neutral facts about the film. As a result of this work, around two thousand reviews were labelled as 'positive' or 'negative'. Another notable study compares machine-learning algorithms to determine the most accurate sentiment-analysis model, using the classic dataset of 50,000 reviews from IMDb [11]. This dataset was created in 2011 to demonstrate the superiority of word vector representations over the bag-of-words approach [12]. Among the studies focusing on Russian-language film reviews, it is worth mentioning a recent corpus of over 100,000 reviews collected from Kazakhstan's largest film website [13]. The study compares classical approaches to the tasks of sentiment classification and film rating prediction (based on TF-IDF with a linear SVM) with transformer-based models (mBERT, XLM-RoBERTa and RemBERT), concluding that modern transformer-based approaches outperform linear methods in terms of accuracy metrics.

It should be noted that Telegram has increasingly attracted scholarly attention. For example, [14] developed a methodology for constructing a corpus of Ukrainian-language texts from public Telegram channels. Other studies have addressed more specialized research questions. For instance, [15] analyzed 215 Dutch-speaking Telegram channels in which far-right, anti-government, and conspiracist narratives were circulated. Drawing on 371,951 messages published between 2017 and 2021, the authors combined network analysis with qualitative discourse analysis and identified interconnected themes, including conspiracy narratives related to COVID-19, QAnon, and the 'Great Reset'. Also, Willaert [16] conducted a computational analysis of 2,724,187 messages from 492 public Telegram channels across five thematic networks. They showed that channels may have distinct narrative profiles, ranging from coherent sequences of interconnected posts to more fragmented patterns of communication. Their findings suggest that Telegram publications cannot be treated as a homogeneous text genre, since their discourse functions may vary significantly.

It should also be stated that, to the best of our knowledge, no attempt has been made to create and process a corpus of film reviews on Telegram. Therefore, the aim of this study is to (1) compile a new corpus of Russian-language film criticism on Telegram; (2) analyze quantitative characteristics, including publication counts and lexical frequencies; (3) compare dictionary-based and machine-learning-based approaches to sentiment analysis on the collected data, using manual annotation to evaluate their performance; and (4) identify the main topics addressed in the critical texts included in the corpus. The compiled corpus is publicly available [17].

II. Materials and Methods

A. Corpus Collection and Preprocessing

The material for the study consists of posts from 40 Telegram channels devoted to cinema and published between 2017 and 2026. The corpus was compiled by exporting chat histories from Telegram. Although the total number of posts in the corpus reaches 135,539, empty publications lacking text were excluded from analysis; thus, 128,613 texts were processed. No filtering by text length was applied, nor were any temporal restrictions imposed on the publications. We did not restrict the sample to posts of a particular size, aiming instead for a more complete overview of Telegram publications about film. For example, in the thematic analysis of publications, post length emerges as a significant feature distinguishing entire groups of texts.

As for the selected sources, the study focused on 40 Telegram channels whose authors are film critics writing in Russian. The selection was made to ensure that the sample included channels run by professional film critics and film scholars that were active at the time of data collection (the final version of the corpus was assembled on 19.02.2026). For the purposes of this study, professionals were defined as those who write about cinema not only on Telegram, but also in newspapers, magazines and on websites.

Table I lists the channels analysed in this study, their number of subscribers, the author's name and their number of posts. The number of subscribers varies across channels: the study included both relatively small channels (450–1,000) and large ones (15–37,000); the average number of subscribers is 4,718. Among the channels with the larger audience in the corpus are such projects as 'Kronenberg nefiltrovannyy' (the leader in terms of subscribers and activity), 'Ya videl svechenie ekrana' and 'Kino s Larisoy Malyukovoy'. They are characterised by regular posts on current events in the film industry. There are also other examples of channels with a wide reach: for instance, Stanislav Zelvensky ('zelvenskiy') (10,423) and Mikhail Nefedov ('Cinema insolite') (16,913) write about their personal impressions without any obvious systematic approach.

Finally, the study included texts by 14 women and 26 men. The diversity of the sample is no coincidence: we endeavoured to collect texts of varying purpose, length and authorship, anticipating a broad range of topics and rich linguistic diversity.

TABLE I. Telegram channels selected for the study (sorted by number of subscribers)

Rank	Channel	Subscribers (19.02.2026)	Author(s)	Number of posts
1	Kronenberg nefiltrovannyy	37,199	Andrey Zagudaev	25,957
2	Ya videl svechenie ekrana	17,109	Maksim Ershov	3,825
3	Cinema insolite	16,913	Mikhail Nefedov	2,961
4	zelvenskiy	10,423	Stanislav Zelvenskiy	591
5	Kino s Larisoy Malyukovoy	9,580	Larisa Malyukova	6,973
6	Kabinet doktora K.	7,441	Vsevolod Korshunov	3,423
7	odin raz uvidet	6,796	Alisa Tayozhnaya	1,301
8	Plakhov	5,992	Andrey Plakhov	571
9	Usy Nikandera	5,788	Kirill Artamonov	10,116
10	Dnevnik Patersona	5,441	Vadim Bogdanov	7,102
11	Chacun son cinema	5,357	Alia Khuzyagaleeva	3,103
12	Dziga-apparatshchik	5,125	Egor Kozkin	3,649
13	Parni iz Chitalni	5,123	Evgeniy Tkachev	12,811
14	Ruki Bressona	4,128	Vladimir Rostovskiy	18,832
15	Oh boy	4,098	Tamara Khodova	2,431
16	Alisa v strane kino	3,911	Alisa Tayozhnaya	1,776
17	kinovoys	3,245	Betsi Isakova	709
18	Director's Card	3,008	Oleg Kovalev	2,081
19	Zapiski sotsiologa pro kino	2,718	Anastasia Starzhinskaya	5,056
20	Kinokritika eks-kurilshchika	2,512	Ivan Afanasev	637
21	Age Of Cinema	2,458	Elena Zarkhina	3,554
22	ya i Orson Uells	2,391	Pavel Pugachev	1,132
23	KINski	2,132	Marusia Buneeva	1,009
24	KinoGover	2,048	Vasily Goverdovskiy	1,327
25	The First Picture Show	1,876	Marat Shabaev	1,672
26	Sedmoe iskusstvo	1,780	Nikita Demchenko	614
27	Kinokrotik	1,708	Olga Belik	2,786
28	Chernichnye kinonochi	1,544	Nastya Gladilshchikova	1,011
29	Zvonite Gaydayu!	1,408	Nikolay Kornatskiy	374
30	Defekt Kuleshova	1,166	Sergey Kuleshov	1,494
31	plastikovyy Sinatra	1,148	Darya Tarasova	538
32	Vudinnormalen	1,095	Elena Fomina	1,259
33	Evgeniy Mayzel	1,064	Evgeniy Mayzel	189
34	Psikho i Ya	959	Karina Nazarova	404
35	Kulturnoe kino	877	Zhora Amian	1,261
36	Terminatarkovsky	783	Mikhail Morkin	519
37	Ptitsa Dodo	754	Pavel Voronkov	988
38	AF	694	Anton Fomochkin	1,132
39	uvidimsya v Montoke	465	Olesya Novikova	52
40	What she said	463	Masha Rakitina	319

We fully exported textual data from each Telegram channel in JSON format. The resulting corpus comprises 135,539 Russian-language publications, of which 128,613 non-empty; these contain 9,059,785 tokens. Posts were preprocessed by removing stop words, tokenizing texts with the NLTK package in Python, converting them to lowercase, cleaning punctuation, digits, and special symbols, and lemmatizing using the *pymorphy3* library. The corpus contained 143,833 unique lemmas. The cleaned and lemmatized texts were then processed by a model for sentiment detection (*blanchefort/rubert-base-cased-sentiment*). For each post, a sentiment label (*positive/negative/neutral*) was assigned.

B. Sentiment Analysis

Sentiment was determined using two approaches: a dictionary-based method and a machine-learning-based method. The first involves 'the use of emotion-annotated dictionaries that assign lexemes in the text (provided they match dictionary entries) … a value indicating class membership ('positive', 'negative', etc.)' [18]. The second relies on machine learning and consists in 'applying a model pretrained on a labeled corpus to new data, or training a custom classifier' [18]. In our case, the RuBERT-based model was trained on a corpus of 351,797 texts, which included posts from social media and reviews of goods and services. The model is designed to analyze the sentiment of short texts in Russian, which is one of the reasons it was selected for this study. Texts exceeding the model's input length were truncated before classification.

For the dictionary-based approach, we selected RuSentiLex [19, 20], a general-purpose sentiment lexicon for Russian. The final sentiment value for a post was calculated by computing the proportion of positive or negative words relative to the total number of sentiment-bearing words in the text; in cases of equal proportions, posts were classified as neutral.

Decisions on excluding or reclassifying lexemes were made by manually reviewing their concordance lines within the corpus. A word was reclassified only when its predominant contextual usage in the corpus was clearly different from its default RuSentiLex polarity. In order to validate this procedure, a second Russian-speaking annotator independently reviewed a random sample of 22 reclassified words, each accompanied by up to 18 corpus concordances. Raw agreement was 95.5% (21/22), which supported the original decisions. Thus, after excluding 47 words from the negative dictionary and 24 lexemes from the initial positive list (71 in total), the dictionaries contained 21,750 and 5,471 words, respectively.

It should be mentioned that GPT-based models were not used to analyse sentiment across the corpus, as their evaluation would require a separate experimental setup and was outside the scope of the comparison conducted in the study.

C. Topic Modeling

As regards topic modeling, one of the most common tasks in thematic modeling is identifying thematic trends in large text corpora [21]. This study employs the Structural Topic Model (STM) presented in [22]. Notably, the corpus was initially processed using a Latent Dirichlet Allocation (LDA) model, but the accuracy of the resulting topics was unsatisfactory. A key advantage of STM is its ability to incorporate metadata (covariates) when generating topics. In our modeling, channel was used as a covariate. This approach enables the identification of thematic trends across channels.

D. Ethics of Data Collection

The corpus was compiled exclusively from publicly accessible Telegram channels, which can be read without a subscription, in accordance with Telegram's Terms of Service and the permissions granted for access to the Telegram API. All 40 channels are run by professional film critics who publish under their own names as public figures. Furthermore, the authors' names are part of publicly available data and are treated as such throughout the study. The dataset is released for non-commercial research use. Images and videos attached to posts were not collected and were not included in the published dataset, which consists only of the text of the posts and the metadata (channel, author, date).

The corpus and the R analysis code are publicly available (see Section X).

III. Statistical Overview: Quantitative Trends in Posts

A. Dynamics of Posting Activity and Peak Dates

To begin, we will discuss the general quantitative trends within the corpus. Firstly, the number of texts published has increased significantly in recent years: in 2022, it doubled compared with 2021, and in 2025 it reached a peak of 31,730 posts. The lowest number of posts in the corpus was published in 2017 (1,061); thereafter, the number of posts grew year on year, with a slight dip in 2023 (20,029 posts compared with 20,210 in 2022). The corpus also includes posts from 2026, although at the time of data collection there were only 3,863 of them. The steady growth in activity on film channels may be consistent with the increased popularity of Telegram as a content platform in the first half of the 2020s.

When examining the combined trends in posts by year, a clear pattern emerges. The highest number of posts occurs on a specific day in virtually every year, with activity on the channels rising briefly on that date just once a year (Fig. 1).

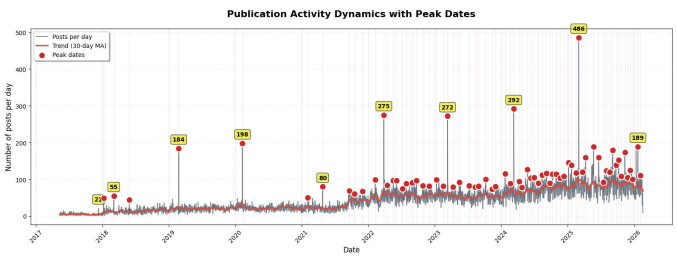


Fig. 1. Peak periods of posting activity

We link the sharp surge in the number of posts to one of the major events in the world of cinema – the Oscars. Due to the time difference, the daily peak in posts is recorded one day later than the date of the ceremony. Table II compares the

exact dates with the highest number of posts over time. As the corpus was compiled prior to the most recent Oscar ceremony to date (15.03.2026), the peak day in 2026 is not included in the table.

TABLE II. PEAK DATES FOR POSTING ACTIVITY

Year	Peak date	Number of posts	Date of the Oscars ceremony
2017	11.12.2017	21	26.02.2017
2018	08.01.2018	49	04.03.2018
2019	25.02.2019	184	24.02.2019
2020	10.02.2020	198	09.02.2020
2021	26.04.2021	80	25.04.2021
2022	28.03.2022	275	27.03.2022
2023	13.03.2023	272	12.03.2023
2024	11.03.2024	292	10.03.2024
2025	03.03.2025	486	02.03.2025

It should be noted that a manual review of posts confirmed that on the specified dates between 2019 and 2025, the number of posts dedicated to the 'Oscars' was extremely high, which cements the event's status as one of the most significant in the Russian-speaking film criticism community.

B. Frequency Dictionaries

After removing stop words, extraneous characters and links, and following lemmatisation of the texts in the corpus, frequency dictionaries of unigrams and bigrams were created. Table III shows the fifteen most frequent words across all publications.

TABLE III. THE 15 MOST FREQUENT LEMMAS IN THE CORPUS

Rank	Word	Number of occurrences
1	'film' (film)	98822
2	'kino' (cinema)	36192
3	'rezhissyor' (director)	25291
4	'serial' (series)	20967
5	'istoriya' (story)	20870
6	'kartina' (picture)	17447
7	'glavniy' (main)	16997
8	'rol' (role)	16207
9	'zhizn' (life)	15700
10	'smotret' (to watch)	13992
11	'mir' (world)	12083
12	'sezon' (season)	11405
13	'festival' (festival)	11301
14	'rabota' (work)	11076
15	'posmotret' (to watch)	10709

The most frequent lexemes in the corpus are directly linked to the themes of the channels and form an integral part of the vocabulary of anyone writing about cinema. Thus, 'film' is by far the most frequent word in the corpus, as are 'kino' and 'serial'. In terms of meaning, 'kartina' and 'rabota' are also closely related to these; these words most often appear in the sense of 'work of art'. Notably, film critics focus more on films than on TV series, judging by the lower (yet still high) frequency of the latter term. Interestingly, the abstract nouns 'zhizn' and 'mir', unlike the other words, refer to the intrinsic content of the works, while the verbs 'posmotret' and 'smotret'—the only verbs in the list—naturally govern the most frequent nouns in the corpus (e.g. 'posmotret film').

Table IV lists the fifteen most frequent bigrams. It is worth noting that most high-frequency bigrams refer to fundamental aspects of films and the filmmaking process (roles, release, a festival), while the collocations 'khoroshiy film' and 'lyubimiy film' indicate the frequent direct expression of a subjective attitude towards a film.

TABLE IV. THE 15 MOST FREQUENT BIGRAMS IN THE CORPUS

Rank	Bigram	Number of occurrences
1	'glavnaya rol'' (lead role)	5,718
2	'noviy film' (new film)	3,876
3	'khoroshiy film' (good film)	3,062
4	'vtoroy sezon' (second season)	2,549
5	'glavniy geroy' (main character)	2,214
6	'rossiyskiy prokat' (Russian release)	1,775
7	'film predstavit'' (to present a film)	1,688
8	'den' rozhdenie' (birthday)	1,603
9	'dokumentalniy film' (documentary film)	1,439
10	'lyubimiy film' (favorite film)	1,227
11	'tretiy sezon' (third season)	1,201
12	'pervaya seriya' (first episode)	1,166
13	'perviy sezon' (first season)	1,132
14	'vtoroy plan' (background)	1,122
15	'kannskiy kinofestival'' (Cannes Festival)	1,090

IV. SENTIMENT ANALYSIS OF POSTS

A. Comparison of Sentiment Classification Methods

The two automatic methods used in the article were evaluated by a human annotation of 516 posts, which were randomly sampled from the corpus and labeled by a single annotator (the author) based on the overall tone conveyed by the post as a whole. To assess the reliability of these labels, a second native-speaking annotator independently labeled a random subsample of 30 posts, without access to the original annotations. The annotators' decisions were largely similar: raw agreement was 86.7%, with Cohen's Kappa = 0.77. We have to admit that given the small subsample, this estimate might not indicate conclusive agreement.

Table V compares the sentiment distributions obtained from the lexicon-based method, the RuBERT classifier, and 516 manually annotated posts. The model classified the majority of publications as texts with a neutral tone (63%), and there were more negatively toned texts (21%) than positive ones (16%); the lexicon method instead finds a majority of positive posts (52.5%), fewer neutral ones (30.6%), and a similar share of negative ones (16.5%). Thus, the lexicon method yields about 36 percentage points more positive posts, 32 points fewer neutral posts, and 5 points fewer negative posts than the classifier. As for the manual annotation, it classified 53% of the 516 sampled posts as neutral, 38% as positive, and only 9% as negative.

TABLE V. SENTIMENT DISTRIBUTION: LEXICON VS. RUBERT VS. MANUAL ANNOTATION

Class	Lexicon (n=128,613)	RuBERT (n=128,613)	Manual (n=516)
Positive	52.5%	16.3%	38%
Neutral	30.6%	62.5%	53%
Negative	16.5%	21.3%	9%

Table VI reports post-level evaluation metrics against the manual annotation of 516 posts. Confidence intervals were obtained via bootstrap resampling (5,000 iterations). Although RuBERT achieves marginally higher accuracy and macro-F1, a McNemar's test comparing the two methods' per-post correctness against the manual labels found no statistically significant difference ($p = 0.39$), indicating that the two methods perform comparably overall despite disagreeing on which individual posts they classify correctly.

TABLE VI. POST-LEVEL EVALUATION AGAINST MANUAL ANNOTATION (N=516)

Method	Accuracy	Macro-F1	95% CI (Accuracy)	95% CI (Macro-F1)
Lexicon (RuSentiLex)	0.498	0.426	[0.453, 0.541]	[0.382, 0.468]
RuBERT classifier	0.527	0.440	[0.484, 0.570]	[0.395, 0.486]

A possible reason automatic methods over-detect negativity is that automatic annotation takes into account negative vocabulary in the titles of films and TV series, and also fails to capture the author's approving or critical tone in lengthy film reviews. In this respect, manual annotation is more similar to the dictionary-based method, which yields a noticeable bias towards positive posts. However, even here dictionaries are not ideal, as a similar bias towards positive vocabulary is possible due to titles, detailed descriptions and ironic remarks. The latter can often complicate the process of determining a text's tone, as mentioned in [23].

Table VII shows the confusion matrices. The lexicon recovers most positive posts (recall 68.7%) but labels 119 of the 271 neutral posts as positive, whereas RuBERT assigns 101 of the 198 positive posts to the neutral class (recall on positive 28.3%). Both methods struggle with the rare negative class (recall 27.7% and 44.7%; precision 14.9% and 18.4%): most posts predicted as negative are in fact positive or neutral.

TABLE VII. CONFUSION MATRICES AND PER-CLASS PRECISION (N=516)

Manual	Lex. Pos	Lex. Neu	Lex. Neg	RuBERT Pos	RuBERT Neu	RuBERT Neg
Positive (n=198)	136	32	30	56	101	41
Neutral (n=271)	119	108	44	24	195	52
Negative (n=47)	23	11	13	6	20	21
Precision, %	48.9	71.5	14.9	65.1	61.7	18.4

B. Most Frequent Positive and Negative Words

The most frequent positive lemmas are 'novyj' (new, 30,080) and 'khoroshij' (good, 28,098), followed by clearly evaluative adjectives such as 'lyubimyj' (favorite, 6,567), 'velikij' (great, 5,268), and 'interesnyj' (interesting, 4,807); 'nastoyashchij' (real/true, 4,825) and 'vazhnyj' (important, 4,602) are more context-dependent and can occur in negative or neutral phrases (e.g., 'nastoyashchij proval', 'vazhnoe zamechanie'). The most frequent negative lemmas revolve around strongly negative concepts—'vojna' (war, 4,782), 'smert'' (death, 4,206), 'uzhas' (horror, 2,123), 'ubijstvo' (murder, 1,986), 'ubijtsa' (murderer, 1,818), and 'strakh' (fear, 1,577) — which possibly reflects critics' frequent engagement with genre cinema (crime, war, and horror films); among evaluative adjectives, 'plokhoj' (bad, 3,934) leads, while 'prostoj' (simple, 3,012), 'slozhnyj' (complex, 2,110), and 'strashnyj' (scary, 1,979) do not always carry a negative connotation.

V. RESULTS OF THEMATIC MODELING. IDENTIFICATION OF MACRO-THEMES

Topic modeling was performed in R using the stm package. Documents were processed with textProcessor(), with stopword removal and lemmatization disabled, as these steps had already been applied during preprocessing; the vocabulary was trimmed using prepDocuments(). The model was estimated with $K = 15$ topics, and channel identity as the only covariate. Year of publication was not included in the final model because of technical problems during estimation; temporal trends were instead computed later by averaging per-post topic probabilities by year (Section VIII). Each post was assigned to its single dominant topic (the one with the highest estimated probability) for the theme- and channel-level analyses. The random seed was fixed (123); semantic coherence and exclusivity averaged −111.64 and 9.52 across the 15 topics. As these metrics have no universal benchmark, they are reported descriptively. It is also important to note that K was chosen for interpretability, following [24]; a systematic comparison across alternative values of K would further strengthen the model-selection procedure.

Subsequently, the 15 topics were consolidated into 10 macro-themes based on overlap in their thematic content and top FREX words. Topics 2, 5, and 8 were combined into 'Russian Cinema, Culture, Film Production' because all three centre on Russian cinema and culture ('teatr', 'rossiya', 'rossiyskiy', 'moskva'). Topics 10 and 12 were combined into 'Analyses and Interpretations' because both deal with analytical long-form texts; Topic 10 is characterised by psychological vocabulary ('zhizn'', 'istoriya', 'syuzhet', 'travma'), while Topic 12 focuses on philosophical and existential themes ('mir', 'voyna', 'smert'', 'real'nost''). Topics 11 and 14 were combined into 'Personalities and New Releases' because both focus on film personalities. Lastly, Topics 13 and 15 were combined into 'Evaluative Reviews' because both centre on evaluative and review-based discourse; Topic 13 captures informal evaluative reactions ('ponravit'sya', 'krutoy'), while Topic 15 focuses on film craft ('stil'', 'kadr'). The remaining six topics (1, 3, 4, 6, 7, 9) were retained as separate macro-themes.

For a substantive interpretation of the themes, four key indicators were analyzed:

1) *HighestProb*: assigned to words with the highest probability of occurring within a topic.
2) *FREX*: assigned to words combining high frequency within a topic and exclusivity relative to other topics; the importance of this indicator is discussed in [25].
3) *Lift*: a measure of how many times greater the probability of words occurring in the topic is compared to the corpus.

4) *Score*: an integrated measure of significance.

As an example, let us look at the values in Table VIII, presenting these metrics for one of the topics identified by the model:

TABLE VIII. STM METRICS FOR 'ANALYSES AND INTERPRETATIONS'

Metric	Words in the Topic
Highest Prob	*'zhizn'', 'istoriya', 'drug', 'zritel'', 'syuzhet', 'pytat'sya', 'rebyonok', 'sem'ya', 'kazat'sya', 'otnoshenie'*
FREX	*'boleznennyj', 'emotsional'nyj', 'travma', 'realistichnyj', 'zhiznennyj', 'slabost'', 'nasilie', 'udavat'sya', 'perezhivanie', 'povorot'*
Lift	*'albanets', 'gubitel'nost'', 'zaukhara', 'zatsiklivat'', 'itp', 'kerrigan', 'cahill', 'miza', 'narechie', 'obvalivat'sya'*
Score	*'geroinya', 'problema', 'zhizn'', 'sem'ya', 'personazh', 'zritel'', 'pytat'sya', 'chuvstvo', 'syuzhet', 'syuzhetnyj'*

This theme centres on analyses and reviews, which are characterised by a comparatively large volume and abstract vocabulary. Unlike other themes, general concepts such as 'zhizn'' (life), 'drug' (friend) and 'sem'ya' (family) feature frequently; interestingly, these words appear in the frequency list alongside vocabulary used to discuss films ('syuzhet' (plot), 'istoriya' (story)). The eighth theme is also characterised by words with psychological connotations ('boleznennyj', (painful), 'emotsional'nyj' (emotional), 'perezhivanie' (distress)), which are frequently used to interpret works of art.

Before analysing the lexical content of the topics, it is necessary to highlight a weakness in the results of thematic modeling, namely 'noisy' output, which manifests itself in lexemes that make it difficult to interpret the topics. As the authors of a study on the thematic features of discussions on 'LiveJournal'—who applied the Latent Dirichlet Allocation—note, information noise may arise due to the stochastic 'bag-of-words' approach [26]. With this method, which also forms the basis of STM, word order is not taken into account directly, and the text is represented as a set of words with their frequencies indicated; as a result, the model may produce units that are frequent but of little significance to the content of the topic. Furthermore, incorrect lexemes may appear in the results due to incorrect lemmatisation of first names, surnames, film titles, as well as words artificially created by the authors. For instance, in our STM results, the lexeme 'layvlya' is identified by the model as exclusive to the fifth macro-theme. Further verification revealed that it derives from the surname of the actress Blake Lively, who is mentioned 48 times within this topic (with multiple mentions in several posts about her). Nevertheless, there are grounds to believe that the frequency of individual names does not determine the thematic focus of the texts in the formed groups. The exception is macro-theme 9 ('Personalities and New Releases'), the posts within which are thematically similar, partly due to frequent mentions of figures from the film industry.

Table IX lists the ten macro-themes obtained after merging semantically overlapping topics. The first theme included texts devoted to international film festivals and new cinema releases (e.g. High Prob: 'festival', 'kanny', 'venetsiya', 'prem'era'; Score: 'berlinale', 'zhuri', 'konkurs').

The second macro-theme comprises texts on Russian cinema, culture and the filmmaking process (High Prob: 'Rossiya', 'russkiy'; FREX: 'minkul't', 'kinotavr'; Score: 'moskva', 'teatr').

All the texts in the third group are united by their presence in the media landscape, interaction with the audience and the dissemination of material about cinema that extends beyond the social network (Score: 'podkast', 'retsenziya', 'tekst', 'vypusk').

Theme 4 brought together texts promoting film screenings, as well as any invitations to the audience (Score: 'spetspokaz', 'bilet', 'registratsiya').

Topic 5 consists mainly of posts about English-language cinema. This includes, for example, texts about the Oscars, which is the premier film event in the USA. (Highest Prob: 'Oskar', 'amerikanskij'; FREX: 'oskarovskij', 'statuetka').

The sixth theme brings together texts on mainstream cinema, that is, commercial films, blockbusters or popular arthouse projects. A unifying feature is the prominence given in the texts to genre films: horror films, action films and comic-book adaptations (Score: 'franshiza', 'horror', 'boevik').

Texts about TV series have also been grouped into a separate theme – the seventh (Score: 'epizod', 'sezon', 'serial'). This theme also includes news items about upcoming projects in the film industry (Highest Prob: 'prem'era', 'trejler').

While the eighth theme centres on Analyses and lengthy reviews, Topic 9 comprises posts focusing on actors, directors and other figures from the world of cinema, as well as new film releases. Among the most frequent and unusual words are the surnames of famous directors and actors ('Tarantino', 'Anderson', 'Robbie', 'Aster').

Finally, the last topic comprises the largest body of texts. First and foremost, this theme consists of film reviews and evaluative comments on films. The posts share the blog-style format of film reviews: relatively short length, a clear authorial stance, and analysis of individual scenes and artistic choices. In particular, this is emphasised by a core set of film terms that are significant within the topics.

TABLE IX. THE FINAL LIST OF 10 MACRO-THEMES

Topic Number	Topic Title	Number of Posts
1	New Releases and Festivals	6576
2	Russian Cinema, Culture, Film Production	17666
3	Materials on Cinema	13722
4	Advertisements and Promotions	6734
5	English-Language Films	10757
6	Popular Films	18042
7	TV Series and Premieres. Film News	6913
8	Analyses and Interpretations	15899
9	Personalities and New Releases	11813
10	Evaluative Reviews	20491

VI. Analysis of the tone within macro-themes

A. Positive Sentiment Across Themes

The overall distribution of sentiment within the themes turned out, as expected, to differ between the two methods. With regard to positive texts, the lexical tagging identified topics 3 ('Materials on cinema') and 4 ('Advertisements and promotions') as the groups with the highest concentration of positive posts (more than 65%). In contrast, topics 6 ('Popular films'), 7 ('TV series and premieres. Film news') and 10 ('Evaluative Reviews') had the lowest proportion of positive publications (around 45%). The picture was somewhat different for the automated method. The fourth topic ('Advertisements and promotions') is far from leading in terms of positive texts (a discrepancy of –54.9 p. p.), while for the fifth topic ('English-language films'), the automated method yielded 44.9 p. p. fewer positive posts. Furthermore, 'Evaluative Reviews' (Topic 10) were rated less positively by the model than by the dictionaries (–27 p. p.). At the same time, methodological similarity was found in the topic 'Materials on cinema', which in both cases significantly outperforms almost all other topics in terms of the proportion of positive posts.

B. Neutral Sentiment Across Themes

Across all ten macro-themes, automatic tagging yields a higher proportion of neutral assessments than the dictionary-based method (on average +36.5 p. p.). The first theme ('New Releases and Festivals') stands out as one of the leaders in terms of the proportion of neutral texts (over 70%). The content of Topic 1 supports the model's findings: texts in this group often take the form of reports or news bulletins covering films not yet released in cinemas. We can assume that, in such cases, the authors of the texts strive for greater objectivity and a more neutral tone. Furthermore, we note that the model identified a significant degree of neutrality in topics 4–7 (over 70%), whereas the lexical method revealed no more than 40% neutral posts in these groups. While neutrality is likely to be expected for the fourth topic ('Advertisements and promotions'), such a result is difficult to interpret in the other groups ('English-language films', 'Popular films', 'TV series and premieres. Film news'). Finally, the model is least likely to produce neutral assessments within the tenth theme ('Evaluative Reviews'); this is predictable, given that emotionality and evaluativeness were among the criteria used to identify the theme.

C. Negative Sentiment Across Themes

As for negative tone, the eighth theme ('Analyses and Interpretations') emerged as the leader for both methods (around 30% of all publications in the topic). For the lexical method, this can be explained by a higher concentration of negative vocabulary due to the longest average text length (157.9 words per post). In this case, a comparison of the methods reinforces the conclusions regarding the general characteristics of the texts in the group: as the posts here are stylistically and functionally closer to reviews, their critical tone is heightened. Thus, the eighth theme should be regarded as a collection of posts that most closely resemble what might be termed 'classical' film criticism. The heightened negativity is also confirmed by the distinctive vocabulary associated with the theme – 'boleznennyj' (painful), 'nasilie' (violence), 'slabost'' (weakness) and 'emotsional'naya travma' (emotional turmoil) are words that are exclusive to this theme and, at the same time, negative in tone.

The next most negative topic in terms of tone was the group of texts on Russian cinema and culture. Interestingly, for both methods, this topic ranked among the top three most negative. This may reflect a more critical treatment of the Russian film industry and culture among the critics in the corpus. The high proportion of texts with a negative tone may also be linked to socio-political difficulties, censorship restrictions (FREX: 'tsenzura' (censorship)) and the crisis in the contemporary Russian film distribution sector.

VII. Length of posts within macro-themes

Average post length also varies substantially by theme. Theme 8 ('Analyses and Interpretations') has by far the highest average length (158 words per post), consistent with its discursive character (roughly 16,000 posts in the corpus fall into this category). Texts on Russian culture and cinema took second place in terms of average length – 93.9 words per post. Theme 4 ('Advertisements and Promotions') ranks third (78.4 words on average): a typical screening announcement needs to specify the venue, date, time, guest speaker, and registration details, which naturally lengthens the post.

It is worth noting that most posts in the corpus consist of fewer than 60 words on average; the very format of a Telegram post lends itself to brevity, so long texts – even amongst critics – should be regarded as something of an anomaly. Finally, the topic 'TV series and premieres. Film news' has the lowest average post length (41.8 words). It mainly contains concise news items about upcoming projects, current events and premieres. The news-oriented nature of the category thus influences the statistically short format of the posts it includes.

VIII. Trends in the popularity of macro-topics (2017–2026)

As already mentioned, the advantage of the STM model we have chosen lies in its consideration of metadata; therefore, in the final section of our review of the results, we will focus on channel identity as a prevalence covariate. Also, year of publication was analysed: topic probabilities were averaged by year.

Turning first to temporal trends, it is worth noting that analysing the evolution of topics in text corpora is of broad interest and allows us to track trends in fields where the flow of information is constantly updated [21]. In our case, such an analysis is particularly relevant, as the discourse surrounding contemporary cinema is characterised by its fluid nature.

For each publication, the probabilities of it being classified under each of the 10 macro-themes were recorded; subsequently, all posts were grouped by year of publication, and for each year the arithmetic mean of the probabilities for each theme was calculated. Thus, the values obtained reflect

the average ‘popularity’ of a theme within the corpus for the corresponding year.

The graph (Fig. 2) shows that the range of topics covered by film critics’ channels has, on the whole, become more diverse. While from 2017 to 2019 the main topics were ‘Russian cinema, culture, film production’ and ‘Evaluative Reviews’, there was subsequently an increase in the probability of posts relating to popular films and TV series (2019–2021). Since 2020, there has also been a gradual increase in other topics, alongside a decline in the prominence of review articles and topics relating to Russian cinema. As is well known, 2020 saw a sharp rise in activity on Telegram during the COVID-19 pandemic; this may explain the broadening of the range of topics that film critics began to write about on their channels during this period. For instance, since 2020, the prominence of the topic ‘Advertisements and promotions’ has been steadily rising, as has the likelihood of encountering analytical texts (topic 8, ‘Analyses and Interpretations’).

Interestingly, over the last few years, all 10 macro-topics have been distributed more evenly across the corpus than before; each occurs with a probability of between approximately 3% and 12%. This suggests that the texts are diverse in content, and there is no strong bias towards any particular topic, as was the case prior to 2021. Finally, the most likely posts in terms of subject matter in recent years remain those about Russian cinema and culture, and publications sharing materials on film (probability > 10%).

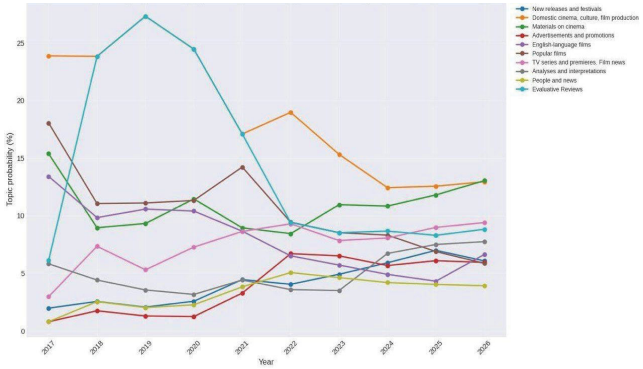


Fig. 2. Trends in the probability of topics on Telegram channels about cinema (2017–2026)

IX. Distribution of Macro-Themes Across Film Critics’ Telegram Channels

The study examined 40 Telegram channels of varying formats and levels of activity; consequently, it is difficult to comment on the thematic distribution within each of them. However, the thematic activity within each channel was analyzed. Table X shows the percentage distribution of each topic within the top five most active channels in the corpus.

First, we shall note the thematic characteristics of the channels with the highest number of posts. Posts on the ‘Kronenberg nefil’trovannyj’ channel (25,957) account for a significant proportion of topics 5, 6, 7 and 9; these are united by a focus on current films and events, as well as on popular cinema. Consequently, posts on this channel are more often short and news-oriented, a fact confirmed by the low proportion of the eighth topic (in-depth Analyses).

A thematically similar project is the ‘Ruki Bressona’ channel, where texts about new releases are also common; however, it contains more promotional posts and event announcements (around 40% of the total content of theme 4), and other themes—such as analytical pieces—are more prominent. The ‘Parni iz Chital’ni’ channel focuses more on popular cinema (theme 6), Russian culture (theme 2) and film reviews (theme 10), while ‘Usy Nikandera’ is characterised by its emphasis on critical reviews (around a fifth of theme 10 is found on this channel).

Having cited a few examples, we note that meaningful conclusions can be drawn for those channels whose activity is sufficiently high. If, however, the number of posts is significantly below average, the author’s contribution to a particular topic is barely noticeable. Therefore, it is more difficult to compile a thematic profile for the channels of certain well-known critics in the corpus (e.g. Andrei Plakhov, Stanislav Zelvensky).

TABLE X. Distribution of posts across the five most active channels for each topic (values show each channel’s share of the topic; column totals across all 40 channels equal 100%). KN: Kronenberg nefil’trovannyj; RB: Ruki Bressona; PiC: Parni iz chital’ni; UN: Usy Nikandera; DP: Dnevnik Patersona.

#	KN	RB	PiC	UN	DP
1	12	10.9	4.5	1.7	11.4
2	7.3	17.4	12	0.9	7.5
3	12.4	5.9	9.4	5	3.8
4	4.1	36.5	1.3	6.2	2.6
5	55.3	8.1	4.9	1.9	3.7
6	53.4	9.9	12.7	2.7	4.7
7	37.1	27	3.7	1.5	3.8
8	4.1	15.5	5.9	3.4	3.9
9	24.9	10.9	4.3	1.7	7.2
10	8.3	13.3	12.6	20.6	6.1

X. Conclusion

The paper presents a corpus of Russian-language film criticism on Telegram, and its computational analysis. In particular, the dynamics of posts between 2017 and 2026 were calculated and channel activity was determined. The study revealed that the regular annual spikes in activity on Telegram channels occur on the day following the Academy Awards ceremony every year from 2019 to 2025.

Sentiment analysis involved the use of two methods: machine-based and lexical; in this way, it was possible to compare the results of each and obtain a more comprehensive picture, while assessing the shortcomings of both approaches. While the machine-based model underestimated the positivity of many posts, the lexical method proved to be far more ‘optimistic’. Manual annotation of a small set of random texts helped us conclude that neither method is by any means perfect at determining the tone of film reviews. It should also be noted that an important part of dictionary-based sentiment analysis involved preparing dictionaries to eliminate lexical ambiguities; as a result, lexical items that initially appeared in both the positive and negative dictionaries were filtered out.

A significant part of the study involved thematic modeling using STM. The model made it possible not only to identify 15 main themes in the corpus, which were reduced to 10 macro-themes, but also to analyze them in conjunction with channel as a covariate, as well as identify thematic patterns across several years.

Consequently, the 10 macro-themes, whose lexical composition was examined in this study, represented a broad range of content within the film criticism corpus. Evaluative reviews made the widest collection of posts (20,491). Publications presenting film-related content proved to be among the most positive according to both sentiment analysis methods, while long analytical posts contained the largest proportion of negative texts. Another topic with a higher negativity level was 'Russian Cinema, Culture, Film Production'. Furthermore, the study revealed that a more balanced distribution of themes has been observed in the corpus in recent years. Among other things, the channels' specialisation across various topics was examined, and it appears that, in future, it would be useful to consider this data in terms of the tone of posts on individual Telegram channels. Furthermore, a more detailed analysis could be carried out by grouping data into subsets to enable a more focused study (for example, to characterise the posts of a specific channel in a particular year). It would also be extremely useful to expand the research sample by including numerous other film-related channels not covered in this study, as well as film reviews from other sources (not Telegram), and to compare the results using the same indicators.

To conclude, this study has taken the first important steps towards a systematic description of Telegram film criticism. The combination of several computer-based methods has helped to examine the corpus from various angles and provide insight into the main trends across a substantial number of publications. Finally, the corpus is available on Hugging Face [17], and the R code for the analysis is available on GitHub [27]; the repository includes a README with reproduction steps and the list of required R packages.

REFERENCES

[1] TGStat, "Official Telegram statistics," Telegram, Sep. 26, 2025. Web: https://t.me/TGStat/491.

[2] E. I. Bolshakova, K. V. Vorontsov, N. E. Efremova, E. S. Klyshinsky, N. V. Lukashevich, and A. S. Sapin, *Automatic Natural-Language Text Processing and Data Analysis (in Russian).* Moscow: HSE Publishing House, 2017.

[3] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis – a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[4] S. Volkova, G. Coppersmith, and B. Van Durme, "Inferring user political preferences from streaming communications," in *Proc. ACL (1)*, 2014, pp. 186–196.

[5] X. Wang et al., "A depression detection model based on sentiment analysis in micro-blog social network," in *Proc. PAKDD Workshops.* Berlin, Heidelberg: Springer, 2013, pp. 201–213.

[6] S. Volkova and E. Bell, "Account deletion prediction on RuNet: a case study of suspicious Twitter accounts active during the Russian-Ukrainian crisis," in *Proc. NAACL-HLT*, 2016, pp. 1–6.

[7] A. Shah, "Sentiment analysis of product reviews using supervised learning," *Reliability: Theory & Applications*, vol. 16, no. 1, pp. 243–253, 2021.

[8] M. I. Hossain, M. Rahman, T. Ahmed, and A. Z. M. T. Islam, "Forecast the rating of online products from customer text review based on machine learning algorithms," in *Proc. Int. Conf. on Information and Communication Technology for Sustainable Development (ICICT4SD)*, IEEE, 2021, pp. 6–10.

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[10] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. 42nd Annual Meeting of the ACL*, 2004, pp. 271–278.

[11] M. A. Jassim, D. H. Abd, and M. N. Omri, "Machine learning-based new approach to films review," *Social Network Analysis and Mining*, vol. 13, Art. no. 40, 2023.

[12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annual Meeting of the ACL: Human Language Technologies*, Portland, OR, USA, 2011, pp. 142–150.

[13] R. Yeshpanov, "100,000+ movie reviews from Kazakhstan: Russian, Kazakh, and code-switched texts," arXiv:2605.08600, 2026.

[14] M. Malysheva, "Corpus of texts from public channels of the Telegram messenger: Creation and practical application," *Odesa National University Herald.* Series: Philology, vol. 29, no. 2(30), pp. 47–56, 2024, doi: 10.18524/2307-8332.2024.2(30).320406.

[15] T. Willaert, S. Peeters, J. Seijbel, and N. Van Raemdonck, "Disinformation networks: A quali-quantitative investigation of antagonistic Dutch-speaking Telegram channels," *First Monday*, vol. 27, no. 5, 2022, doi: 10.5210/fm.v27i5.12533.

[16] T. Willaert, "A computational analysis of Telegram's narrative affordances," *PLOS ONE*, vol. 18, no. 11, Art. no. e0293508, 2023, doi: 10.1371/journal.pone.0293508.

[17] A. Shornikov, "Telegram-Film-Criticism," Hugging Face Datasets, 2026. Web: https://huggingface.co/datasets/Alex-Shornikov/Telegram-Film-Criticism.

[18] M. A. Kirina and L. D. Telnina, *Automatic Assessment of Students' Impressions Using Sentiment Analysis Methods (in Russian).* St. Petersburg: HSE University, 2026.

[19] RuSentiLex: a sentiment lexicon of the Russian language, version 2017. Web: http://www.labinform.ru/pub/rusentilex/index.htm.

[20] N. Loukachevitch and A. Levchik, "Creating a general Russian sentiment lexicon," in *Proc. Language Resources and Evaluation Conf. (LREC)*, 2016.

[21] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, X. Tong, and H. Qu, "TextFlow: towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, 2011, doi: 10.1109/TVCG.2011.239.

[22] M. A. Kirina, "A comparison of topic models based on LDA, STM and NMF for qualitative studies of Russian short prose," *Vestnik NSU. Series: Linguistics and Intercultural Communication*, vol. 20, no. 2, pp. 93–109, 2022, doi: 10.25205/1818-7935-2022-20-2-93-109.

[23] E. G. Borisova and Yu. K. Pirogova, "Modeling non-trivial conditions of message understanding (the case of irony)," in *Proc. Annual Int. Conf. "Dialogue"*, 2013, pp. 148–162.

[24] A. Sciandra, M. Trevisani, and A. Tuzzi, "Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment," in *Proc. Statistics and Data Science Conference*, Pavia University Press, 2023, pp. 61–66.

[25] J. M. Bischof and E. M. Airoldi, "Summarizing topical content with word frequency and exclusivity," in *Proc. 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 201–208.

[26] N. A. Almaev and O. V. Murasheva, "Thematic analysis of discussions using Latent Dirichlet Allocation" (in Russian), *Social and Economic Psychology*, vol. 7, no. 1(25), pp. 47–68, 2022.

[27] A. Shornikov, "telegram-film-criticism: R code for the analysis," GitHub, 2026. Web: https://github.com/avshornikov/telegram-film-criticism.