Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media

Chan Young Park*
Carnegie Mellon University
chanyoun@cs.cmu.edu

Julia Mendelsohn*
University of Michigan
juliame@umich.edu

Anjalie Field*
Stanford University
anjalief@stanford.edu

Yulia Tsvetkov
University of Washington
yuliats@cs.washington.edu

Abstract

NLP research on public opinion manipulation campaigns has primarily focused on detecting overt strategies such as fake news and disinformation. However, information manipulation in the ongoing Russia-Ukraine war exemplifies how governments and media also employ more nuanced strategies. We release a new dataset, VoynaSlov, containing 38M+ posts from Russian media outlets on Twitter and VKontakte, as well as public activity and responses, immediately preceding and during the 2022 Russia-Ukraine war. We apply standard and recently-developed NLP models on VoynaSlov to examine agenda setting, framing, and priming, several strategies underlying information manipulation, and reveal variation across media outlet control, social media platform, and time. Our examination of these media effects and extensive discussion of current approaches' limitations encourage further development of NLP models for understanding information manipulation in emerging crises, as well as other real-world and interdisciplinary tasks.

1 Introduction

On February 24, 2022, Russia began an open military invasion of Ukraine. At the time of writing, this ongoing conflict has killed thousands of people and displaced millions.[1] The conflict has also manifested in ongoing *information warfare*, as Russian, Ukrainian, and ally forces attempt to shape online narratives of the war.[2] Even before these events, Russian-backed entities have aimed to influence opinions outside (Arif et al., 2018; Starbird et al., 2019) and inside Russia (Field et al., 2018a; Golovchenko et al., 2018; Rozenas and Stukal, 2019). More broadly, researchers consider the manipulation of public opinion over social media as "a critical threat to democracy" and have identified computational propaganda in over 80 countries (Bradshaw et al., 2021).

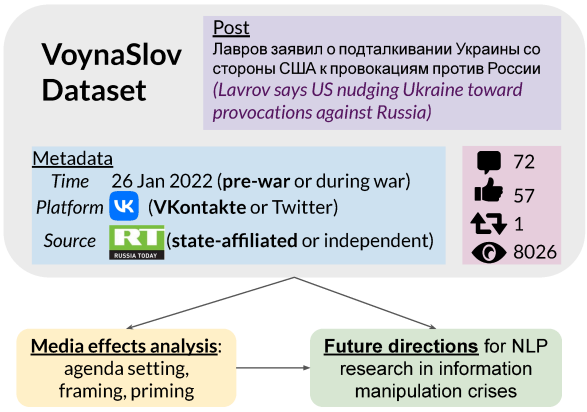

Figure 1: We create VoynaSlov, a dataset of Russian news organizations' social media posts and public responses. Above is an example VK post in VoynaSlov from the state-affiliated outlet *Russia Today*, along with metadata and engagement metrics. We use VoynaSlov to analyze media effects of agenda setting, framing, and priming. Both the dataset and our analyses contribute to our discussion of future directions for NLP research.

Because these campaigns often rely on text-based news and social media content, NLP can be a valuable tool in combating them. In this work, we examine the usability of NLP approaches for combating information manipulation campaigns through the release of an in-progress data set, VoynaSlov, focused on the 2022 Russia-Ukraine war (§2). This dataset itself addresses one challenge for research in this space: the need for real-world data. In contrast to contemporaneous Twitter data sets (Haq et al., 2022; Chen and Ferrara, 2022), our corpus is explicitly designed to capture Russian-government-backed information manipulation; we collect posts by state-affiliated and independent Russian media outlets and reactions to them on Twitter, which is more dominant in Europe and the U.S., and VKontakte (VK), one of the most widely-used social media platforms in Russia (Makhortykh

*Equal contribution
[1]UNHCR, CNN
[2]Atlantic

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5238 - 5264

and Sydorova, 2017).[3] Unlike approaches that crowd-source data to study a specific task (Thorne et al., 2018a; Pérez-Rosas et al., 2018), we derive research tasks directly from real-world data.

The dominant NLP research paradigm in information campaigns has focused on automated fact-checking or propaganda and fake news detection (Thorne et al., 2018a; Oshikawa et al., 2020; Martino et al., 2020; Zhou and Zafarani, 2020; Guo et al., 2022). However, this work typically involves supervised approaches and pre-annotated data, which are not available in emerging situations, and only captures one extreme form of media manipulation. Instead, we draw on a common paradigm of information manipulation from communications research and examine signs of *agenda setting*, *framing*, and *priming* in VoynaSlov (§3). For each media effect, we investigate the utility of the most common and recently developed NLP approaches. Our analysis first reveals evidence of manipulation tactics in our data, showing that VoynaSlov presents an avenue for studying them, and second, exposes open challenges in current NLP approaches towards uncovering, analyzing, and mitigating information manipulation campaigns.

We conclude by highlighting broader limitations of extant NLP approaches, discuss why model performance advancements have not yet translated to deployable technology in crises, and propose directions for future work to close this gap (§4). Our contributions, visualized in Figure 1, are a new data set of Russian media activity, which we use to analyze media effects, and an in-depth discussion of challenges and opportunities in NLP research on information manipulation campaigns. We hope to facilitate research on information warfare and ultimately enable reduction and prevention of disinformation and opinion manipulation.

2 VoynaSlov

VoynaSlov contains posts from Russian news outlets on VK and Twitter, which primarily feature breaking news or summaries of original articles. Here, we describe data collection and statistics.

List of News Outlets We identified Russian media outlets and their Twitter and VK handles starting from a seed list.[4] We then selected other media accounts followed by the seed outlets on Twitter, repeating until convergence. Twitter identifies state-affiliated Russian media accounts with a badge[5], which we use to label outlets as *state-affiliated* or *independent*. The resulting list was manually verified by a fluent Russian speaker and includes 23 state-affiliated and 20 independent outlets (Appendix B). However, we note that independent outlets may not be truly independent from state influence, particularly due to restrictions on free speech since the invasion.[6] We collect data as early as January 2021, over a year before the war, as many believe the invasion was planned far in advance and the media may have preemptively planted narratives.[7]

	Media Posts		**Public Reac.**	
	SA	Ind	SA	Ind
VK (Pre-war)	333K	143K	11M	3M
VK (Wartime)	94K	27K	6M	430K
Twitter (Pre-war)	41K	33K	-	-
Twitter (Wartime)	109K	36K	17M	

Table 1: Number of posts/comments/tweets by state-affiliated (SA) and independent (Ind) media in VoynaSlov.

VoynaSlov-VK We collect VK posts from identified media accounts with the VK Open API.[8] Table 1 provides a detailed breakdown of the 21M+ posts collected. For each post, we collect the number of views, likes, the presence of images, videos, and links, and comments to capture *public reaction*.

VoynaSlov-Twitter We similarly collect tweets and metadata such as like and retweet count from Russian media accounts. We capture public reaction with the Twitter search API and iteratively craft search terms. Starting from a small seed list, we collect an initial set of tweets. We augment our seed list with frequent terms from this initial set judged to be relevant to the war. After several rounds, our final list contains 264 terms and hashtags (Appendix A). Since the Twitter API only

[3] https://www.linkfluence.com/blog/russian-social-media-landscape

The data, available at https://github.com/chan0park/VoynaSlov, currently contains >38M posts and will continue to be updated.

[4] "Mass media in Russia", Wikipedia

[5] About government and state-affiliated...labels on Twitter

[6] Russia Takes Censorship to New Extremes, Stifling War Coverage (NYT)

[7] e.g., "Prelude to the 2022 Russian invasion of Ukraine"

[8] https://vk.com/dev/openapi All VK media account pages are publicly available. Our data release provides only posts/comment IDs to abide by VK's terms and conditions. Full data can be restored using VK Open API as long as it remains available at the time of collection.

supports search with a 7-day limit, we collect 17M public tweets from 24 Feb - 31 May 2022.

Data Statistics Table 1 presents basic data statistics. We provide additional metrics in Appendix C and summarize here. Due to Twitter's 280 character constraint, VK posts tend to be longer than tweets, and independent media posts are significantly longer than state-affiliated posts on both platforms. Compared to independent outlets, state-affiliated outlets include much more multimedia, which can be powerful framing devices (Powell et al., 2015). Most state-affiliated and independent tweets include external links, which enhance users' perceptions of trustworthiness and credibility (Morris et al., 2012; Wang and Mark, 2013). However, there is a stark difference on VK, where 76.3% of independent media posts include external links compared to just 26.5% of state-affiliated posts.

VoynaSlov suggests that state-affiliated media dominates VK, but independent media dominates Twitter. On average, state-affiliated VK accounts have 26K posts, over twice as much as independent media. This pattern is reversed on Twitter, where independent accounts are slightly more active. VK's publicly-available data includes view counts, presenting a unique opportunity to study exposure (Tewksbury et al., 2001). Not only are state-affiliated outlets more active on VK, but their content reaches a larger audience (18K vs 10K views per post).

Popularity cues, e.g., likes, comments, and retweets, can serve as indicators of the success of the media's opinion manipulation strategies (see §3.3). Independent media posts receive more engagement on Twitter, but state-affiliated posts receive more engagement than independent posts on VK. However, independent posts on VK still have engagement rates if we account for their smaller audiences. As discussed in §1, VK is more widely used in Russia and enables analyzing internal Russian information manipulation campaigns as well as reactions of people likely to be in Russia. Our data enables comparisons between VK and Twitter and could reveal differences in strategies used by state-affiliated Russian media when targeting domestic and international audiences.

3 Facilitation of NLP Research on Media Opinion Manipulation

We demonstrate how VoynaSlov can facilitate research on media opinion manipulation by focusing on three media effects: agenda-setting (§3.1), framing (§3.2), and priming (§3.3), though we note disputes over the distinctness and convergence of these concepts (Price and Tewksbury, 1997; Scheufele, 2000; Ghanem and McCombs, 2001). For each media effect, we first provide background and review existing NLP approaches. We then apply current state-of-the-art models from the most dominant NLP paradigms on VoynaSlov, presenting evidence that this data can support examinations of these effects. We conclude each subsection with open challenges exposed by our analyses.

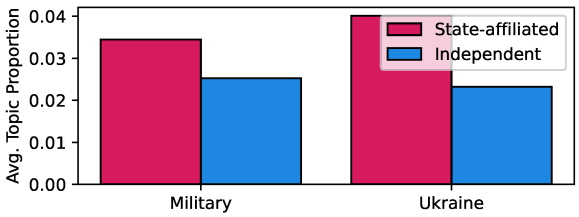

Figure 2: War-related topic proportions for state-affiliated and independent media outlets, as learned by a 30-topic CTM. We display the two topics we identified as most war-related: "Military" (top words: округа:area, авиации:aviation, военного:military, военно:military, флота:navy, учения:military exercise, су:Sukhoi Su, военнослужащие:military personnel, противника:enemy, сил:forces), and "Ukraine" (top words: днр:DPR, лнр:LPR, мариуполя:Mariupol, мирных:peaceful, жителеи:residents, украинские:Ukrainian, народнои:folk, новости:news, донбасса:Donbass, мариуполе:Mariupol). We report all topics in Figure 10.

3.1 Agenda Setting

Background *Agenda setting*, first introduced by McCombs and Shaw (1972), suggests that the importance attributed to issues by audiences is strongly correlated with the emphasis that mass media place on them (Scheufele and Tewksbury, 2007). An actor seeking to manipulate public opinion can influence how important an audience considers specific issues by reducing or increasing their representation in the media. As news topics are event-driven and agenda setting strategies are unknown in a new corpus, NLP approaches to uncovering them use statistical and unsupervised methods, including word statistics or Bayesian models and evaluating against external indicators of events (Tsur et al., 2015; Field et al., 2018b).

Results in our Data VoynaSlov facilitates examination of agenda setting by including posts

from various outlets and labels of outlets as state-affiliated or independent. Because we need unsupervised analyses of *what* topics are covered, we identify topic modeling and word frequencies as the most imminently usable NLP methods.

We employ two different topic models: a structured topic model (Roberts et al., 2016, 2019, STM) and a contextualized neural topic model (Bianchi et al., 2021a,b, CTM). The STM is a popular LDA-style probabilistic model that improves upon prior approaches by allowing users to incorporate arbitrary metadata. The CTM is based on a variational autoencoder (Srivastava and Sutton, 2017) and appends pre-trained sentence embeddings (Reimers and Gurevych, 2019) to bag-of-words document representations, reducing the bag-of-words assumptions made by traditional models. We train both models over VK posts from January 1, 2021 to May 15, 2022. For the STM, we include affiliation (state or independent) and time (days) as topic prevalence covariates. Appendix D provides details.

Figure 2 shows selected topic proportions estimated by the CTM, averaged over posts from state-affiliated and independent outlets. Appendix D reports full results for both models. Both models show differences in topic distributions in state-affiliated and independent outlets, suggesting this data offers opportunities for examining how coverage differs by outlet affiliation.

Many speculate the extent of state-affiliated media's war coverage and suggest that people in Russia have little knowledge of the invasion.[9] Omitting coverage constitutes agenda setting. The most war-related topics are CTM topics 5 and 14 (Figure 2) and STM topic 19 (Figure 11); the prevalence of these topics sharply increases in both types of news outlets in late February (Appendix D). However, these three topics have higher prevalence estimates in state-affiliated than independent media. In contrast, Figure 3 uses word statistics to more directly examines how often each media type mentions the war. A much higher proportion of independent posts mention war-related terms (e.g., "war", "operation"), especially since the invasion, which is consistent with our observations from manually reading randomly sampled posts.

Open Questions Based on these results, we highlight 3 main limitations in current approaches (topic modeling and word statistics) for uncovering agenda-setting strategies, *uninpretability*, *instability*, and *over-simplification*. The difficulty of interpreting topic modeling and related approaches has long been acknowledged (e.g., (Chang et al., 2009)) and remains an open challenge despite the continued popularity of these methods. In our data, not all topics are coherent, and even in coherent topics, comprehensiveness is difficult to determine. For example, the most Ukraine-related CTM topic (Topic 5) references the two largely-unrecognized breakaway states in eastern Ukraine: *DNR* and *LNR*, suggesting that this topic captures explicitly pro-Russia coverage of events, which we expect to be more common in state-affiliated outlets. There is no straightforward mechanism to prevent or easily recognize the one-sidedness of topics.

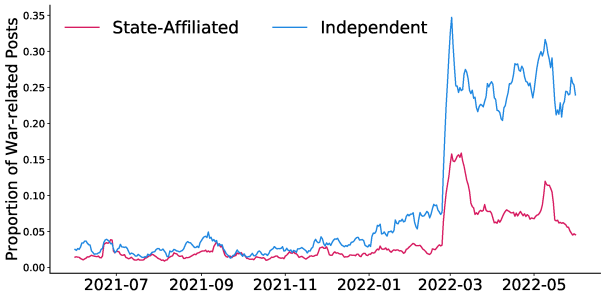

Figure 3: Proportion of posts that mention war-related terms (e.g. "war", "operation") in state-affiliated and independent outlets over time. Independent VK posts use these terms more frequently, especially following the invasion.

Relatedly, results vary even under similar models. Topic modeling is sensitive to pre-processing decisions (Denny and Spirling, 2018), and word frequencies depend entirely on the choice of words. While word statistics and manual analysis show evidence that independent outlets discuss the war more frequently than state-affiliated outlets, topic models suggest the opposite. Instability makes results difficult to trust, and more research is needed to improve consistency and reliability.

Finally, simplifying assumptions likely limit conclusions. Word-level metrics fail to account for context, and most topic models, including STMs, make bag-of-words and independence assumptions. While the CTM relaxes bag-of-words assumptions with sentence embeddings, a disadvantage compared with the STM is that it does not parameterize topics with metadata. Combining contextualized embeddings with flexible neural architectures could provide avenues for relaxing assumptions (Card et al., 2018; Zhao et al., 2021). However, using embeddings pretrained on external data risks in-

[9]Examples: Time, CNN

troducing false findings derived from the external data, rather than from the target analysis corpus (Field and Tsvetkov, 2019; Shwartz et al., 2020).

Identifying agenda setting requires examining topics in unseen corpora. Alternative methods do exist, such as embedding clustering (Sia et al., 2020), and if researchers have specific hypotheses, hand-coding articles or constructing lexicons may be possible. Nevertheless, unsupervised word-level metrics remain go-to approaches for topic analysis, and interpretability, stability, and reducing simplifying assumptions remain open challenges.

3.2 Framing

Background In media studies, whereas agenda setting refers to *what* topics are discussed, framing is based on the assumption that *how* those topics are discussed can influence the way audiences understand them (Scheufele and Tewksbury, 2007). Framing has origins in sociology (Goffman, 1974) and psychology (Tversky and Kahneman, 1981) as well as communication (Entman, 1993). Psychologists tend to focus on *equivalence* frames: different presentations of logically-identical information (Scheufele and Iyengar, 2012), such as using the phrases "90% employment" vs. "10% unemployment" (Chong and Druckman, 2007; Tewksbury and Scheufele, 2019). In contrast, *emphasis* frames present "qualitatively different yet potentially relevant considerations" (Chong and Druckman, 2007, p.114), such as focusing on free speech vs. public safety in new coverage of a protest. Frames can also be *issue-specific*, which facilitates highly detailed analyses, or *generic*, which facilitates replicability and generalizablility (De Vreese, 2005).

Much NLP research has focused on detecting *generic* frames, using the Media Frames Corpus (MFC) (Card et al., 2015; Johnson et al., 2017; Field et al., 2018a; Khanehzar et al., 2019), moral foundations (Mokhberian et al., 2020; Roy et al., 2021), or episodic and thematic frames (Mendelsohn et al., 2021). Due to high expert annotation costs, there has been considerably less attention to *issue-specific* frames, but several recent works showcase how they can enrich our understanding of discourses (Morstatter et al., 2018; Liu et al., 2019; Mendelsohn et al., 2021). NLP research also mirrors the debate over whether frames should be identified inductively as they emerge from the data under study, or a-priori based on existing theories in a deductive manner (De Vreese and Lecheler, 2012): some models are (sometimes weakly) supervised (Morstatter et al., 2018; Khanehzar et al., 2019; Roy and Goldwasser, 2020), while others are unsupervised (Kwak et al., 2021; Nicholls and Culpepper, 2021; Yu and Fliethmann, 2021).

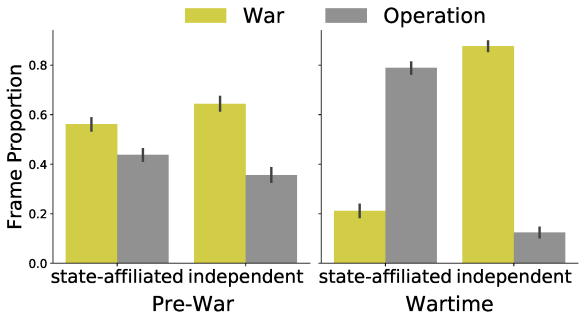

Figure 4: Proportion of media posts containing the "war" vs. "operation" equivalence frames, normalized by the number of posts using either. Since the war started, independent outlets have used the "war" frame much more frequently, while state-affiliated outlets exhibit a strong preference for the "operation" frame.

VoynaSlov offers unique opportunities to study *frame-building*, how social forces (e.g., organizational pressures or journalists' ideologies) influence what frames are cued by media coverage (De Vreese, 2005), along several dimensions, notably ownership (state-affiliated or independent), platform (Twitter or VK), and time. We first compare two *issue-specific equivalence* frames: use of words denoting "war" vs. the euphemism "military operation". Then, we develop a state-of-the-art model to analyze *generic emphasis* frames from the MFC (Card et al., 2015).

Results in our Data Whereas Figure 3 depicts how often outlets mention the war at all, Figure 4 focuses on what terminology they use by examining "war" vs. "operation" *issue-specific equivalence* frames. Since the onset of the war, independent outlets have exhibited a strong preference for "war" while state-affiliated ones more often use the "operation" euphemism. These findings are consistent with other accounts describing that the Russian government downplays the severity and aggression of the invasion and eventually even banned media from using the terms "war" and "invasion".[10]

Next, we train sentence-level classifiers to detect 15 *generic emphasis* frames using the annotated MFC, standard data for frame analysis in NLP (Card et al., 2015). We construct classifiers based

[10] CNN

	Data	Model	F1
In-domain	MFC	XLM-R$_L$	67.5
Zero-shot	Immigration	XLM-R$_L$	52.7
	Same-sex	XLM-R$_L$	50.4
	Tobacco	XLM-R$_L$	51.0
	VoynaSlov	XLM-R$_L$	33.5

Table 2: Macro-F1 results of trained MFC classifiers in in-domain, zero-shot, and VoynaSlov setups.

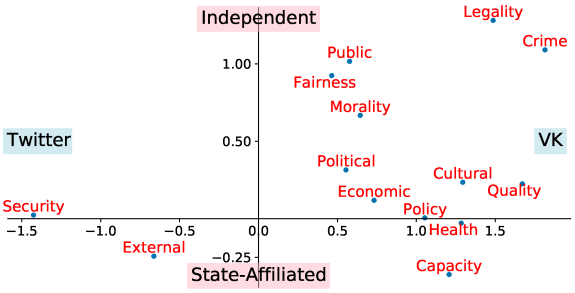

Figure 5: Association between framing and media ownership (y-axis; independent vs. state-affiliated) and social media platform (x-axis; Twitter vs. VK). The plotted values represent β coefficients when using ownership and platform as features in a logistic regression model to predict the presence of each frame. For example, the "crime" frame (upper right) is most strongly associated with independent media and VK. All values with magnitude > 0.5 are significant ($p < 0.01$).

on large pre-trained language models shown to be the state-of-the-art (Kwak et al., 2020; Akyürek et al., 2020). Unlike prior work, which evaluates in-domain, we more realistically simulate both in-domain and zero-shot scenarios with the MFC, and also evaluate with VoynaSlov. As the MFC is organized by policy issue, we simulate zero-shot classification by leaving one issue as a test set (e.g. immigration) and using remaining data for training and development (e.g., same-sex marriage and tobacco). To evaluate over VoynaSlov, a native Russian speaker annotated randomly sampled sentences from VoynaSlov-VK.[11]

Unsurprisingly, performance significantly drops in the zero-shot setting within the MFC corpus, an even more so in VoynaSlov, which features content in a different language[12], cultural context, style, and format than the MFC news articles (Table 2). Nevertheless, we analyze frame-building with the predicted MFC frames.[13] Following Mendelsohn et al. (2021), we fit separate mixed-effects logistic regression models with each frame's presence as a binary dependent variable. Fixed effects include ownership (state-affiliated vs. independent) and platform (Twitter vs. VK), and we control for specific media outlet and date as random effects.

Cued MFC frames vary across platform and media ownership (Figure 5). Compared to state-affiliated outlets, independent outlets are more likely to use *Legality* and *Crime & Punishment*, which possibly indicates questioning legal precedent of the invasion or criminal activity of the military. Frames that capture human rights (*Fairness & Equality*, *Morality*) and citizens' views (*Public Sentiment*) are also significantly associated with independent media. Regarding platform effects, most frames are associated with VK, which may reflect lack of platform character limits: VK enables more in-depth posts that explicitly cue MFC frames. However, *External Regulation & Reputation* and *Security & Defense* are used significantly more on Twitter. Since both frames face "outward" by focusing on Russia's relationships with other countries, this result supports our speculation that Russian media use Twitter to reach people outside of Russia, and VK to reach people within Russia. However, it is difficult to draw more specific conclusions because Figure 5 does not reveal what each frame means in the context of VoynaSlov; indeed, the MFC typology may obscure more meaningful framing patterns (Mendelsohn et al., 2021).

Open Questions We focus on two open questions exposed by our data and analysis: *Unclear typology* and *Domain-specificity*. Even in other disciplines, there is no consensus on whether frames should be specific or generic, equivalence or emphasis, and inductive or deductive, and not only what typology is appropriate for specific research questions and corpora, but also what best reflects the psychology processes by which people are actually influenced by framing (De Vreese and Lecheler, 2012). Generic emphasis frames have become the dominant typology in NLP, likely because this approach aligns with standard NLP paradigms of classification and reusable data. However, the difficulty of interpreting Figure 5 suggests the MFC frames may not be the most relevant in VoynaSlov, despite this being the easiest typology to imminently operationalize. While less-supervised approaches exist, they often use word statistics or hierarchical topic models, which lead to the same

[11] See Appendix E for model and F for annotation details.

[12] When we apply the MFC classifier over VoynaSlov for both evaluation and analysis, we translate original Russian texts to English as MFC only contains English data.

[13] See Appendix E.3 for more information.

interpretability, stability, and simplification concerns as in §3.1 (Nguyen et al., 2013; Roberts et al., 2016; Demszky et al., 2019; Bhatia et al., 2021).

Even with an established typology, framing is highly context-dependent, as it is a "bridging concept between cognition and culture" (Van Gorp, 2007, p.61). The need to capture subtle and nuanced content is an ongoing challenge in NLP research: models often overfit to shallow lexical features and generalize poorly to new domains (Daume III and Marcu, 2006), which Table 2 exemplifies. While domain-adaption is a large field in NLP, it is unclear how well these approaches work in detecting nuanced concepts, and model complexity may reduce deployability. Although surfacing framing strategies remains challenging, particularly in an emerging crisis outside of the U.S. political context, Figures 4-5 show signs of *frame-building* in our corpus, suggesting that VoynaSlov offers avenues for future framing research.

3.3 Priming

Background Priming typically refers to the effects of framing and agenda setting (Entman, 2007). Some researchers use the term to specifically refer to "changes in the standards that people use to make political evaluations" (Iyengar and Kinder, 1987), which is associated closely with agenda setting (Scheufele and Tewksbury, 2007; Moy et al., 2016). For example, news coverage of particular issues encourages audiences to base judgements of leaders and governments on these issues at the exclusion of others, including during elections (Scheufele and Tewksbury, 2007). The effects of framing, or audiences' adoption of frames presented in news as ways to understand issues, can then be termed *frame setting* (Moy et al., 2016). We take a broad definition of the term *priming* and consider both agenda setting and framing effects in this section.

Little work in NLP has focused on priming. Some aspects of how readers respond to news coverage fall outside the scope of text analysis (Zubiaga et al., 2016) and may be better examined through surveys, historical polls, or election results (Price et al., 1997; Valkenburg et al., 1999; Zhou and Moy, 2007). User reactions on social media, including likes, shares, and comments, can also offer some insight into how framing and agenda setting strategies are received. In this section, we show that the inclusion of reactions in VoynaSlov offers an avenue for studying priming, but that this line of research raises technical and ethical challenges.

Views	Likes	Reposts	Engagement
Has Video (.24)	Has Video (.56)	Policy (.40)	Public Sent. (.31)
Has Image (.20)	Public Sent. (.35)	Has Video (.37)	Has Video (.23)
Public Sent. (.13)	Morality (.27)	Morality (.28)	Morality (.17)
Crime (.12)	Security (.21)	Qual. of Life (.24)	Has Link (.16)
Fairness (.11)	Has Image (.18)	Capacity (.22)	Security (.08)

Table 3: Frame-setting results on engagement metrics. The numbers in parentheses indicate each feature's coefficient of trained regression models.

Results in our Data We investigate the effects of MFC frames on user engagement with mixed-effects linear regression models. Independent variables include the presence of each frame, ownership (state-affiliated or independent), and if a post has an image, video, or link (each coded as binary factors). Random effects include specific outlet and date. We consider four outcomes: numbers of views, likes, and reposts (all log-scaled), and engagement rate, defined as the sum of like, repost, and comment counts normalized by the view count.

Table 3 shows the variables most strongly associated with each engagement metric. Including multimedia, especially videos, is strongly predictive of user engagement. Civilian-focused frames (e.g. *Public Sentiment*, *Morality*), are linked to higher engagement in all four measurements. However, as in §3.2, these correlations are sometimes difficult to interpret. For example, *Policy* is most strongly associated with more reposts, but we are unable to decipher what this frame captures and what the implications may be for media manipulation campaigns.

We further investigate how frames adopted by media posts affect readers' frame usage. Figure 6 shows the average frame proportion of user comments, depending on the frames of original posts and media state-affiliation. On average, users leave significantly fewer comments with *Political*, *Public Sentiment*, and *Fairness & Equality* frames on state-affiliated media, which might be related to Russian laws imposing strict censorship. Instead, comments on state-affiliated posts more often employ *Economic* and *Quality of Life* frames which might reflect what readers prioritize or feel comfortable discussing.

Open Questions We identify three primary open questions: *Data validation*, *Privacy*, and *Technology Misuse*. While we investigate priming through user reactions and comments, this approach is fun-

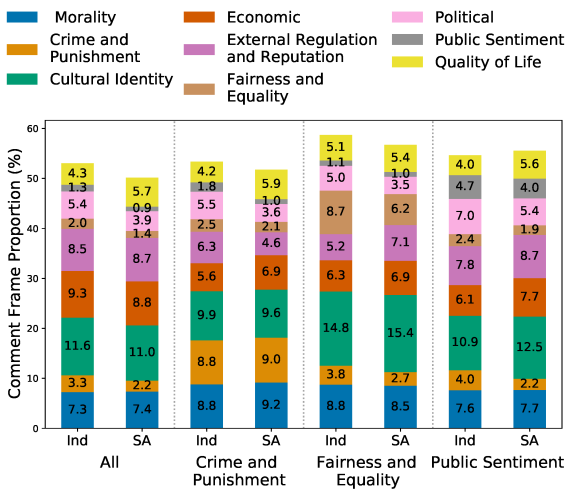

Figure 6: Comparison of frame proportion of comments of independent (Ind) and state-affiliated (SA) media, for each frame used by posts.

damentally limited. Priming relates to cognitive influence on individual users, which is impossible to capture through aggregate metrics. Additionally, creating fake or disingenuous accounts to promote particular content is a known information manipulation strategy (King et al., 2017; Arif et al., 2018), and studying unverified aggregated data could be capturing content by these accounts. Even if posts are made by genuine social media users, we cannot determine how reflective social media activity is of true cognitive states, individual attitudinal changes, or macro-level public opinion shifts.

Deeper investigations of framing and agenda setting effects on individual users could better address this data validation issue, but raises concerns of *privacy* and the potential for *user-targeting*. Russian state surveillance of social media activity is well-established and there have been cases of civilians being arrested based on social media activity (Mejias and Vokuev, 2017; Gabdulhakov, 2020). Furthermore, big data and computational social science research has been confronted with ethical challenges due to the lack of informed consent from participants and unawareness of how their data is being used by researchers (Fiesler and Proferes, 2018; Lazer et al., 2020). Privacy-preserving modeling of social phenomena surfaced by language remains an open challenge in NLP. Recent social science research has linked social media activity with traditional surveys that abide by principles of informed consent and ethical research (Eady et al., 2019), which could be a viable path forward in NLP as well.

Finally, research on priming has greater *misuse* potential than framing or agenda setting. Framing and agenda setting analyses identify manipulation strategies already in use, and thus are unlikely to inform malicious actors on how to generate propaganda. In contrast, priming research could directly inform malicious actors on which strategies are effective, though this has not deterred political psychology work on priming (e.g. Conway et al., 2017). The misuse potential poses a paradox: uncovering effective framing and agenda setting strategies is more important than ineffective ones, but analyzing their effectiveness can lead to wider adoption. These challenges make studying priming from an NLP perspective difficult. We suggest that VoynaSlov facilitates research on understanding some aspects of the visibility and adoption of framing and agenda setting, but that more in-depth analyses may fall outside the scope of NLP research.

4 Discussion

There are numerous opportunities for NLP research to have positive impacts in identifying and mitigating information manipulation campaigns. As a first step, we release VoynaSlov, focused on Russian media activity on social media before and during the 2022 Russia-Ukraine war. Grounded in political communication research on media effects, we apply both traditional and state-of-the-art NLP techniques to analyze the language underpinning information manipulation. Indeed, we uncover variations in agenda setting and framing strategies across time (pre-war or wartime), social media platforms (VKontakte or Twitter), and media control (state-affiliated or independent). We encourage future work to continue to explore the plethora of social science theories and NLP techniques to analyze the data in VoynaSlov. Furthermore, we hope that VoynaSlov will aid future efforts in using NLP to address not only the 2022 Russia-Ukraine war, but emerging crisis situations more broadly.

Through both our data collection and analysis, we learn that vast improvements in model performance on core NLP tasks have not yet translated into deployable technology capable of addressing information manipulation campaigns. Our work suggests several reasons for this discrepancy, and we foreground such limitations in order to set forth concrete directions for future work. First, much prior work has focused on developing supervised models to detect fake news and propaganda in iso-

lated media texts, including the establishment of shared tasks and standardized data sets (Thorne et al., 2018b; Da San Martino et al., 2020; Shaar et al., 2021). While there are settings where these approaches can be useful, they typically require carefully-labeled data that are unavailable in emerging settings, and we did not identify them as applicable in our analysis of VoynaSlov. Similarly, NLP advancements in model-pretraining (Brown et al., 2020) are difficult to deploy over emerging data: most pre-training data is in English, and it is difficult to disentangle patterns in the target data from ones learned during pre-training (Field and Tsvetkov, 2019; Shwartz et al., 2020). Pre-annotated data and pre-trained models can be immensely valuable for analyzing the past but have limited utility in understanding ongoing events.

Second, even technological approaches that aim to target ongoing events, just as systems for aiding human fact-checkers (Nakov et al., 2021) typically consider only the most extreme and overt form of opinion manipulation: disinformation and fake news. However, our results build upon existing communications studies, demonstrating that media manipulation often constitutes more subtle strategies, such as selectively covering (or avoiding) issues (§3.1) and changing minor word choices to influence audiences (§3.2). Little NLP work has examined *agenda setting* and *priming* at all. While substantial work has focused on *framing* (§3.2), it disproportionally focuses on U.S. politics, with few applications to non-English languages, other social contexts, or information warfare.

Third, NLP research on media manipulation has primarily examined isolated news texts without additional context, neglecting the larger hybrid media ecosystem comprised of intricate interactions between journalists, media organizations, political actors, social media platforms, and civilians (Chadwick, 2017). VoynaSlov attempts to facilitate research in this area with content from a specific context and includes multiple outlets and platforms. Nevertheless, it remains challenging to truly comprehend the media's motivations for how they present the news and their desired effects on public opinion, which then enables specific and nuanced analyses of manipulation strategies.

Although we emphasize limitations in existing NLP approaches, we conclude by asserting that NLP has a unique opportunity to uncover information manipulation campaigns and contribute to social science research. Media effects have been rigorously studied within social science, but common approaches, including focused analyses of small sets of articles and human experiments with constructed stimuli in highly-contrived settings, are insufficient for assessing the scale and societal impact of media manipulation. Computational methods can be representative of the full media environment and capture more realistic audience responses to news content shared on social media.

This work aims to shift the paradigm for research on automated opinion manipulation to encompass broader tactics, have grounding in social science theory, and incorporate emerging context. We believe these expansions will enable NLP to have positive impact *during*, rather than *after*, ongoing crisis situations. We hope that our release of VoynaSlov, our analysis of media effects, and our discussion of open NLP challenges facilitate the detection and ultimately the prevention of information manipulation.

5 Limitations

Our work includes the release of a new data set, data analysis using state-of-the-art NLP models, and a discussion of open challenges in this space. The comprehensiveness of our data is limited by decisions about the data collection process, including which news outlets to focus on and which keywords and hashtags to use when collecting tweets. While we take steps to broaden the coverage of our data, such as multiple rounds of identifying news outlets and relevant terms, collection biases could reduce the reliability of any analyses conducting with this data. Our data set cannot be considered to capture all relevant content from this time period.

Throughout §3 we focus on highlighting the limitations of current NLP approaches in this setting, and we refer to this section for details. We acknowledge that our discussing of challenges and limitations is itself limited by the discussion framework. Structuring our analysis using different social science theories could lead to different results. We additionally focus on entirely text analysis and do not discuss limitations related to other types of media, such as images or video (Beskow and Carley, 2019). Finally, our discussion of limitations is based on our choice of NLP methodology to use over our data. While we attempt to select state-of-the-art models for the most dominant NLP research paradigms for each media effect, other

methods and paradigms that reduce our discussed limitations may exist.

6 Ethical Considerations

Given the ongoing war and the limitations on free speech in Russia, including the recently passed law that punishes spreading "false information" with up to 15 years in prison, it is possible that our data set contains content that could have physical and legal ramifications for individual users or media outlets. Even in our initial data collection, some VK data was flagged as deleted by moderators. We take several steps to mitigate the impact our work may have on the risk to individuals or media outlets. All of the data collected in this work is publicly available and we do not make any attempt to uncover non-public data. While we do include posts by general users on Twitter, we primarily focus on posts from media outlets and replies to them, where we can assume a lower expectation of privacy. In order to preserve users' ability to delete content, we do not release any raw text data and instead only release post IDs, which other researchers can use to recollect raw data, if it has not been removed. We further note that all data was collected in accordance with social media platforms' terms of service.

Throughout this work, we also avoid using specific examples from the data or referring to individual users. We encourage future work on this data to exercise similar caution, and we do not condone any research that attempts to deanonymize or profile users or identify narratives that could result in individuals being targeted. We refer to Vitak et al. (2016) and Williams et al. (2017) for a more in-depth discussion of ethical considerations of research using social media data.

We also primarily focus on news content posted by Russian media outlets, which we suggest provides avenues for studying disinformation, because of prior work on Russian information manipulation strategies and because Russia is the aggressor in this conflict. However, we note that independent reports have also found evidence of misinformation perpetuating pro-Ukranian narratives.[14] More generally, the authors of this work are situated in the U.S. and our assumptions in this work (e.g. that Russia is the aggressor) reflect this context, but we note that this viewpoint is not universal.

[14] https://www.newsguardtech.com/special-reports/russian-disinformation-tracking-center/

7 Acknowledgments

We thank the anonymous reviewers for their feedback, as well as the Text As Data 2022 audience, especially Sarah Dreier. A.F. and J.M. gratefully acknowledge support from the Google PhD Fellowship. C.Y.P. gratefully acknowledges support from KFAS. Y.T.gratefully acknowledges support from NSF CAREER Grant No. IIS2142739, the Alfred P. Sloan Foundation Fellowship, and the DARPA Grant under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within# blacklivesmatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27.

David M Beskow and Kathleen M Carley. 2019. Social cybersecurity: an emerging national security requirement. Technical report, Carnegie Mellon University Pittsburgh United States.

Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. OpenFraming: Open-sourced tool for computational framing analysis of multilingual data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

Samantha Bradshaw, Hannah Bailey, and P Howard. 2021. Industrialized disinformation: 2020 global inventory of organized social media manipulation. computational propaganda research project.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics.

Andrew Chadwick. 2017. *The hybrid media system: Politics and power*. Oxford University Press.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Emily Chen and Emilio Ferrara. 2022. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between ukraine and russia. *arXiv preprint arXiv:2203.07488*.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Lucian Gideon Conway, Meredith A Repke, and Shannon C Houck. 2017. Donald trump as a cultural revolt against perceived communication restriction: Priming political correctness norms causes more trump support. *Journal of Social and Political Psychology*, 5(1).

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126.

Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal & document design*, 13(1).

Claes H De Vreese and Sophie Lecheler. 2012. News framing research: An overview and new developments. *The SAGE handbook of political communication*, pages 292–306.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew J. Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Gregory Eady, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A Tucker. 2019. How many people live in political bubbles on social media? evidence from linked survey and twitter data. *Sage Open*, 9(1):2158244019832705.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018a. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018b. Framing and agenda-setting in Russian news: a

computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2550–2560, Florence, Italy. Association for Computational Linguistics.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.

Rashid Gabdulhakov. 2020. (con) trolling the web: Social media user arrests, state-supported vigilantism and citizen counter-forces in russia. *Global Crime*, 21(3-4):283–305.

Salma I Ghanem and Maxwell McCombs. 2001. The convergence of agenda setting and framing. In *Framing public life*, pages 83–98. Routledge.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Yevgeniy Golovchenko, Mareike Hartmann, and Rebecca Adler-Nissen. 2018. State, media and civil society in the information warfare over ukraine: citizen curators of digital disinformation. *International Affairs*, 94(5):975–994.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Mario Haim, Anna Sophie Kümpel, and Hans-Bernd Brosius. 2018. Popularity cues in online media: A review of conceptualizations, operationalizations, and general effects. *SCM Studies in Communication and Media*, 7(2):186–207.

Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Tristan Braud, and Pan Hui. 2022. Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*.

Shanto Iyengar and Donald R Kinder. 1987. *News that matters: Television and American opinion*. University of Chicago Press.

Kristen Johnson, I-Ta Lee, and Dan Goldwasser. 2017. Ideological phrase indicators for classification of political discourse framing on twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 90–99.

Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.

Gary King, Jennifer Pan, and Margaret E. Roberts. 2017. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *12th ACM Conference on Web Science*, pages 305–314.

Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. Frameaxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*.

Mykola Makhortykh and Maryna Sydorova. 2017. Social media and visual framing of the conflict in eastern ukraine. *Media, war & conflict*, 10(3):359–381.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *The Public Opinion Quarterly*, 36(2):176–187.

Ulises A Mejias and Nikolai E Vokuev. 2017. Disinformation and the media: the case of russia and ukraine. *Media, culture & society*, 39(7):1027–1042.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *International Conference on Social Informatics*, pages 206–219. Springer.

Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):1–18.

Patricia Moy, David Tewksbury, and Eike Mark Rinke. 2016. *Agenda-Setting, Priming, and Framing*, pages 1–13. John Wiley Sons, Ltd.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Viet-An Nguyen, Jordan L Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Tom Nicholls and Pepper D Culpepper. 2021. Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 38(1-2):159–181.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pablo Porten-Cheé, Jörg Haßler, Pablo Jost, Christiane Eilders, and Marcus Maurer. 2018. Popularity cues in online media: Theoretical and methodological perspectives. *SCM Studies in Communication and Media*, 7(2):208–230.

Thomas E Powell, Hajo G Boomgaarden, Knut De Swert, and Claes H de Vreese. 2015. A clearer picture: The contribution of visuals and text to framing effects. *Journal of communication*, 65(6):997–1017.

Vincent Price and David Tewksbury. 1997. News values and public opinion: A theoretical account of media priming and framing. *Progress in communication sciences*, pages 173–212.

Vincent Price, David Tewksbury, and Elizabeth Powers. 1997. Switching trains of thought: The impact of news frames on readers' cognitive responses. *Communication research*, 24(5):481–506.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2019. stm: An r package for structural topic models. *Journal of Statistical Software*, 91(2):1–40.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Shamik Roy, María Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958.

Arturas Rozenas and Denis Stukal. 2019. How autocrats manipulate economic news: Evidence from russia's state-controlled television. *The Journal of Politics*, 81(3):982–996.

Dietram A. Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass Communication and Society*, 3(2-3):297–316.

Dietram A Scheufele and Shanto Iyengar. 2012. The state of framing research: A call for new directions. *The Oxford handbook of political communication theories*, pages 1–26.

Dietram A Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 82–92, Online. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017*.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

David Tewksbury and Dietram A Scheufele. 2019. News framing theory and research. *Media Effects: Advances in Theory and Research: Fourth Edition*, pages 51–68.

David Tewksbury, Andrew J Weaver, and Brett D Maddex. 2001. Accidentally informed: Incidental news exposure on the world wide web. *Journalism & mass communication quarterly*, 78(3):533–554.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *SCIENCE*, 21:1–30.

Patti M Valkenburg, Holli A Semetko, and Claes H De Vreese. 1999. The effects of news frames on readers' thoughts and recall. *Communication research*, 26(5):550–569.

Baldwin Van Gorp. 2007. The constructionist approach to framing: Bringing culture back in. *Journal of communication*, 57(1):60–78.

Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 941–953, New York, NY, USA. Association for Computing Machinery.

Yiran Wang and Gloria Mark. 2013. Trust in online news: Comparing social media and official media use by chinese citizens. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 599–610.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168. PMID: 29276313.

Qi Yu and Anselm Fliethmann. 2021. Frame detection in german political discourses: How far can we go without large-scale manual corpus annotation? In *1st Workshop on Computational Linguistics for Political Text Analysis*, pages 13–24.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Yuqiong Zhou and Patricia Moy. 2007. Parsing framing processes: The interplay between online public opinion and media coverage. *Journal of communication*, 57(1):79–98.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

A Search Terms for Tweet Collection

Table 4 describes the number of total search terms we curated in each update. We describe each update in the following paragraphs.

Version 0 An initial round of defining hashtags and keywords; we manually collected general keywords and hashtags including (1) entity names, e.g. *Russia*, *Ukraine*, names of cities from both Ukraine and Russia, (2) war-related terms, including *war*, *peace*, *UkraineRussianWar*, *RussianUkraineWar*, and (3) we include the same keyword phrases in Russian and Ukrainian languages. After we sampled 1K tweets with these general keywords, we sorted all hashtags in the data sample to augment this initial seed list.

Version 1 In a manual analysis of an initial data sample, we identified additional frequently mentioned entities, e.g. additional cities in Ukraine, names of politicians in Ukraine and Russia, stance-bearing pro-Russia and pro-Ukraine hashtags (e.g., *#IstandwithRussia*, *#stopputin*), additional hashtags referring to the Second World War (#), pro-war and anti-war hashtags (*#StopWar*, #). We note that while the overall sentiment in tweets was bearing more solidarity with Ukraine, the set of hashtags and keywords is diverse, in terms of languages (English, Russian, Ukrainian), stance (pro-Russa, pro-Ukraine, pro-war, anti-war), and in addition it includes mentions of external entities involved (NATO, Belarus, USA).

Version 2 We pulled an additional sample of 5K tweets using terms from Version 1 for a manual analysis of missing seed terms and to obtain the ranking of the terms and keywords by their frequency in the sample.

Version 3 (Final) We analyzed the tweets collected through the first 24 hours and sorted hashtags by frequency. We then manually annotated top 665 hashtags (until freq-150) and added to the list 81 most frequent conflict-related hashtags.

Versions	V0	V1	V2	V3 (final)
# of keywords	87	98	184	264

Table 4: The size of search term list in each update.

Pro-War Search Terms (8):
#crimeanspring, #мненестыдно, #русскиеидут, #deadrussiansoldiers, #istandwithputin, #imwithrussia, #своихнебросаем, #proudtoberussian
Anti-War Search Terms (12):
#нетвойне, #stoprussia, #nowar, #stopputinnow, #stopwarinukraine, #nowarwithukraine,#saveukraine, #stopputin, #нетвойнесукраиной, #stopthewar, #stopwar, #stoprussianaggression

Pro-Ukraine (19):
#standwithukraine, #fkputin, #нетпутину, #slavaukraini, #stoprussia, #saveukraine, #fklukashenko, #stoprussianaggression, #staywithukraine, #своихнебросаем, #stopputinnow, #stopwarinukraine, #nowarwithukraine, #helpukraine, #stopputin, #banrussiafromswift, #славаукраїні, #istandwithukraine, #istandwithzelenskyy
Pro-Russia (8):
#donbasstragedy, #crimeanspring, #русскиеидут, #istandwithputin, #istandwithrussia, #imwithrussia, #своихнебросаем, #proudtoberussian

Final Search Terms (264):
#Украина, #нетвойне, #Ukraine, #Россия, #украина, #Харьков, #НетВойне, #Херсон, #мариуполь, #война, #НетПутину, #ukraine, #mariupol, #україна, #Україна, #StopWar, #UkraineWar, #нетвойнесУкраиной, #Russia, #россия, #Путин, #StopTheWar, #путин, #StopRussianAggression, #Киев, #Мариуполь, #NoWarWithUkraine, #UkraineRussie, #StandWithUkraine, #Зеленский, #РФ, #RussiaUkraineConflict, #SaveUkraine, #StopRussia, #Сумы, #UkraineInvasion, #stopputin, #СлаваУкраїні, #UkraineRussiaCrisis, #Гостомель, #UkraineConflict, #FlyAway, #войска, #ДНР, #NoWar, #Одесса, #Харків, #Kiev, #Ukraina, #России, #Херсоне, #путинубий-

ца, #протесты, #Donbass, #нацизм, #Одеса, #геноцид, #Mariupol, #eu, #europe, #фашизм, #Odesa, #Odessa, #ЛНР, #лукашенко, #Москва, #IStandWithUkraine, #Мелитополь, #невойне, #протесты, #нацизм, #фашизм, #геноцид, #война, #WWII, #nuclearwar, #санкции, бомбит, нацизм, войска, диверсионные, Удары, армия, пиздец, мир, мирные, #DearsForPeace, МыНеМолчим, #newsua, #newsru, #НетвойнеУкраиныпротивДонбасса, #санктпетербург, #зеленський, #ДаПобеде, #SWIFT, #київ, #мынемолчим, #тихийпикет, #ЕС, #russianinvasion, #Противійни, #ПУТИН_ВИНОВЕН, #донбасс, #EuroMaidan, #Ирпень, #беларусь, #Maidan, #МойЛуганск, #StayWithUkraine, #Zelenskiy, #НетБезумию, #питер, #CoupdEtat, #Протесты, #бандеровцы, #всу, #Кремль, #BanRussiafromSwift, #бомбардировки, #Лавров, #Rusya, #МОСКВА, #АрмияРоссии, #SanctionRussiaN, #российское_вторжение, #ДавайЗаМир, #НоваяКаховка, #Irpin, #worldwar3, #Moscow, #дапобеде, #переговоры, #русские, #ООН, #Евросоюз, #путинхуйло, #терроризм, #Минобороны, #WWIII, #митинг, #РусскаяВесна, #DonbassWar, #янемолчу, #moscow, #РоссияУбивает, #русскийсолдат, #времяпомогать, #Шойгу, #россияне, #ЗаПрезидента, #армия, #наДонбассевойна8лет, #МнеНеСтыдно, #русскиймир, #россияукраина, #ЯМыПутин, #ЕдинаяРоссия, #DeadRussianSoldiers, #ВКСРоссии, #КремлевскиеСМИ, #Русскиелюди, #КризиснаДонбассе, #денацификация, #Putler, #русскийТопот, #россиявставай, Путин, Россия, Украина, Киев, Путину, Украины, Россияне, АЭС, США, НАТО, Зеленский, #Chernihiv, #Kherson, #Украина, #Ukraine, #Россия, #украина, #Харьков, #Херсон, #мариуполь, #Киев, #Мариуполь, #Зеленский, #РФ, #Russia, #россия, #Путин, #путин, #ДНР, #Харків, #Kiev, #России, #Ukraina, #Херсоне, #донецк, #Луганск #СвоихНеБросаем, #DonbassTragedy, #See4Yourself, #Think4Yourself, #WeRemember, #IstandwithRussia, #Novorossiya, #Donbass, #РаботайтеБратья, #Welcome2Crimea, #Crimea, #CrimeanSpring, #IStandWithPutin, #своихнебросаем, #русскиеидут, #imwithrussia, #ProudToBeRussian, #нетвойнесУкраиной, #StopTheWar, #StopRussianAggression, #NoWarWithUkraine, #StandWithUkraine, #UkraineRussie, #SaveUkraine, #StopRussia, #СлаваУкраїні, #UkraineRussiaCrisis, #UkraineInvasion, #stopputin, #NoWar, #путинубийца, #RussiaUkraineConflict, #UkraineConflict, #StopPutinNow, #StopWar, #StandWithUkraine, #SlavaUkraini, #HelpUkraine, #invaision, #РоссияБЕЗпутина, #PutinIsFalling, #PutinWarCrimes, #StopWarInUkraine, #resist, #SlavaUkrayini, #FreeBelarus, #FKPutin, #FKLukashenko, #UkraineInvasion, #правдаовойне, #IStandWithZelenskyy, #IStandWithUkraine, #StopWarInUkraine, #PutinWarCriminal, #ClosetheSkyoverUkraine, #AdolfPutin, #PutinHitler, #RussiaInvadedUkraine, #нетвойне, #НетВойне, #НетПутину, #UkraineWar

B Twitter/VK Handles of Russian News Outlets

Media Name	Twitter Handle	VK Handle
TV Rain	@tvrain	tvrain
Alexei Navalny	@navalny	navalny
IStories	@istories_media	istories.media
OVD-Info	@OvdInfo	ovdinfo
Novaya Gazeta	@novaya_gazeta	novgaz
DW (Deutsche Welle)	@dw_russian	
BBC Russia	@bbcrussian	bbc
MediaZona	@mediazzzona	mediazzzona
Radio Liberty	@SvobodaRadio	svobodaradio
The Insider	@the_ins_ru	theinsiders
Forbes Russia	@ForbesRussia	forbes
Meduza	@meduzaproject	meduzaproject
Current Time TV	@CurrentTimeTv	currenttimetv
RTVI	@RTVi	rtvi
Voice of America	@GolosAmeriki	golosameriki
Snob Project	@snob_project	snob_project
Echo of Moscow	@EchoMskRu	
FBK	@fbkinfo	
Reuters Russia	@reuters_russia	
Znak.com	@znak_com	

Table 5: List of Independent media and their handles on Twitter and VK.

C Data Statistics and Analysis

C.1 Analysis: Post Content

Length As a consequence of Twitter's 280-character constraint, VK posts are on average sig-

Media Name	Twitter Handle	VK Handle
RT (Russian)	@RT_russian	rt_russian
RT (English)	@RT_com	
TASS (Russian)	@tass_agency	tassagency
TASS (English)	@tassagency_en	
Sputnik News	@SputnikInt	sputnikint
Sputnik (Radio)	@ru_radiosputnik	sputnik_radio
RIA Novosti	@rianru	ria
RIA Novosti (Breaking News)	@riabreakingnews	
PRIME	@1prime_ru	1prime
Ministry of Defence	@mod_russia	mil
Ruptly	@Ruptly	ruptly
Moscow 24	@infomoscow24	m24
inoSMI	@inosmi	inosmi
Life	@lifenews_ru	life
5TV	@5tv	tv5
Vesti	@vesti_news	vesti
Russia-1	@tvrussia1	russiatv
RBC	@ru_rbc	rbc
Gazeta.Ru	@GazetaRu	gazeta
Rossiyskaya Gazeta	@rgrus	rgru
Ukraina.ru	@ukraina_ru	ukraina_ru_official
Redfish	@redfishstream	
MIA Rossiya Segodnya	@pressmia	
Margarita Simonyan	@M_Simonyan	
Zubovski 4	@zubovski4	
DVostok	@media_dv	
Vladimir Soloviev	@VRSoloviev	

Table 6: List of State-affiliated media and their handles on Twitter and VK.

	VK				Twitter		
	Media		Public		Media		Public
	SA	Ind	SA	Ind	SA	Ind	*Twit.*
Posts per account	26K	11K	34.7	59.8	5.6K	5.8K	23.1
Word count	26.1	50.8	14.9	16.7	19.2	23.3	19.2
Image/video (%)	70.2	21.4	8.2	12.3	50.9	28.0	9.6
Link (%)	26.5	76.3	0.2	0.9	78.8	75.3	7.1
Likes	81.9	66.8	2.0	2.3	39.4	249.4	4.3
Comments/RTs	39.3	25.9	-	-	11.2	60.0	399.7
Views	18K	10K	-	-	-	-	-

Table 7: Statistics of VoynaSlov. The difference between state-affiliated and independent media was statistically significant ($p < 0.05$) for all metrics in both Media Posts and Public Reaction.

nificantly longer than Twitter posts. Interestingly, independent media posts are significantly longer than state-affiliated media posts on both platforms. This pattern is consistent in comments on VK.

Images and videos can themselves be powerful framing devices (Powell et al., 2015), and images posted to VK in particular have been used to understand opposing representations and interpretations of the Russia-Ukraine conflict (Makhortykh and Sydorova, 2017). On both VK and Twitter, state-affiliated media posts include much more multimedia (images and video) than independent media posts (70.3% vs. 21.5% on VK and 59.6% vs. 31.9% on Twitter, respectively).

External links In contrast to embedded multimedia, a slightly different pattern emerges for the inclusion of external links, which have been shown to enhance users' perceptions of trustworthiness and credibility on social media (Morris et al., 2012; Wang and Mark, 2013). The majority of both state-affiliated and independent media posts on Twitter include external links (70.5% and 72.5%, respec-

tively), possibly again a consequence of Twitter's constraint affordance. However, there is a stark difference on VK, where 76.3% of independent media posts include external links compared to just 26.4% of state-affiliated posts. As expected, a much lower proportion of public Tweets and public comments to media posts on VK contain embedded multimedia or external links. Public tweets collected via hashtags have slightly higher rates of including multimedia and links compared to VK comments, but are much lower compared to media posts (9.9% of public tweets include images or video, and 6.9% include external URLs).

C.2 Analysis: Activity and User Engagement

Analyses of account activity and user engagement suggests that state-affiliated media dominates VK, but independent media dominates Twitter.

Account Activity On average, each state-affiliated media account included in VoynaSlov-VK has nearly 25K posts, more than twice as much as independent media which averages 11K posts per account. This pattern is reversed on Twitter, where independent accounts are slightly more active than state-affiliated accounts.

We also observe a high degree of self-sorting among users who comment on VK media posts: 74.4% comment only on state-affiliated posts, 16.8% only on independent posts, and only 8.8% of users have commented on both types of media posts. In other words, most people who comment on state-affiliated posts never comment on independent posts and vice versa. While we do not have user-level data about media exposure, this pattern suggests that information from state-affiliated and independent media reach disparate audiences.

Views Unlike most platforms studied by NLP and computational social science researchers, VK's publicly-available data includes view counts (i.e. impressions) and thus presents a unique opportunity to study incidental exposure to media content (Tewksbury et al., 2001). Not only are state-affiliated outlets more active on VK than independent outlets, but also each post on average reaches a larger audience (17K vs 10K views, respectively).

Interactive engagement metrics Popularity cues, such as the numbers of likes, comments, and retweets, can serve as an indicator of the success of the media's agenda-setting, framing, and propaganda strategies. These popularity cues have further consequences: they can be used to recommend content on social media platforms and thus impact users' media diets, and they can act as heuristics for people trying to decide what media content is credible, accurate, and important (Haim et al., 2018; Porten-Cheé et al., 2018). Consistent with the idea that the Twitter public sphere is more globally-oriented, independent media posts receive more engagement on Twitter than state-affiliated posts. In contrast, state-affiliated media posts on VK receive more engagement than independent posts. However, we note that independent posts on VK still have a higher rate of engagement if we account for their smaller audiences (view counts).

C.3 Volume over Time

In February and March 2022, immediately after the war began, the volume of posts by media accounts and comments in both VoynaSlov-VK and VoynaSlov-Twitter significantly increased (Figure 8 and Figure 7). However, on March 4, Putin signed a new bill called "fake news laws" which punishes spreading "false information" with up to 15 years in prison. Consequently, many independent Russian media outlets including TV Rain and Radio Liberty temporarily suspended operations, while others announced that they were stopping coverage of the invasion because of the signed bill; these independent outlets include Colta.ru, Snob Project, Znak.com, and Novaya Gazeta.[15] The impact of the censorship is also evident in our data set, as we see a significant decrease in the volume of independent media accounts' posts and comments to independent media starting March 2022.

We also note that state-affiliated media accounts became extremely active on Twitter after the war started, even when compared to their own activity on VK. For instance, the number of state-affiliated tweets in the first half of May greatly surpasses the volume from the first half of April, but the opposite trend is observed on VK. This suggests a recent shift in Russia's state-affiliated media strategy: they are focusing more efforts on reaching and spreading (dis)information to a global audience through Twitter, rather than a primarily Russian audience through VK.

While we can divide media posts and their comments according to state-affiliated and independent outlets, we do not have user-level information about

[15] https://www.amnesty.org/en/latest/news/2022/03/russia-kremlins-ruthless-crackdown-stifles-independent-journalism-and-anti-war-movement/

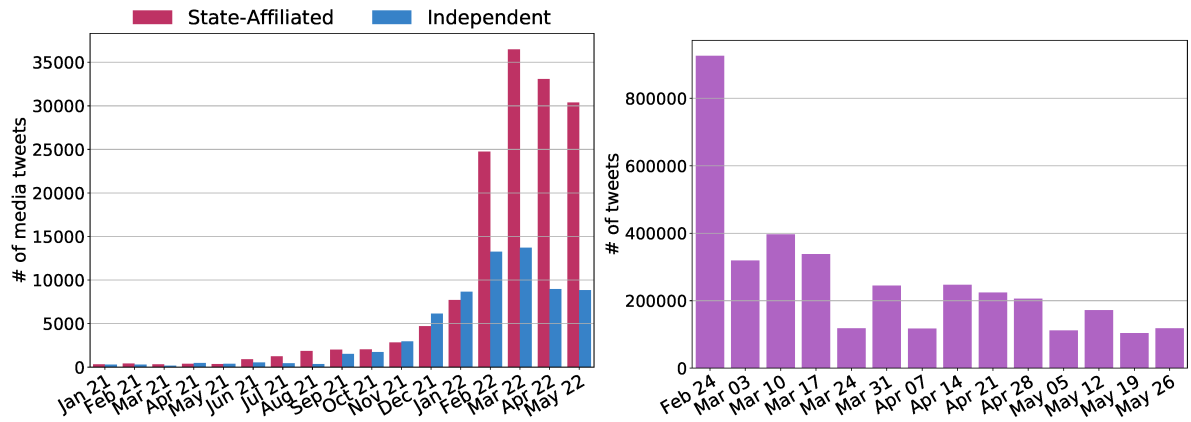

Figure 7: Total volume of tweets from media accounts (left) and general accounts (right) in VoynaSlov-Twitter.

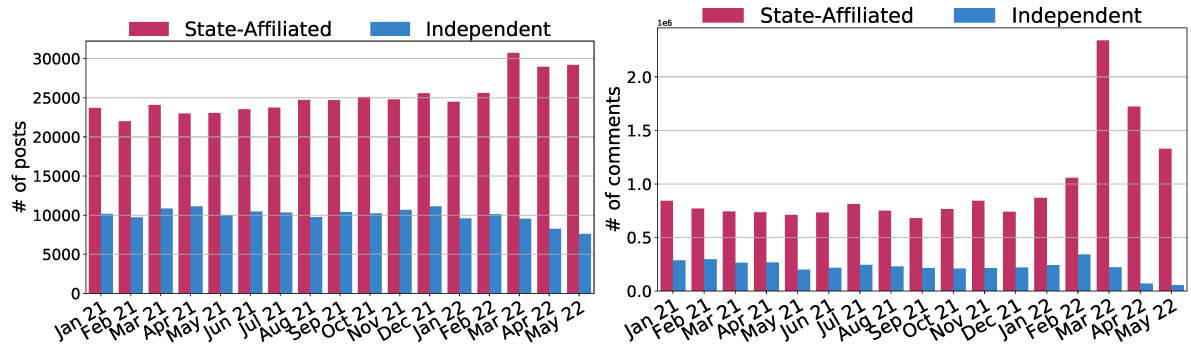

Figure 8: Total volume by month of posts (left) and comments (right) in VoynaSlov-VK.

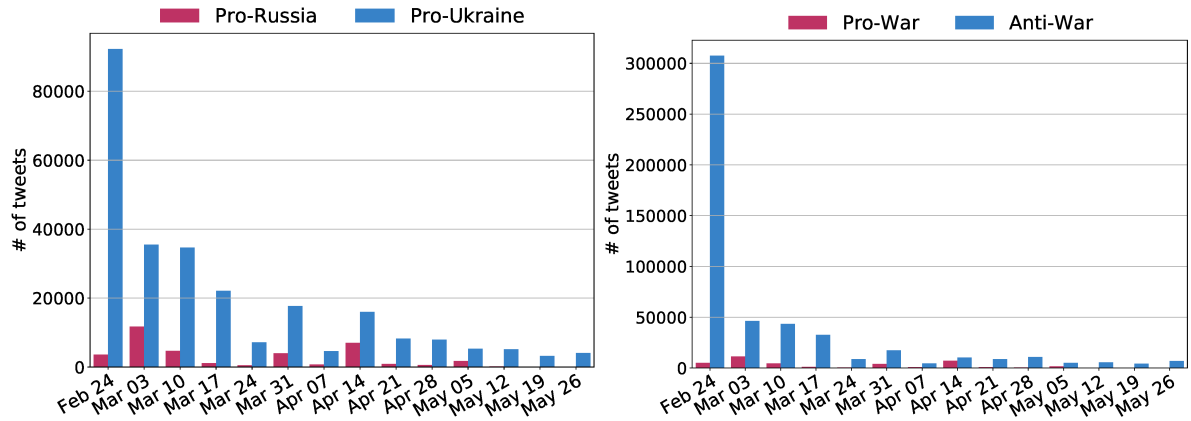

Figure 9: Weekly volume of tweets that contain Pro-Russia/Pro-Ukraine (left) and Pro-War/Anti-War (right) hashtags during the war.

individual Twitter users' stances towards Russia or the war. However, among the search terms we used for data collection, we curated a list of terms that show clear association with certain stances (Pro-Russia/Pro-Ukraine/Pro-War/Anti-War), which can be found in Appendix A. We then measure the volume of tweets that contain such stance-related terms (Figure 9). The results show that Pro-Ukraine and Anti-war tweets are consistently more prominent on Twitter. Russian-speaking Twitter users tend to be more pro-Ukraine and Anti-war compared to Russian residents according to a recent poll result that shows 81% of Russian people sup-

port the Russian military operation in Ukraine.[16] Considering the fact that VK is more widely used inside of Russia than Twitter, our results suggest researchers should exercise caution in generalizing opinions on Twitter to the entire Russian population.

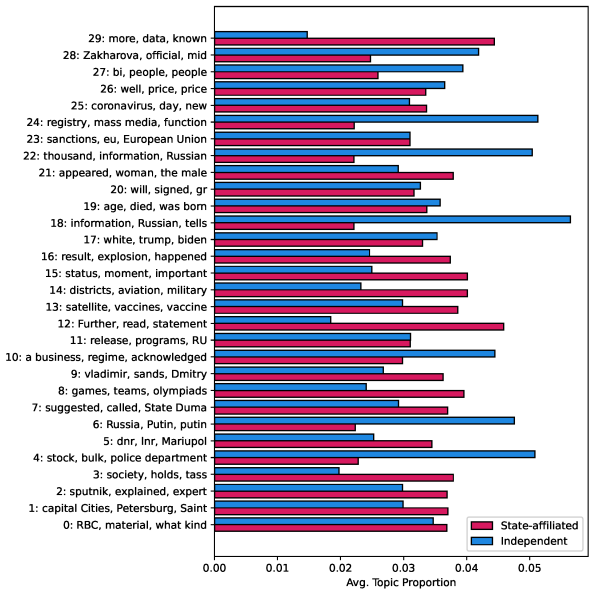

Figure 10: Topic proportions for state-affiliated and independent media outlets, as learned by a 30-topic CTM. The y-axis lists the highest-probable words for each learned topic. The x-axis reflects the estimated topic proportion, averaged across all state-affiliated or independent news outlets.

D Topic Model Parameters and Additional Data

We train the CTM model with 30 topics and take contextualized embeddings from PARAPHRASE-MULTILINGUAL-MPNET-BASE-V2.[17] Contextual embeddings are derived from the first 128 tokens in each document. We preprocess data by removing stopwords and words that occur in $> 99.5\%$ of documents. We train for 50 epochs. We train the STM with 20 topics, and set news outlet affiliation and days since the corpus-collection start as topic-prevalence covariates (e.g., $prevalence = \sim kind * s(date)$). We train for 75 epochs using Spectral initialization. Both models are trained on data through May 15, 2022. For both models, we fixed the number of topics based on which output looked most coherent out of 10, 20, and 30 topics.

Table 8 lists the most probable words in each topic identified by the CTM model and Table 9 lists the same for the STM model. Figures 10 and 11 show topic prevalence as associated with-affiliated or independent outlets. Figure 12, Figure 13, and Figure 14 show topic prevalence over time for the topics most related to Ukraine and the war.

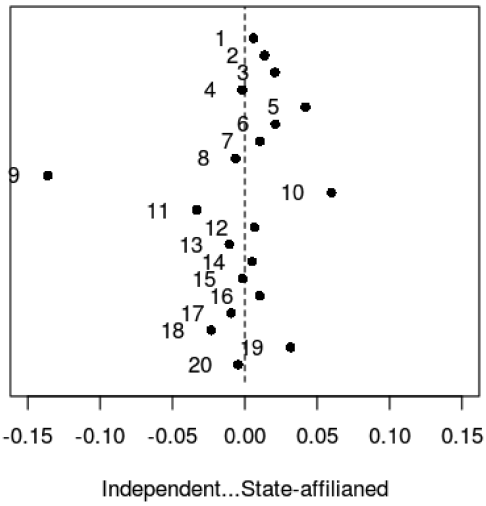

Figure 11: Topic proportion shifts between independent and state-affiliated outlets estimated by a 20-topic STM model, where time and outlet affiliation are incorporated in the model as topic prevalence covariates.

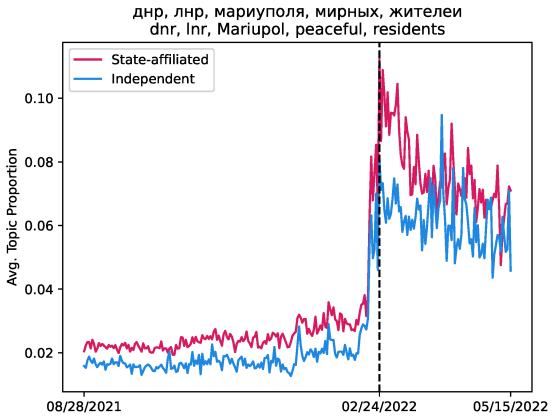

Figure 12: Average estimated proportion of documents related to Topic 5, learned by the CTM model. Topic 5 is the learned topic with the highest probability of Ukraine related terms. Estimated topic prevalence increases sharply in the days preceding the invasion.

[16] https://www.levada.ru/en/2022/04/11/the-conflict-with-ukraine/

[17] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

0	рбк, материале, какие, life, рассказываем, разбираемся, читайте, forbes, бизнес, чаще	RBC, material, what, life, we tell, understand, read, forbes, business, often
1	столицы, петербурге, санкт, столице, москве, днем, утро, друзья, градусов, ожидается	capital, petersburg, st, capital, moscow, afternoon, morning, friends, degrees, expected
2	sputnik, объяснил, эксперт, радио, ру, би, рассказал, рассказала, интервью, си	sputnik, explained, expert, radio, ru, bi, told, told, interview, si
3	общество, проводит, тасс, экономика, ruptly, политика, reuters, премьер, мире, михаил	society, conducts, tass, economy, ruptly, politics, reuters, premier, world, michael
4	акции, навального, овд, инфо, протеста, поддержку, задержанных, алексея, акциях, новои	actions, Navalny, OVD, info, protest, support, detainees, aleksey, actions, new
5	днр, лнр, мариуполя, мирных, жителей, украинские, народнои, новости, донбасса, мариуполе	DPR, LPR, Mariupol, peaceful, residents, Ukrainian, folk, news, Donbass, Mariupol
6	россии, путина, путин, это, функции, иностранного, агента, политолог, выполняющим, почему	Russia, Putin, Putin, this, functions, foreign, agent, political scientist, doing, why
7	предложили, назвали, госдума, предупредили, депутаты, предлагают, доступ, законопроект, новую, предлагает	proposed, named, State Duma, warned, deputies, offer, access, draft law, new, offers
8	игр, сборнои, олимпиады, команды, олимпиаде, спортсменов, чемпионата, золото, команда, победу	games, teams, olympiads, teams, olympics, athletes, championship, gold, team, victory
9	владимир, песков, дмитрии, президент, путин, зеленскии, секретарь, кремль, переговоров, лукашенко	Vladimir, Sands, Dmitry, President, Putin, Zelensky, Secretary, Kremlin, negotiations, Lukashenko
10	дело, режима, признал, приговор, свободы, годам, обвинение, лишения, бывшего, грозит	case, regime, admitted, verdict, freedom, years, accusation, deprivation, former, threatens
11	выпуск, программы, ru, смотрим, программе, шоу, utm, россия, эфир, телеканале	release, programs, ru, watch, programme, show, utm, russia, air, TV channel
12	далее, читать, заявление, сделал, важное, принял, сделали, сделала, приняли, обратился	further, read, statement, did, important, accepted, did, did, accepted, applied
13	спутник, вакцины, вакцину, коронавируса, вакцина, вакцин, вакцинации, воз, омикрон, здравоохранения	satellite, vaccine, vaccine, coronavirus, vaccine, vaccine, vaccination, who, omicron, health
14	округа, авиации, военного, военно, флота, учения, су, военнослужащие, противника, сил	county, air, military, military, navy, exercise, su, servicemen, enemy, forces
15	статус, момента, важных, продолжается, отказались, сказала, выход, собираются, временем, руководство	status, moment, important, continues, refused, said, exit, going to, by the time, leadership
16	результате, взрыв, произошел, борту, погиб, пострадавших, аварии, предварительным, находились, пожара	result, explosion, occurred, board, killed, injured, accident, preliminary, were, fire
17	белого, трампа, байден, трамп, байдена, сша, джо, администрация, энтони, белым	white, trump, baiden, trump, baiden, usa, joe, administration, antony, whites
18	информации, российским, рассказывает, сообщение, массовои, материал, функции, иностранного, иностранным, агента	information, Russian, tells, message, mass, material, functions, foreign, foreign, agent
19	возрасте, умер, родился, скончался, жизни, ссср, рождения, актер, роли, советского	age, died, born, deceased, life, ussr, birth, actor, roles, soviet
20	будут, подписал, qr, смогут, выплаты, правительство, поддержки, закон, должны, могут	will, signed, qr, may, payments, government, support, law, must, may
21	появилось, женщина, мужчина, девушка, instagram, летняя, женщину, мать, ребенка, житель	appeared, woman, man, girl, instagram, summer, woman, mother, child, inhabitant
22	тысяч, информации, российским, рублеи, сообщение, массовои, миллиона, около, материал, млн	thousand, information, Russian, rubles, message, mass, million, about, material, million
23	санкции, ес, евросоюз, против, евросоюза, отношении, пакет, российских, запрет, дипломатов	sanctions, eu, eu, against, eu, regarding, package, Russian, ban, diplomats
24	реестр, сми, функцию, писали, выполняет, требует, иноагентов, нко, иноагента, закон	registry, media, function, wrote, performs, requires, foreign agents, nco, foreign agent, law
25	коронавирусом, сутки, новых, последние, число, случаев, умерли, случая, заболевших, максимум	coronavirus, days, new, last, number, cases, dead, cases, cases, max
26	курс, цена, стоимость, выросла, выросли, цены, вырос, tesla, рост, цен	rate, price, value, up, up, prices, up, tesla, up, price
27	би, люден, люди, это, си, которые, воины, власти, несколько, время	bi, people, people, this, si, which, warriors, powers, several, time
28	захарова, официальныи, мид, мария, представитель, информации, российским, сообщение, массовои, материал	Zakharova, official, mid, maria, representative, information, Russian, message, mass, material
29	подробнее, данные, известно, прокомментировали, обратился, ситуации, выступил, стало, оценили, отреагировали	more, data, known, commented, applied, situations, speak, became, appreciated, reacted

Table 8: Highest probability words for each topic learned by 30-topic CTM model

1	рассказал, далее, интервью, наш, андрей, главный, эксперт	told, further, interview, our, andrey, chief, expert
2	читать, своих, заявили, прокомментировал, слова, эксперт, возможность	read, own, stated, commented, words, expert, opportunity
3	россии, новости, ситуации, россию, заявление, насчёт, газа	russia, news, situations, russia, statement, about, gas
4	изза, ранее, сми, дело, данным, человека, сообщили	because of, earlier, media, case, data, person, reported
5	httpsliferup, видео, области, фото, смотрим, дома, тасс	httpsliferup, video, areas, photo, look, houses, tass
6	страны, глава, сообщил, сергей, мид, новые, стран	countries, head, informed, sergey, mid, new, countries
7	детей, стали, рассказали, результате, известно, погибли, одной	children, became, told, result, known, died, one
8	власти, суд, решение, москвы, связи, мая, такое	authorities, court, decision, moscow, communications, may, such
9	иностранного, агента, функции, выполняющим, информации, российским, иностранным	foreign, agent, function, performer, information, Russian, foreign
10	подробнее, заявил, россия, путин, владимир, президент, александр	more, stated, russia, putin, vladimir, president, alexander
11	рублей, тысяч, навального, новой, января, делу, акции	rubles, thousands, bulk, new, january, deeds, shares
12	россиян, около, coronavirus, мире, тыс, стране, дней	Russians, around, coronavirus, world, thousand, country, days
13	также, которые, могут, будут, компании, пока, которых	also, which, can, will, companies, yet, which
14	москве, дня, стало, февраля, апреля, рассказываем, нашем	Moscow, days, became, February, April, we tell, our
15	сша, против, президента, путина, считает, российского, безопасности	usa, against, president, putin, believes, Russian, security
16	коронавируса, covid, сутки, новых, последние, россии, коронавирусом	coronavirus, covid, day, new, latest, russia, coronavirus
17	года, лет, году, день, несколько, жизни, год	years, years, year, day, multiple, life, year
18	это, почему, людей, люди, очень, рассказывает, своей	this is, why, people, people, very, tells, his
19	украины, украине, российских, минобороны, российские, российской, подробности	ukraine, ukraine, russian, defence, russian, russian, details
20	время, который, словам, которая, которой, михаил, своего	time, which, according to, which, which, Michael, his

Table 9: Highest probability words for each topic learned by 20-topic STM model

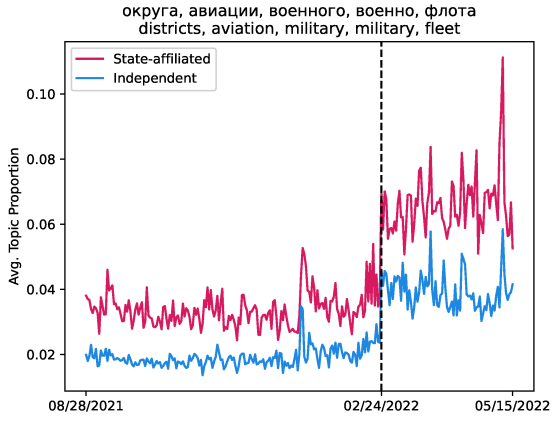

Figure 13: Average estimated proportion of documents related to Topic 14, learned by the CTM model. Topic 14 is the learned topic with the highest probability of military related terms. Estimated topic prevalence increases sharply in the days preceding the invasion.

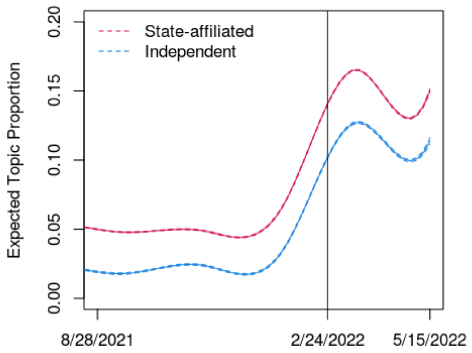

Figure 14: Expected topic proportions of Topic 19 learned by the STM model overtime. Topic 19 is the learned topic with the highest probability of Ukraine and military related terms (e.g., Ukraine, Russian, defense), and expected proportion of this topic increase sharply around the time of the invasion in both state-affiliated and independent news outlets.

E MFC Data Preprocessing & Classifier Training & Label Generation

E.1 MFC Data Preprocessing

We used MFC v4.0 (Card et al., 2015), which contains MFC frame annotations of 27.7k articles across five social issues (death penalty, gun control, immigration, same-sex marriage, tobacco). For each article, the data set provides annotations of two different granularity: article-level and span-level MFC frames. Here, we chose to use span-level annotations to construct sentence-level labels as a sentence is a universal unit of texts; thus the trained models can be applied to both articles and comments in VoynaSlov (Naderi and Hirst, 2017). Also, training models based on large language models with article-level annotation could be challenging, as the cost of training exponentially grows as the input gets longer.

In order to convert the span-level labels in the MFC data to sentence-level, we first mapped each span annotation to its corresponding sentence[18] in the article by finding a sentence that overlaps the most with each span. We then apply a rigorous filtering and kept the sentences and labels that at least two annotators agreed on. Since some articles have only one annotator, this process significantly reduces the number of data samples in the final training data[19]. And for the cross-validation purpose, we randomly split the final training data into 10-folds and used eight folds among them as training, and one as development, and the last one fold as test set. In zero-shot classification scenarios, we held out the data from one social issue as test set and use the ramining data as training and development sets. Table 10 shows the final number of examples in different experiment settings.

Experiment, Data	# train	# dev	# test
In-domain, MFC unfiltered	2.1M	27K	27K
In-domain, MFC	105K	13K	13K
Zero-shot, MFC immigration	100K	11K	21K
Zero-shot, MFC same-sex	102K	11K	19K
Zero-shot, MFC tobacco	106K	12K	13K

Table 10: The number of training instances for each (Experiment set-up, test data) combination. All datasets consist of sentence-level MFC frame labels.

[18] All articles were segmented into sentences using NLTK's sentence tokenizer.

[19] We conducted preliminary experiments with the more extensive, non-filtered data, but the trained models performed significantly worse.

E.2 MFC Classifier Training

Model Prior work has shown that classifiers based on large pre-trained language models achieve state-of-the-art performance. We built our MFC classifiers based on pre-trained language models in light of the findings. We specifically experimented with four different models: Roberta base (Roberta$_B$), Roberta large (Roberta$_L$), and XLM-R base (XLM-R$_B$), and XLM-R large (XLM-R$_L$) model. Roberta$_B$ and XLM-R$_B$ both are the same size (125M parameters), and Roberta$_L$ and XLM-R$_L$ are also the same size (355M parameters). We used 2-layer feedforward neural network, with the hidden size same as the base transformer model, for the final classifier layer. The layer outputs 15-dimension logits to generate multi-class predictions over 15 frame classes. We set the probability threshold for binary label conversion as 0.5.

Hyperparameters The average length of sentences in the MFC sentence-level data is 37, with a standard deviation of 16. We set the maximum length of model inputs to 70 tokens, and truncated the inputs if they were longer than the limit. The learning rate and batch size were set to 5e-6 and 64, respectively. We used the AdamW optimizer with a schedule of linearly decreasing the learning rate after the 5% of warm-up steps. The model was trained with the cross-entropy loss for multi-class classification. The final model for each experiment was selected based on the macro F1 score over the development set. Each model was trained for 20 epochs, and in most cases, the best models were found before reaching ten epochs. In terms of training time, one epoch took less than 7 minutes for smaller models and 10 minutes for larger models when trained on a single GPU machine with A6000. We did not do any hyperparameter search for our experiments, and performed exactly ten runs per experiment for 10-fold cross-validation.

Evaluation Results We computed F1 for each frame category following the standard multi-label evaluation procedure and used the macro averages across classes as our main metric. Table 11 describes the evaluation results of four different baseline models and two different evaluation settings. The Roberta$_L$-based classifier achieves the best performance of 68.1. One of few existing work we can compare against is Naderi and Hirst (2017). They also converted the MFC data to sentence-level labels in a similar fashion and trained sentence-level

	Data	Model	Macro-F1
In-domain	MFC	Roberta$_B$	66.3±0.33
		XLM-R$_B$	64.6±0.34
		Roberta$_L$	**68.1**±0.54
		XLM-R$_L$	67.5±0.53
Zero-shot	Immigration	XLM-R$_L$	52.7±0.36
	Same-sex	XLM-R$_L$	50.4±0.64
	Tobacco	XLM-R$_L$	51.0±0.33
	VoynaSlov	XLM-R$_L$	**33.5**±0.72

Table 11: Macro-F1 results of trained MFC classifiers. We conducted 10-fold cross-validation for each experiment and tested two scenarios: in-domain and zero-shot classification by leaving one social issue (Immigration, Same-sex, Tobacco) out of training data and using it as a test set. We also evaluate the classifiers with VoynaSlov.

MFC frame classifiers based on various LSTM models. The best performance reported in the paper was 53.7 and our model unsurprisingly well surpasses the performance and establishes the new state-of-the-art.

The comparison across models over the in-domain MFC test set shows that multilingual models perform worse than monolingual models with the same number of parameters, which was also evidenced in Akyürek et al. (2020). It suggests that XLM-R enables seamless multilingual transfer at the cost of performance on English data. However, interestingly, the performance drop was lower with the larger multilingual model (XLM-R$_L$), suggesting that practitioners might want to consider large models when using multilingual models.

E.3 Generating MFC labels of VoynaSlov

Akyürek et al. (2020) has examined several strategies of multilingual transfer of frame classifiers, and concluded that translating non-English target data to English and then applying the model (trained with only English data) over the translated text generally achieves the best performance. Following their suggestion, we translated all posts and comments in VoynaSlov to English using the publicly available Russian-English machine translation model[20]. After translating, we apply the final model based on XLM-R$_L$ on each sentence in the translated English text and generate MFC frame labels. Eventually, we are interested in the post-level and comment-level MFC frame labels of texts in VoynaSlov, which might consist of more than one sentence. We aggregated labels of sentences in a post/comment by majority voting with a random tie-breaking (i.e., hard voting) and one final frame label was assigned to each post/comment.

F Human Annotation of VoynaSlov

To measure how well the trained MFC frame classifiers work on VoynaSlov, we annotated a small subset of Russian sentences in our data with the MFC frame labels. Although our trained classifiers generate sentence-level frame labels, in the end, we ultimately want to use the labels of articles/posts/comments for analyses. We thus sampled the examples at the post-level, instead of individual sentence-level; We randomly sampled a post and added sentences in the post to the annotation set until we reached the desired number of sentences (50 each from state-affiliated and independent media). 103 sentences from 49 articles were finally selected. At the time of annotation, we provided the full post (i.e., all sentences in the post) along with the target sentence to annotate[21].

We recruited one native Russian speaker to annotate the sampled Russian sentences with the 15 MFC frame labels. We ensured that the annotator have had past experiences annotating for similar tasks and has good understanding of the concept of framing. We acknowledge that the frame is inherently subjective and having more annotators could have resulted in a better quality evaluation set. However, we collected the in-domain annotation to get a broad sense of the generalizability. During the annotation, we additionally presented the English translation of a target sentence (Appendix E.3) and asked to annotate the quality of the translation to make sure the machine translation models are trustworthy and we can trust the inferred labels. The screenshots of the annotation instruction and the annotation interface are in Figure 15 and Figure 16.

We present label distribution of frame and translation quality labels of 103 annotated sentences in Table 12 and Table 13. The annotation results over the MFC frames show that the MFC frame labels are imbalanced and so does the model performance. The final model performs especially poor on rel-

[20] We used the publicly released model, facebook/wmt-ru-en, downloaded from the HuggingFace model repository.

[21] We acknowledge that there is a discrepancy in amount of available information between our models and the human annotator. We believe the sentence-level annotations that consider the full context reflect what actual readers will perceive from reading the sentence more accurately. We leave incorporating global contexts for the sentence-level frame classifiers for future research.

Frame	#Label	F1
Other	30	6.5
Political	24	38.7
Health and Safety	14	62.9
Crime and Punishment	12	60.0
Security and Defense	11	54.6
Legality, Constitutionality, Jurisdiction	9	16.7
External Regulation and Reputation	9	33.3
Capacity and Resources	6	0
Quality of Life	4	28.6
Economic	2	33.3

Table 12: The proportion of MFC frame labels in the annotated VoynaSlov data. F1 indicates the macro-F1 score of the final model measured for each frame class. The five frame classes that are not in this table did not attain any annotation label.

Translation Quality	#Label
Good	59
OK	43
Bad	1

Table 13: The proportion of MFC frame labels in the annotated VoynaSlov data.

atively abstract frames such as Capacity and Resources, Other, and Legality, Constitutionality. On the other hand, the translation quality annotations suggest that the most of generated English translations of VoynaSlov (99%) are either good or decent enough to maintain the frame label in the original Russian text (i.e., OK).

1 Instructions

The data you are annotating consists of VK posts made by media accounts from 2021 Jan-2022 May. For each post, and each sentence in the post, we'll ask you to mark the framing used in the sentence, according to following 15 categories:

Category name	Description	Example
Economic	Financial implications of an issue	"IRS issues tax rules for married gay couples"
Capacity & Resources	The availability or lack of time, physical, human, or financial resources	"Global warming may leave West in the dust"
Morality & Ethics	Perspectives compelled by religion or secular sense of ethics or social responsibility	"More lives sacrificed on altar of Second Amendment"
Fairness & Equality	The (in)equality with which laws, punishments, rewards, resources are distributed	"Democracy means equality for all."
Legality, Constitutionality & Jurisdiction	Court cases and existing laws that regulate policies; constitutional interpretation; legal processes such as seeking asylum or obtaining citizenship; jurisdiction	"U.S. court panel rules in favor of D.C. gun law"
Crime & Punishment	The violation of policies in practice and the consequences of those violations	"Texas Inmate Executed for Killing"
Security & Defense	Any threat to a person, group, or nation and defenses taken to avoid that threat	"To Deter Terror, Show No Mercy"
Health & Safety	Health and safety outcomes of a policy issue, discussions of health care	"FDA to ban many stop-smoking products"
Quality of Life	Effects on people's wealth, mobility, daily routines, community life, happiness, etc.	"Chaos, not closure, for family of murder victim"
Cultural Identity	Social norms, trends, values, and customs; integration/assimilation efforts	"Why do some people think America has a gun problem?"
Public Sentiment	General social attitudes, protests, polling, interest groups, public passage of laws	"Support Down in Poll On Gun Restrictions"
Political	Focus on politicians, political parties, governing bodies, political campaigns and debates; discussions of elections and voting	"Urged by Right, Bush Takes On Gay Marriages"
Policy Prescription & Evaluation	Discussions of existing or proposed policies and their effectiveness	"Proposed gun control laws are a farce"
External Regulation & Reputation	Relations between nations or states/provinces; agreements between governments; perceptions of one nation/state by another	"Halts Cuban Immigration Talks; Worsening of Ties Seen"
Other	All other sentences that do not belong to any of the above categories	

Please note that **there could be more than one framing label** used in each sentence. Also, for some sentences, we don't ask you to annotate anything as they are less likely to be using any framings at all.

For each sentence, focus on the primary framings that the tweet expresses, rather than all possible emotions (e.g. most sentences involve only 1-2 framings, and the average is 1.2).

We additionally present English translations of each sentence, and ask you **evaluate how well they are translated. There are three options**:

- **Good**: the English translation keeps most of the original sentence's meaning, and thus framing does not change by translating.
- **OK**: the translation is not perfect, but not bad enough to change the framing entirely.
- **Bad**: the translation is so bad that the meaning is significantly altered, and primary framing label of the translated text is different from the one from original sentence.

Annotations will be automatically saved each time you click "submit", so you don't have to worry about saving your results.

Figure 15: Screenshot of annotation instruction provided to the annotators. We borrowed the description from Mendelsohn et al. (2021) and the examples were selected from the sentence-level MFC training data we constructed.

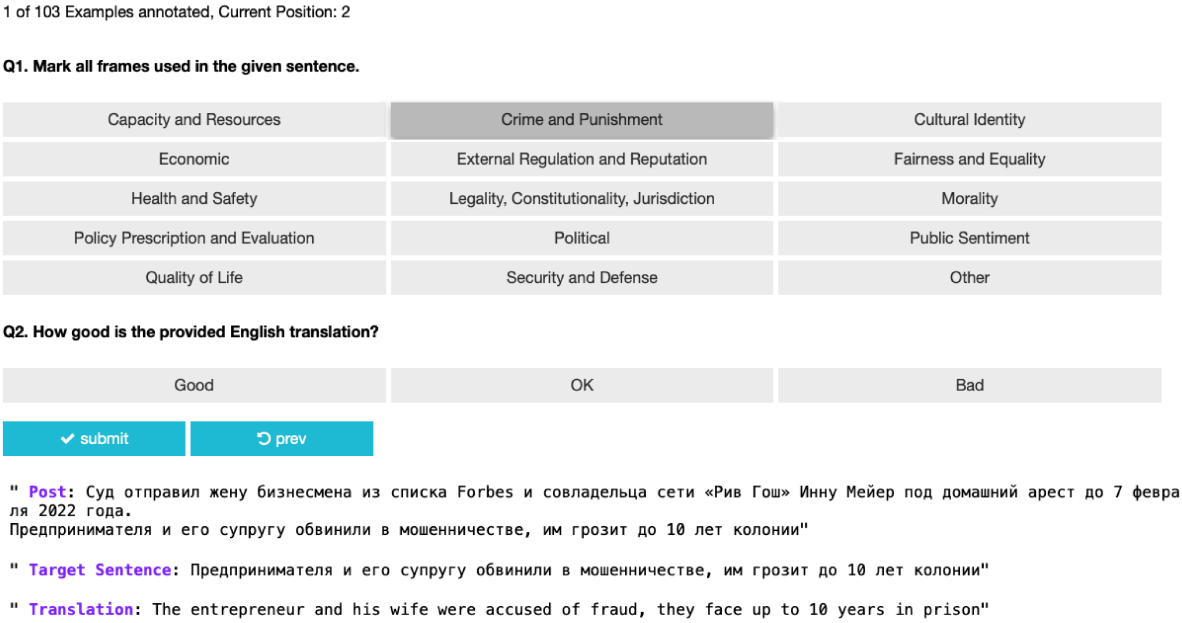

Figure 16: Screenshot of annotation user interface. For each target sentence, we provide the full context (*Post*) and its English translation. Annotators mark their answers to two questions regarding the MFC frame and translation quality.