Human Computation (2024) 11:1:1-28

ISSN: 2330-8001, DOI: 10.15346/hc.v11i1.1

Framing is Mightier than the Sword: Detection of Episodic and Thematic Framing in News Media

PANAGIOTIS MAVRIDIS, DELFT UNIVERSITY OF TECHNOLOGY

OANA INEL, DELFT UNIVERSITY OF TECHNOLOGY

XANDER WILCKE, VRIJE UNIVERSITEIT AMSTERDAM

MYKOLA MAKHORTYKH, UNIVERSITY OF BERN

MARKUS DE JONG, VRIJE UNIVERSITEIT AMSTERDAM

HONORATA MAZEPUS, LEIDEN UNIVERSITY

ANTOANETA DIMITROVA, LEIDEN UNIVERSITY

JESSE DE VOS, NETHERLANDS INSTITUTE FOR SOUND AND VISION

ALESSANDRO BOZZON, DELFT UNIVERSITY OF TECHNOLOGY

TOBIAS KUHN, VRIJE UNIVERSITEIT AMSTERDAM

ABSTRACT

Many people today watch news videos to get informed. However, news videos can frame information differently and be prone to bias, which might lead to miscommunication. Bias is ubiquitous and inherent in interactions between news consumer groups, but framing can introduce additional bias in news communications. For example, citizens who interpret the news have different political orientations and, thus, understand them differently. Experts can be more capable of detecting biased or differently framed information. However, the ever-increasing amount of news videos also makes it difficult for experts alone to analyze. While automated methods exist for identifying different types of bias, frame detection approaches, namely episodic and thematic framing, are scarce and focused on texts. In this work, we address the issue of scalable thematic and episodic frame detection in news videos through crowdsourcing and machine learning techniques. We design a crowdsourcing task for annotating thematic and episodic framing in videos with the help of domain experts in political and social sciences. We then use the annotations from experts and crowds to investigate whether machine learning methods can scale up the annotation process by automatically labeling videos on episodic and thematic framing. Our results indicate that framing analysis is challenging for both humans and machines, with high disagreement amongst experts and crowd annotators. Nevertheless, our results suggest that machine learning has potential by combining crowd and expert annotations and building upon them a classifier.

1. INTRODUCTION

In today's "high-choice" (Van Aelst et al., 2017) information environment, news is increasingly distributed and consumed online. This ease of production and dissemination of digital news exposes the audience to viewpoints coming from a wide range of actors, which includes traditional media (e.g., news channels such as Al Jazeera or CNN), as well as citizen journalists (e.g., bloggers or influencers), where citizen journalism refers to news items produced by non-professionals (Wall, 2015). This change has significant implications for framing societal matters, such as making some of their aspects more salient or promoting a particular interpretation of the subject (Entman, 1993). It is, hence, important to understand how a specific news subject is presented, especially considering that certain ways of presentation or framing are more suitable for manipulating public opinion.

Framing shapes and is shaped by the different perceptions individuals and societies have. Frames can transmit biases that potentially result from established cultural and societal practices, including tactical opinion manipulations (Entman, 2007; Morstatter et al., 2018). Such biases are common for polarizing subjects, which are presented and interpreted differently within different communities. For instance, international crises stir prominent emotional reactions from audiences and are framed differently by various groups involved in them (Makhortykh and Sydorova, 2017; Pantti, 2016; Fengler et al., 2020). The typology of thematic and episodic framing (Iyengar, 1996) distinguishes between frames that depict an issue by putting it into a broad context (thematic) or specific instances or cases (episodic).

Political communication research considers episodic framing, which involves personal stories and experiences, to have a broader impact by provoking a stronger affective reaction. Thematic framing might evoke less emotional responses (Aarøe, 2011). Detecting the frame types used in a news story is essential for revealing its intended impact and affective strength. Thematic and episodic frames are crucial for communicating news, and the differences between them can also lead to a biased representation of the framed subject. News consumers can perceive episodic framing of a subject as a very particular case, and they might not understand its depth or spread. Conversely, thematic framing depicts content in a general or abstract manner, and news consumers might not be able to transpose themselves into the issue and see the connections. Thus, frame detection is integral for measuring how balanced the media representation of a particular subject is and countering possible manipulations.

Unlike news credibility (Bhuiyan et al., 2020) and fake news detection (Stefanone et al., 2019; Flintham et al., 2018), which have been studied extensively, less research has been done on framing bias in online news (for an exception, see (Morstatter et al., 2018)). Hitherto, news media framing, and, more specifically, thematic and episodic framing, have been less studied so far. Frame identification is a knowledge-intensive task, typically performed by experts in political and social sciences (Aarøe, 2011; Boräng et al., 2014; Cremisini et al., 2019). Given the ever-increasing volume of news, expert frame identification suffers from scalability issues (Wijekoon et al., 2019). Crowdsourcing proved to be an effective tool to annotate news articles' credibility (Shariff et al., 2014), trustworthiness (Castillo et al., 2013) or bias (Iyyer et al., 2014; Lim et al., 2020). However, to the best of our knowledge, crowdsourcing approaches for framing analysis have not yet been studied. Furthermore, framing analysis studies are typically performed on texts (Aarøe, 2011; Boräng et al., 2014), while news videos are less researched (Dimitrova et al., 2017). Computational methods to automatically detect frames are also more widely used for texts (Park et al., 2011). Nevertheless, considering the complexity of classifying frames (Entman, 1993), it is crucial to keep humans "in-the-loop" and study how to combine crowdsourcing and automated approaches to detect frames.

In this paper, we address the problem of scalable thematic and episodic frame detection in news videos through crowdsourcing and machine learning. Our case study is an international conflict, the Crimea Crisis, which was selected due to the political relevance of the annexation of Crimea by Russia, the most dramatic violation of international borders since the end of the Second World War. How this conflict is understood and presented by the media is crucial for international politics. Considering that the choice of frames can facilitate or impede the mobilization of unpopular political decisions (e.g., economic sanctions or military interventions) during such crises, it is necessary to understand how news media uses thematic and episodic frames to represent them. Thus, we emphasize that the importance of being able to detect episodic and thematic frames relates to these types of frames having different impacts on the audience's opinions about the issue which is being framed (Gross, 2008; Aarøe, 2011). It is of particular relevance to detect episodic and thematic framing when reporting on polarising subjects such as armed conflicts or civil rights (Khaldarova and Pantti, 2016).

We seek to answer the following research question: *"To what extent can crowdsourcing and machine learning effectively be employed to identify thematic and episodic framing in video news?"*, by employing a cone-shaped experimental methodology. We first employ experts in social and political sciences, with vast experience in conducting research in the area of framing analysis, to annotate a small number of videos and to help us define a suitable annotation template. Then, we use this annotation template to annotate more videos using crowdsourcing. Finally, we use these results to train various classifiers and investigate whether we can use machine learning methods to scale up the annotation process. Through continuous collaboration among domain experts, citizens, and machine learning techniques, we provide the initial necessary means to analyze biases and framing in news at scale, which can further facilitate the development of tools, frameworks, and metrics to reliably quantify and evaluate bias, diversity, or fairness in news media. This paper makes the following contributions:

- an annotation study to identify episodic and thematic framing in news videos about the Crimea Crisis; designed with social and political scientists specializing in Eastern Europe;
- a quantitative analysis and discussion of expert and crowd annotations to understand the complexities of identifying frames in news;
- a comparison study to investigate the ability of various classification models to detect framing type in news videos, using expert and crowd labels;
- a dataset of 120 videos about the Crimea Crisis from various news channels annotated by three experts (a subset of 58 videos) and crowd annotators (N = 338) with frames, sentiment, and trustworthiness values; for political science scholars, an important open question is whether particular frames make news more trustworthy for recipients or evokes specific sentiments (Park, 2012).

2. RELATED WORK

We first review existing work on framing analysis, particularly on episodic and thematic framing. Then, we present an overview of human-centered approaches for framing analysis, followed by automatic techniques. We conclude with our contribution to framing analysis.

2.1. Framing analysis

Introduced by Goffman (1974), the concept of framing attracted critical acclaim in social sciences, being referenced as *“the most utilized mass communication theory of the present era”* (Bryant and Miron, 2004). Framing denotes how the presentation of a specific subject, e.g., anti-abortion legislation, influences the audience’s judgments and choices. Framing is realized via individual frames, i.e., distinct patterns of representation and interpretation that highlight some aspects of an issue and omit or play down others (Entman, 1993), distributed through mass media. While historically focused on texts, such as newspaper pieces or politicians’ speeches, framing is increasingly adopted to analyze visual content, mainly regarding armed conflicts and disasters where images serve as powerful means of communicating complex phenomena (Bleiker, 2018) and mobilizing the audience (Makhortykh and González Aguilar, 2020).

Multiple typologies of frames were proposed to facilitate framing research: generic/issue-specific (De Vreese, 2005), human impact/powerlessness/economics/moral values/conflict (Neuman et al., 1992), thematic/episodic (Iyengar, 1994). We focus on the latter typology that distinguishes between frames depicting issues via specific instances or cases, i.e., episodic frames, and frames depicting issues by putting them into a broader context, i.e., thematic frames. Iyengar (1996) illustrates them using news reporting on poverty: the story about an individual ending up on the street as episodic framing and the report on recent trends in poverty rates as thematic framing. Episodic frames tend to stir stronger emotional reactions by referring to individual cases and stories that resonate with the audience (Gross, 2008; Aarøe, 2011). Both types of frames are important for public mobilization, but thematic frames are more effective when no emotional reaction is elicited (Aarøe, 2011) or when the greater treatment responsibility is attributed to the authorities (Hart, 2011).

Several studies examined episodic and thematic framing in healthcare (Kang et al., 2017), natural disasters (Muralidharan et al., 2011), or political issues (Semetko and Valkenburg, 2000; Dimitrova, 2006). While many approaches focused on printed and digital press (Muralidharan et al., 2011; Papacharissi and de Fatima Oliveira, 2008; Dimitrova, 2006), online videos and TV news (Kang et al., 2017; Semetko and Valkenburg, 2000; Dimitrova et al., 2017) were less researched. Episodic and thematic frames differ in terms of coverage when looking across media outlets (Papacharissi and de Fatima Oliveira, 2008) and when comparing online videos with TV news and newspapers (Kang et al., 2017). Concerning terrorism issues, U.S. newspapers cover more episodic frames and U.K. newspapers more thematic frames (Papacharissi and de Fatima Oliveira, 2008). Similarly, online videos on Attention Deficit Hyperactivity Disorder (ADHD) contain more thematic frames than episodic frames, in contrast to TV news and newspapers (Kang et al., 2017).

2.2. Human annotation for framing analysis

Crowdsourcing is frequently used for evaluating content veracity, such as identifying deceptive opinions (Ott et al., 2011), controversial issues (Mejova et al., 2014), or fake information (Sethi, 2017). Other studies deploy crowd workers to investigate the degree of credibility of news content (Shariff et al., 2014), of newsworthiness (Castillo et al., 2013) or the degree of bias (Iyyer et al., 2014; Lim et al., 2020).

Compared to veracity evaluation, framing analysis through crowdsourcing received less attention. Most studies using human annotations to detect frame types come from political science (Boräng et al., 2014; Aarøe, 2011) and typically rely on one (Cremisini et al., 2019) or multiple annotators (Kang et al., 2017; Boräng et al., 2014; Burscher et al., 2014) with social or political sciences backgrounds. In (Cremisini et al., 2019), a single expert annotated 227 news sources about the annexation of Crimea from 43 countries, using a pro-Russia, pro-Western, and neutral frame typology. In (Burscher et al., 2014), two trained coders annotated 156 political news articles with four frame types (conflict, morality, economic consequences, and human interest) and achieved rather low Krippendorff's α, between 0.21 (morality) and 0.58 (economic consequences).

Text-based framing analysis received considerable attention in contrast to TV news or videos. Dimitrova et al. (2017) performed manual annotation of approximately 600 TV news broadcasted in Belarus, Moldavia, and Ukraine to understand how the Russian and EU influences are framed in terms of subjects, tone, topic, and theme. Kang et al. (2017) analyzed 685 videos on YouTube about ADHD to study episodic and thematic framing (Cohen's κ of 0.83 on 70 videos).

2.3. Automated framing detection methods

Computational methods for episodic and thematic framing detection are scarce, mainly focused on text and supervised approaches. Guo et al. (2021) proposed a computational method using BERT (Devlin et al., 2018) to detect episodic and thematic framing in news headlines (0.95 precision, 0.89 recall). More often, however, research focused on identifying generic news frames. Burscher et al. (2014) proposed two supervised machine learning approaches to automatically code four generic news frames (conflict, economic consequences, morality, and human-interest) with accuracy varying for each frame, from 0.74 (human interest) to 0.89 (morality). To discriminate between ten generic new frames in online news items about the Iraq war, Morstatter et al. (2018) found that a simple linear regression classifier yielded the best results, irrespective of language, whereas a deep LSTM (Graves and Schmidhuber, 2005) performed poorly. Likewise, a linear regression model which uses NLP-generated features as input gave the best results in (Opperhuizen and Schouten, 2020), where authors use expert-annotated frames in Dutch news articles about gas drilling were used to train a classifier (average F1-score varies between 0.718 to 0.889 among 14 frames of interest). Differently, Saez-Trumper et al. (2013) employed an unsupervised method to investigate whether clustering methods can help analyze different types of bias (selection, coverage, and statement bias) in social media news posts from different countries. The authors identified geopolitical influences by deriving basic statistics and projecting these into two dimensions using PCA.

Cremisini et al. (2019) developed a model for identifying the main frame (i.e., pro-Russia, pro-Western, neutral) in news articles covering multiple sources in different countries about the annexation of Crimea. They found, however, that the model learns the regional journalistic style rather than the actual frame. Similarly, Baumer et al. (2015) focused on understanding the type of language perceived as framing-related in texts. Their proposed classifiers achieved comparable results to those of human annotators. Field et al. (2018) used lexical classifiers to predict, across languages, the main article frame using the framing dimensions of the Media Frames Corpus (Card et al., 2015), e.g., political, economical, among others. The NewsCube2.0 platform (Park et al., 2011) automatically identifies possible biases in socially contentious issues and uses frames to present different media viewpoints. NewsCube2.0, however, faces scaling challenges, needing mass participation to identify frames and automate the process.

2.4. Summary of contributions

We focus on identifying and detecting episodic and thematic frames in online videos about the Crimea Crisis, a recent political, international issue. The Crimea Crisis is a suitable case study for framing analysis because reporting on it is led by two completely different and rather incompatible understandings of what happened during the annexation, represented by two contrasting discursive frames. Russia sees it as a peaceful reunification of a region that used to be Russian and where citizens voted to rejoin the Russian federation. Episodic frames and personal stories were used to create such understandings. The EU and Ukraine regard the actions of Russia as an annexation and the Crimean referendum on joining Russia as illegal. Thematic frames focusing on international law and historic agreements on post World War II borders convey this better. Differently from existing approaches, we perform framing annotation both through experts, well-established practice, *and* crowds, which have not been employed so far. We also explore whether machine learning can be used to automatically detect episodic and thematic framing based on these annotations and scale up the annotations, which, to the best of our knowledge, has not been attempted yet on news videos. We analyze audio transcriptions and metadata (title, description, tags) of the videos. Arguably, audio transcriptions pose different linguistic and syntactic challenges than news headlines due to different narration styles in news videos, e.g., reporters use first-person statements, and there can be multiple “narrators”, i.e., a reporter and several interviewees (Bock, 2016).

3. CRIMEA CRISIS CASE STUDY

We first introduce and motivate our case study, the Crimea Crisis, and then describe the video dataset used in our studies.

3.1. Use Case: The Crimea Crisis

The Crimea crisis in February 2014 followed months of civil unrest in Ukraine. Mass protests, that became known as the EuroMaidan, began on November 21, 2013. Activists and citizens in Kyiv and other cities across the country protested the sudden refusal of the pro-Russian President Yanukovych to sign the Association Agreement with the European Union (Onuch, 2015). Ukraine had negotiated the Agreement with the EU for seven years when, under pressure from Russia, Yanukovych refused to sign it in November 2013.

After several phases of protests and increasing violence from the state security forces resulting in more than a hundred deaths, President Yanukovych was ousted and fled to Russia on February 21, 2014. Immediately after this, the newly installed pro-European interim government in Ukraine faced demonstrations from Yanukovych's pro-Russian supporters across the eastern and southern regions of Ukraine – Donbas and Crimea. Using a rhetoric of protecting the rights of Russian co-ethnics aboard (Nitsova et al., 2018), the Russian military assisted Yanukovych's supporters. Russia got increasingly involved in the separatist rebellion in Donbas in the months to follow (Nitsova, 2021). Meanwhile, the Russian military stationed on the Crimean peninsula helped pro-Russian groups to capture the Crimean parliament on February 27, 2014. The pro-Russian groups installed a new government in Crimea and took control of Ukraine's military and security installations. Following a blockade of Ukrainian army units in their bases and a wave of attacks against pro-Ukrainian activists (Korostelina, 2015; Kofman et al., 2017), Russian authorities helped the local legislative organ in Crimea to organize a referendum on Crimea's independence from Ukraine on March 16, 2014. Disregarding international criticism concerning the legality and validity of the referendum, Russian sources reported an overwhelming vote in favor of the independence of the peninsula from Ukraine. Immediately after, the Russian authorities recognized Crimea's secession and almost at the same time admitted the Republic of Crimea into the Russian Federation, de facto annexing the region (Grant, 2015).

The choice of the Crimea Crisis as the use case is motivated by several reasons. First, it is a starting point of a major international crisis in Europe, and thus, a subject of intensive media framing. Similar to the framing of the Ukraine crisis, characterized by a deep divide between the way it is reported by Western/Ukrainian media and Russian media (Boyd-Barrett, 2017; Makhortykh and Bastian, 2020), the Crimea Crisis was interpreted differently by the sides involved in the crisis. News media in Russia presented it as a benign act of protecting the Crimean people from the so-called "Nazi" (Makhortykh, 2018) government in Kyiv, whereas Ukrainian and Western media interpreted it as an illegal action, breaking international law and violating human rights (Aydin and Sahin, 2019; Biersack and O'lear, 2014).

Second, because of the confrontational nature of the crisis, its framing in the West and Russia plays an influential role in mobilizing the public to support potentially unpopular decisions (e.g., the use of armed forces in Russia and economic sanctions in the West). Considering the importance of episodic and thematic frames for shaping public opinion, the understanding of framing strategies used by international news outlets is of paramount importance. Whether geopolitics (thematic framing) or individual life stories (episodic) present the conflict, these strategies affect the public perception of the events in Crimea and raise support or resistance to political decisions concerning the crisis.

Dataset	#Videos	Duration (sec.)			#Channels	#Videos per Channel		Upload Date	
		Avg.	Min.	Max.				Start	End
Crowd	120	154	133	180	6	Al Jazeera English	63	February 2014	April 2014
						RT	28		
						CNN	23		
						BBC News	4		
						DW English	1		
						FRANCE 24 English	1		
Expert	58	166	153	180	5	Al Jazeera English	18	February 2014	April 2014
						RT	15		
						CNN	10		
						BBC News	4		
						FRANCE 24 English	1		

Table 1. Dataset overview.

3.2. Dataset

In our framing analysis experiments, we used videos in English related to the Crimea Crisis topic. The videos were selected from the YouTube-8M Dataset[1] using the keyword "Crimea". Table 1 provides an overview of the dataset. We randomly selected 120 videos (row *Crowd*) published on well-known, popular YouTube news channels (Al Jazeera English, BBC News, CNN, DW English, FRANCE 24 English, and RT), categorized as "News & Politics", based on YouTube's categorization scheme, in the first months of the crisis. We chose videos with lengths between 2 and 3 minutes to *(1)* keep the annotation task within reasonable length and cost, *(2)* avoid annotator fatigue (Dai et al., 2015), and *(3)* comply with recent research (Wu et al., 2018; Lopatecki et al., 2019) acknowledging the decline in people's attention span when watching online videos.

A subset of 58 videos randomly selected from the 120 videos, the *Expert* row in Table 1, was annotated by three experts, with a background in social (one expert) and political (two experts) sciences. We use this dataset to understand the difficulty of performing framing annotation in videos and to evaluate the crowd performance on this task.

4. METHODOLOGICAL APPROACH

The proposed methodological approach, depicted in Figure 1 and detailed below, consists of three sequential parts: framing analysis performed by experts, framing analysis performed by crowd annotators, and machine learning for framing classification.

Expert annotations: were acquired from one social science and two political science scholars to avoid ties. All three experts have vast experience in conducting research in the area of framing analysis and have a good understanding of our use case, the Crimea Crisis. They annotated episodic and thematic frames in online news videos about the Crimea Crisis. The annotation process was

[1] https://research.google.com/youtube8m/

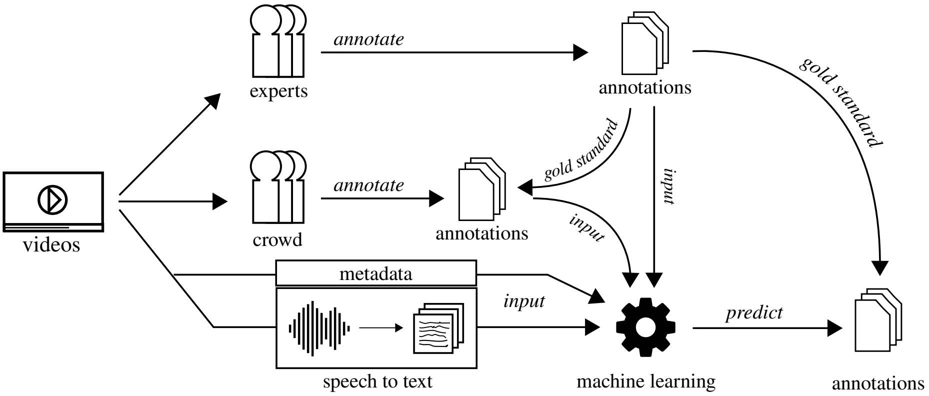


Figure 1. ***High-level overview of our methodology. Expert annotations are used as gold standard for crowd annotations, both of which are used as input for machine learning (as target labels) together with video transcriptions (as input samples) and metadata. We validate the predicted annotations by the machine learning model against the expert annotations.***

the foundation for deriving the dependent variables and the questions in the annotation task. We conducted several discussion rounds with the experts to improve the annotation task.

Crowd annotations: were obtained from crowd annotators who do not necessarily have a political or social science background. Participants were given instructions together with example videos and external sources with information about framing. Then, they answered the questions that we derived together with our experts.[2]

Machine learning: is used to investigate whether classification can scale up the annotation process where manual annotations are expensive or difficult to obtain. We used video transcriptions and metadata as input and expert and crowd annotations as targets. We test and evaluate several different models and configurations against a majority class classifier.

4.1. Expert and crowd annotation studies

We conducted two studies for identifying episodic and thematic framing in videos: the first study with experts (i.e., *expert study*) and the second study with crowd workers (i.e., *crowd study*). We present the studies together and highlight differences between them.

4.1.1. Materials

In the *crowd study*, we used the 120 videos described in Section 3.2 (see row *Crowd* in Table 1). In the *expert study*, we used a subset of 58 videos from this dataset (see row *Expert* in Table 1), randomly selected. All videos have similar characteristics. We use the *Expert* dataset to understand the difficulty of performing video framing annotation and to evaluate the performance of the crowd.

4.1.2. Procedure and Participants

Table 2 shows the questions in the annotation study and their answer space.

[2] We obtained the approval of the ethics committee of one of the institutions involved.

In the *expert study*, three project collaborators with a background in social (one expert) and political sciences (two experts) provided annotations independently and voluntarily. They first watched a video, and then they answered the *Framing* questions reported in Table 2. We defined the questions through several discussions with them, so no instructions were needed.

In the *crowd study*, the crowd annotators first read definitions and guidelines on how to differentiate between episodic and thematic framing. We provided an example video for each framing type and external resources for further reference. Then, they were asked to watch a video and answer the *Framing* questions in Table 2 about the video, one *Demographics* question, two *Control* questions, and two *Experience* questions. The crowd annotators could leave comments regarding their annotation experience for each annotated video. We required a minimum of nine annotations per video, after filtering out the low-quality annotations.

To maximize diversity among crowd annotators, we recruited *crowd annotators* on both FigureEight[3] and Amazon Mechanical Turk[4]. We first ran the study on FigureEight, filtered out the low-quality annotations, and then ran the study on MTurk. On FigureEight, we used the following quality control mechanisms to maintain high annotations' quality: *(1)* level three annotators[5]; *(2)* minimum work time of 6 minutes, i.e., the minimum time needed to watch a gold video and a normal video; *(3)* dynamic judgments set to 0.7 (i.e., we stop collecting judgments for a video when the confidence is above 0.7 or after collecting the desired number of judgments); *(4)* gold questions to filter untrusted annotators at run-time (i.e., gold questions were videos already annotated by experts) and *(5)* control questions to post-filter untrusted annotations (see *Control* questions in Table 2). On MTurk, we did not use gold questions, but we selected annotators with *(1)* HIT approval rate greater than 98%, *(2)* number of HITs approved greater than 500, and *(3)* located in English-speaking countries.

4.1.3. Independent Variables

We have two main independent variables in our studies, referring to annotators, namely *expert* and *crowd* annotators. We also consider the video channel as independent variable. In our channel analysis in Section 5.2, we study the three channels with the most videos, namely Al Jazeera English, RT, and CNN. Nevertheless, we still collect both expert and crowd annotations on the videos posted on the other channels.

4.1.4. Dependent Variables

In the *expert study*, we measure the *inter-rater agreement* (IRR) among *experts* to understand the underlying difficulty of such a complex annotation task. For all videos, we measure the agreement on the following variables: dominant frame, framing score, frame selection modality, sentiment label, sentiment magnitude, and trustworthiness, where trustworthiness refers to how reliable the content of the news video is perceived by study participants. Perceived video sentiment and trustworthiness provide insights regarding media coverage (Park, 2012). More precisely, in some political science debates and political communications, scholars investigate whether the use of episodic or thematic frames makes news more trustworthy for recipients or evokes specific sentiments.

We measure the *agreement* between the experts and the crowd on the dominant framing and framing score to understand whether crowd annotators can perform video framing analysis. Differently than

[3] Currently known as Appen (https://appen.com)

[4] https://www.mturk.com

[5] Best performing annotators c.f. FigureEight

	Variable	Answer Space	Type	Expert	Crowd
1	Annotator gender	male/female	Demographics		✓
2	Video title	Free input	Control		✓
3	Video type	fashion/world news/sports/music	Control		✓
4	Framing scale	7-point scale (1: Clearly Thematic, 4: Balanced, 7: Clearly Episodic)	Framing	✓	✓
5	Dominant frame	Thematic/Episodic(/Balanced)[6]	Framing	✓	✓
6	Frame selection modality	images/words/images and words	Framing	✓	✓
7	Sentiment present	Yes/No	Framing	✓	
8	Sentiment label	Positive/Negative/None (Neutral)	Framing	✓	
9	Sentiment magnitude	5-point scale (1: Very Negative, 3: Neutral, 5: Very Positive)	Framing	✓	✓
10	Video trustworthiness	7-point scale (1: Very Untrustworthy, 4: Neither, 7: Very Trustworthy)	Framing	✓	✓
11	Use of extra materials	Yes, they were necessary / Yes, I was curious and wanted to know more / No, the instructions were very clear	Experience		✓
12	Annotation experience	Free input: paragraph	Experience		✓

Table 2. ***Overview of questions included in the*** **expert study** ***and the*** **crowd study** ***for annotating episodic and thematic framing in videos. The table contains the dependent variable, the answer space, the variable type, and shows whether we measured the variable in the*** **expert study** ***and the*** **crowd study.**

in the *expert study*, in the *crowd study*, the participants can annotate a news video as episodic, thematic, or balanced. We added the option *balanced* in the crowd study to capture cases where annotators perceived the news video as both episodic and thematic. In crowdsourcing studies, it is common to offer annotators an option they can choose when they are either uncertain or they feel like all options or none apply. In addition to this question, in our survey, we have a question where we require a more granular answer in terms of perceived episodic or thematic framing, on a 7-point Likert scale. In the *crowd study*, we also measure the *task experience* and the *clarity* of the instructions.

In both studies, based on the video trustworthiness value, for each *video channel*, we measure the aggregated *trustworthiness* of all the videos from that channel (i.e., the overall trustworthiness of the video channel). Per video channel, we also measure the *frequency* of thematic and episodic framing. Additionally, we measure the *sentiment* per *video channel* and per *framing type*. Finally, we measure the correlation between the type of framing and the framing selection modality to understand whether speech (i.e., what is said in the video) or visual aspects (i.e., what is shown in the video) are more important to identify a framing type.

4.2. Machine Learning Annotations

The goal of our machine learning experiments is to investigate whether supervised classification can scale up the annotation process when human annotations are expensive or difficult to obtain, by predicting if news videos contain primarily episodic or thematic framing (Figure 2). The input to these classifiers are audio transcriptions and textual metadata such as titles, descriptions, and tags. We abstain from using the raw video themselves as input since our dataset is insufficiently large to

[6] The option *balanced* was only used in the *crowd study*.

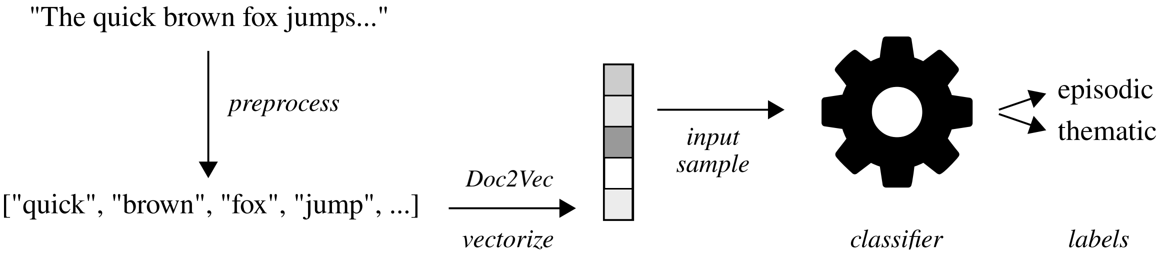


***Figure 2.** Simplified depiction of the machine-learning framework. Preprocessed transcriptions and metadata are vectorized using Doc2Vec and used as input samples to train a classifier. Expert and crowd annotations are used as labels.*

support the additional model complexity that this would entail (See Section 6). We use the experts and crowd annotations on dominant framing type (i.e., episodic or thematic) as labels.

4.2.1. Model Selection

We test five classification models: ridge regression (Suykens and Vandewalle, 1999), Gaussian naive Bayes (John and Langley, 2013), random forest (Ho, 1995), support vector machines (SVM) (Drucker et al., 1997), and neural networks (Rosenblatt, 1962). Ridge regression, SVM, and neural networks often show good performance on various text classification problems (Zhang and Oles, 2001; Wang et al., 2006), whereas naive Bayes and random forest are known to perform well on smaller datasets (Pranckevičius and Marcinkevičius, 2017). For SVM, we use a polynomial kernel as this showed the best performance in preliminary tests. For random forest and neural networks, we employ a grid search to determine the optimal number of estimators and network topology.

We use each model in a stacked architecture composed of a language model followed by the classifier. We use the language model to learn sensible fixed-length vector representations of the transcriptions and metadata and the classifier to fine-tune these representations. Deep language models such as LSTMs (Graves and Schmidhuber, 2005) and transformers (Vaswani et al., 2017) were also considered, but preliminary tests showed poor results (see Section 6).

Doc2Vec (Le and Mikolov, 2014) is chosen as language model in our architecture. Doc2Vec learns document-level representations that capture the semantics of documents (i.e., transcriptions and metadata) as a whole. This is well-suited for framing type classification, as framing typically emerges from the entirety of a document rather than individual sentences or words.

4.2.2. Data Preparation

All transcriptions and textual metadata attributes are preprocessed before vectorization using NLP techniques, including stemming, lemmatization, conversion to lowercase, removal of special, single-letter characters, stop words, and unnecessary white space. We first use the preprocessed data to train the Doc2Vec model. We choose to train this model from scratch because we have over 10k unlabeled samples, which help us embody the semantics of the domain more accurately. The unlabeled samples are only used at this stage and are discarded after training the model. Once completed, we use it to generate vectors for the 120 labelled samples in our dataset, which serve as input for the different classifiers.

As target labels we use the dominant frame types: *episodic* and *thematic*. Target labels are derived from both expert and crowd annotations. We test three label configurations to investigate their individual and combined strength: *(1)* the model is trained purely on expert labels, *(2)* the model is trained purely on crowd labels, and *(3)* we combine the expert and crowd labels into a single set (resolving any disagreement by favouring the expert label) which is then used for training the model. In all three configurations, the expert labels serve as gold standard.

4.2.3. Experiment Design

We use stratified K-fold cross-validation (Dietterich, 1998) and split our dataset in a training and test set for each fold using an 80/20 split to overcome the modest size of our dataset. Since this affects our ability to generalize our results to unseen data, we use regularization methods to minimize overfitting.

We compare the results amongst all models and label configurations and against a majority-class classifier as baseline. For significance testing, we rely on McNemar's asymptotic marginal homogeneity test (McNemar, 1947), which determines whether two binary-classification models have the same distribution of predictions.

5. RESULTS

In this section, we first analyze the inter-rater agreement among experts and then evaluate the crowd performance. We systematically compare experts and crowds to understand how *thematic* and *episodic* framing are depicted and perceived in online news videos and how challenging the annotation process is. Finally, we report on the framing classification performance.

5.1. Inter-rater reliability among experts

In Table 3, we report the percentage agreement and the IRR score among experts using Krippendorff's α (for nominal and ordinal input), on each user study question. Overall, the IRR scores vary considerably across dependent variables. For nominal input, we observe a substantial agreement on the frame selection modality, α=0.66, but only fair agreement on the video sentiment, α=0.33. On the annotation of the dominant frame, we observed a moderate agreement of α=0.42. For the user study questions evaluated on a scale, Krippendorff's α values are significantly lower when considering the input as nominal values. When using the ordinal scale as input, we see, however, moderate (framing score, sentiment magnitude for all videos) to substantial agreement (video trustworthiness and sentiment magnitude for videos in agreement).

We further analyzed the agreement on the dominant frame between every two expert annotators, and we observed high IRR variability, from fair to substantial. While the two experts with a political science background tend to agree substantially, α=0.62, the expert with a social sciences background tends to only fairly agree with them, α=0.29 and α=0.34. When using the answers of the expert with a social sciences background only when a tie-breaker is needed, we observe a substantial agreement with Krippendorff's α=0.65. In Section 6, we further reflect on the difficulty of the annotation process and the causes of disagreement among experts.

5.2. Crowd performance

We gathered 758 judgments from a total of 303 crowd workers on the 58 videos in the *Expert* dataset (see Section 3.2). Each crowd worker annotated around 2.5 videos (minimum 1, maximum 58), besides the gold videos. We post-filtered the judgments based on the two control questions mentioned in Section 4.1 used as attention checks. In the data collected from FigureEight, we

Measure	Dominant Frame	Framing Score	Frame Selection Modality	Sentiment Present	Sentiment Value		Sentiment Magnitude		Trust
					All Videos	Agreed Videos	All Videos	Agreed Videos	
% Agreement	56.90	10.34	79.31	50.0	44.83	86.2	29.31	51.72	17.24
Krippendorff's α (nominal)	0.42	0.19	0.66	0.33	0.34	0.84	0.26	0.51	0.24
Krippendorff's α (ordinal)	-	0.51	-	-	-	-	0.46	0.76	0.72

***Table 3.* Overview of expert annotators agreement on the dependent variables.**

observed that multiple workers' ids submitted judgments from the same IP address, close in time.[7] Thus, we used a third post-filter based on multiple worker ids from the same IP address.

We aggregate the crowd's answers using the majority vote approach, which showed the best results compared to MACE (Hovy et al., 2013) and CrowdTruth (Dumitrache et al., 2018). While the latter two consider the quality of the input and annotators, we hypothesize that the low number of videos annotated per annotator may not be enough to compute reliable quality scores for them. We compute precision (P), recall (R), and F1-score (F1) per framing type and the micro- and macro-performance scores (typically used in multi-class classification) across framing types to see how the crowd performs compared to experts. We analyze the performance of the crowd in terms of identifying the dominant frame (DF-C & DF-E) and the framing score (FS-C & FS-E) of the videos, which is transformed into a dominant frame (values 1-3 become thematic framing, value 4 becomes balanced framing and values 5-7 become episodic framing, after computing the median score among expert and crowd annotators for the framing score).

Table 4 reports these results, before and after filtering out the low-quality annotations. Filtering out judgments based on the workers' IP further improves the overall results, and thus, we only report these. After filtering out these low-quality annotations, we always observed better performance of the crowd. The best F1-score of the crowd (both micro and macro) is always higher than the experts' IRR score, indicating that crowd annotators can perform just as well as experts on this difficult task.

Episodic framing seems to generate more disagreement than thematic framing. Upon empirical and comparative analysis of our annotations (see Section 5.3), we observe that episodic videos are often given balanced framing scores. Thus, in the crowd annotations, we merged the dominant frame *balanced* with the frame *episodic* and computed the crowd performance against the experts (last row in Table 4). This further improved the performance of the crowd: 0.62 F1-score for episodic frame, 0.69 for thematic frame and a combined micro-F1 score of 0.66, which indicates that episodic framing is often found in videos alongside thematic framing.

For the remaining of the analysis, we use the crowd annotations filtered based on the video title, video type, and worker IP, and by merging the *balanced* and *episodic* frames. Thus, we have 605 annotations from 251 annotators.

[7] We understand that many annotators could solve the task from an open space, like a cafe, but we would like to avoid the cases where the same annotator logs in with different ids.

Target	Annotations	Label	P	R	F1	Micro Measures			Macro Measures		
						P	R	F1	P	R	F1
DF-C & DF-E	Unfiltered	Episodic	0.50	0.33	0.40	0.53	0.53	0.53	0.35	0.35	0.34
		Thematic	0.56	0.71	0.63						
	Filtered (type, title, ip)	Episodic	0.64	0.41	0.5	0.62	0.62	0.62	0.42	0.40	0.40
		Thematic	0.63	0.81	0.70						
FS-C & FS-E	Unfiltered	Episodic	0.42	0.69	0.52	0.53	0.53	0.53	0.35	0.47	0.40
		Balanced	0.0	0.0	0.0						
		Thematic	0.63	0.71	0.67						
	Filtered (type, title, ip)	Episodic	0.52	0.81	0.75	0.59	0.59	0.59	0.39	0.52	0.44
		Balanced	0.0	0.0	0.0						
		Thematic	0.65	0.75	0.7						
DF-C & DF-E (Balanced->Episodic)	Unfiltered	Episodic	0.57	0.59	0.58	0.60	0.60	0.60	0.60	0.60	0.60
		Thematic	0.63	0.61	0.62						
	Filtered (type, title, ip)	Episodic	0.64	0.59	0.62	0.66	0.66	0.66	0.65	0.65	0.65
		Thematic	0.67	0.71	0.69						

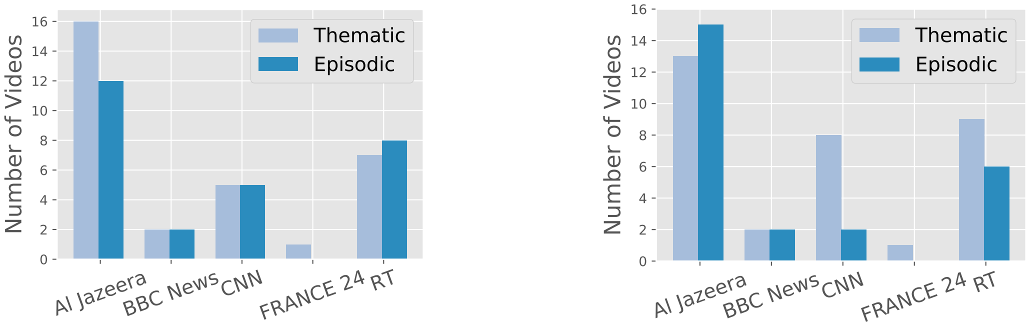
Table 4. Performance of the crowd annotators (C) when compared to the expert annotators (E) for the annotation of dominant frame (DF) and framing score (FS).

5.3. Framing analysis by experts and crowd

Following, we perform a systematic comparison between expert and crowd annotations by focusing on framing analysis, video channel trustworthiness, frame selection modality, and video channel and framing type sentiment.

Disagreement in framing annotation. In Figure 3a we plotted the dominant frame distribution per channel, computed using majority vote on expert annotations. The dataset is fairly balanced: 31 *thematic* and 27 *episodic* videos. For each news channel, the coverage of *episodic* and *thematic* framing is also fairly balanced. Interestingly, on *thematic* framing, all three experts agree 22/31 times (71%), while on *episodic* framing, all experts agree only 11/27 times (41%). In Figure 3b, we show the dominant frame distribution per video channel for the crowd annotators. Compared to experts, we see that all channels except for Al Jazeera contain more thematic videos than episodic videos. We also compared the aggregated crowd framing types with each expert annotator, and we observed that they are more often in agreement with the political sciences experts (60% and 65% of the videos) than with the social sciences expert (55% of the videos).

Further, we assigned a second framing type to each video in *expert data*, based on the median value of the *framing score*: *thematic* for scores in [1:3], *balanced* for a score of 4 and *episodic* for scores in [5:7]. This resulted in the following distribution of videos: 28 *thematic*, 16 *episodic* and 14 *balanced*. Only 3 videos changed their framing type from *thematic* to *balanced*, but 11 videos changed their framing type from *episodic* to *balanced*. This reinforces the idea that *episodic* framing is more difficult to identify compared to *thematic* framing. Furthermore, in 6 (2 for *thematic* and 4 for *episodic*) out of these 14 changes, the expert annotators were in full agreement on the dominant frame. This means that these videos are likely to represent both framing types, and the decision on which framing type is dominant is highly subjective.



(a) Expert annotations.

(b) Crowd annotations.

***Figure 3.** Framing type distribution per channel and across crowd annotations. The bars represent the number of videos annotated as thematic and episodic.*

We notice that the framing scores of experts for *episodic* videos very rarely reach the maximum value of 7, most of them being very close or equal to neutral (4). Thus, based on the framing scores given by the expert annotators and the increased disagreement on *episodic* framing compared to *thematic* framing, we conclude that *thematic* framing is more clearly exhibited in videos compared to *episodic* framing. When analyzing the distribution of video framing scores for crowd annotators, we observe a similar behaviour.

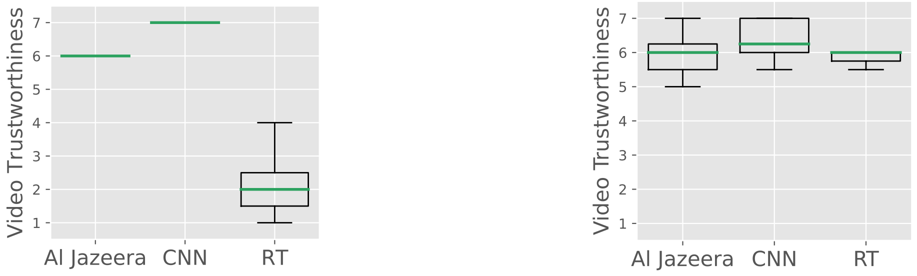
The majority of the crowd annotators evaluated only a few videos, namely less than three. We observe, however, that there could be a learning curve. The percentage of videos annotated in agreement with expert annotators is usually higher for crowd annotators that annotated, for example, more or equal to six videos than less or equal to two videos, 0.60 versus 0.49 on average.

The channels BBC News and FRANCE 24 have very few videos. Thus, for the remainder of the analysis, we exclude these two channels.

Video channel trustworthiness. In Figure 4a, we plotted the aggregated (median) trustworthiness scores per channel given by experts. Overall, the channel RT seems to contain the most untrustworthy videos, compared to all the other channels, difference that is statistically significant c.f. Kruskal-Wallis test (for all raters and their aggregated score, $H(2) = 44.550$, $p < 0.05$). Further, we performed a Dunn post-hoc analysis with Holm-Bonferroni correction to determine which is the channel with a different score distribution. Pairwise comparisons show that the pairs (RT, Al Jazeera) and (RT, CNN News) have a statistically significant difference between the trustworthiness scores, with $p << 0.01$. Thus, we conclude that trustworthiness scores significantly differ for the RT channel compared to the other two channels.

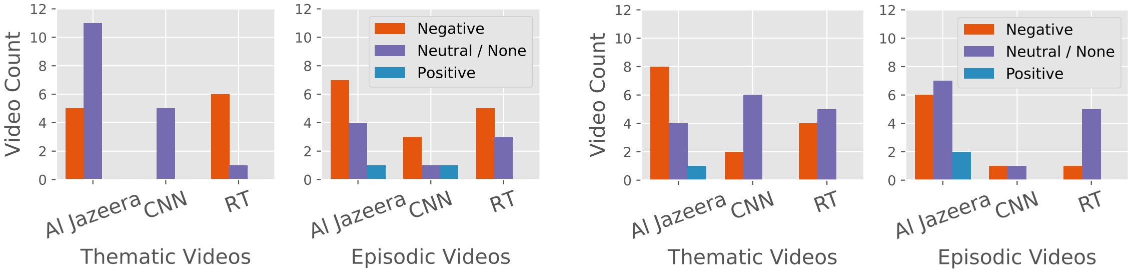
Figure 4b depicts the aggregated trustworthiness score per channel given by crowd annotators. Interestingly, we observe that almost all videos have trustworthiness scores above the median (4), with all channels being rated high. Thus, this means that crowd annotators do not perceive videos or video channels with different levels of trustworthiness, as opposed to experts.

Correlation between framing type and frame selection modality. Among the three options—words,



(a) Expert annotations. ***(b) Crowd annotations.***

***Figure 4.* Distribution of video channel trustworthiness**



(a) Expert annotations. ***(b) Crowd annotations***

***Figure 5.* Distribution of video sentiment score per type of framing and across video channels.**

images, and images and words—experts only picked words (14/53 times) and images and words (39/53 times). Images alone are not informative enough to decide on the framing type. Although images and words are more often picked for both framing types, they seem more informative to identify *episodic* framing, 22/39 cases, while words seem more informative to identify *thematic* framing, 11/14 cases. Using a Chi-square test with Yates correction, we indeed found a correlation between the type of framing and the framing selection modality ($\chi^2(2) = 3.75$, $p < 0.05$), but we can not conclude that the correlation is strong. For the crowd annotations, however, we found no correlation between the framing type and the frame selection modality ($\chi^2(2) = 0.924$, $p > 0.05$).

Sentiment distribution across frames and channels. In Figure 5a we plotted the distribution of sentiment per framing type and per channel for experts. The majority of the *thematic* videos show no sentiment, while fewer are negative. *Episodic* videos tend to cover the entire range of sentiments, with more than half of the videos being negative, followed by neutral and a few positive. For *thematic* videos, we see that only Al Jazeera and CNN channels contain more neutral videos than negative. The opposite trend is observed for the channel RT. Conversely, when looking at *episodic* videos, all channels seem to have more negative videos than neutral.

Figure 5b depicts the crowd sentiment distribution for each framing type and channel. Similarly to experts, most videos are depicted as negative or neutral. The Al Jazeera channel is the only channel

Labels	RR	NB	RF	SVM	NN	MC
Experts	0.615	0.692	0.763	0.615	0.525	0.534
Crowd	0.542	0.458	0.612	0.667	0.547	0.549
Combined	0.478	0.565	0.539	0.739	0.523	0.531

***Table 5.* Mean accuracy over 100 runs for expert, crowd, and combined label configurations. Reported are the performance with ridge regression (RR), naive Bayes (NB), random forest (RF), support vector machines (SVM), and neural networks (NN). MC reports the accuracy of a majority-class classifier.**

containing positive videos (both episodic and thematic). Compared to experts, we see, however, some differences: crowd annotators consider most RT videos as neutral; for the Al Jazeera channel, the distribution of negative and neutral videos per framing type is inverted: thematic videos are more often perceived as negative, while episodic videos are more often perceived as neutral.

5.4. Machine Learning

Table 5 reports the mean test accuracy over $K = 100$ runs for all classification models for all three label configurations. Table 6 lists the statistical significance between the results of the best performing model and that of all others for all three label configurations, whereas Table 7 reports the statistical significance between the crowd and combined label configurations for all models. We omit the results on the expert labels in this latter table since they only cover a subset of the crowd and combined labels. Furthermore, we aim to investigate whether combining expert and crowd labels performs better than using crowd labels alone.

We obtained these results by training a Doc2Vec model for 300 epoch on a corpus of 10,493 labeled and unlabeled documents, learning vector representations of size 25. The 120 vectors for which we have annotations were extracted from this set and used as input for all experiments. For random forest, we varied the number of estimators from 100 to 2,000 with a 100 step increment and found 1,000 estimators to give the best overall results. Likewise for neural networks, which performed best with two hidden layers of 12 and 6 nodes, respectively, and ReLU activation functions. This network was trained using Adam with an initial learning rate of 0.01 and with the cross-entropy loss as criterion until no further significant reduction in loss was encountered. For regularization, the network used a dropout rate of 0.05, an L2-norm of 0.01, and a batch normalization layer before each ReLU call. For the remaining three models, we used the same hyperparameters as those of their original literature (see Section 4.2.1). We repeated all experiments 100 times. We used the average accuracy as the final result, with a confidence level of 0.05.

Overall, our results show that there is no single best classifier that performs well irrespective of which label configuration is being used: with expert labels, the random forest classifier takes the first spot with an accuracy of 0.763, whereas for the crowd and combined label configurations it is the polynomial SVM, with an accuracy of 0.667 and 0.739, respectively. This might happen because the expert label set has fewer samples, making it difficult to train an SVM. Nevertheless, each of these outperforms the majority-class baseline, which, for all three label configurations, divides the classes in a roughly even split (expert: 0.534, crowd: 0.549, combined: 0.531).

Labels	Model	RR	NB	RF	SVM	NN
Experts	RF	0.732	0.739	-	0.705	0.856
Crowd	SVM	0.174	0.093	0.578	-	0.414
Combined	SVM	**0.029**	**0.038**	**0.001**	-	0.071

Table 6. ***Statistical significance between the best performing model and all others for expert, crowd, and combined label configurations. Reported are the p-values for ridge regression (RR), naive Bayes (NB), random forest (RF), support vector machines (SVM), and neural networks (NN).***

Model	p-value
RR	0.272
NB	0.128
RF	0.605
SVM	**0.025**
NN	0.793

Table 7. ***Statistical significance between the predictions obtained using the crowd labels and the combined expert and crowd labels. Reported are the p-values for ridge regression (RR), naive Bayes (NB), random forest (RF), support vector machines (SVM), and neural networks (NN).***

The random forest classifier achieves the highest accuracy overall, 0.763, but does not perform well on either the crowd labels (0.612) or the combined label configuration (0.539). Possibly, this effect might be the difference in disagreement between the label sets: during training, random forest creates decision nodes to assign seen samples to classes, which is more difficult as the variance in underlying distribution increases, requiring more splits. This is an even greater problem for the combined label configuration, which combines the expert and crowd distributions. We see a similar effect for ridge regression. No obvious effect can be observed for the Naive Bayes classifier, which varies greatly per label set, obtaining an accuracy lower than the baseline (0.458), and the neural network, which gives an overall poor performance, similarly to the baseline.

Within the different label sets, only the performance of the SVM in the combined label configuration is statistically significant when compared to all other classifiers (Tab. 6), except for the neural network. This suggests that, save for neural networks, SVMs are indeed the better classifiers on our dataset for this label configuration. In contrast, the differences in performance between the top classifier and all other models, for the expert and crowd label configurations, were not found significant. However, this is expected since the relatively low number of samples in the expert label set lacks the statistical power needed for precise evaluation.

Statistical tests on the performances of the same classifier between the crowd and combined label sets (Tab. 7) indicate that only the difference in accuracy with the SVMs is significant ($p = 0.025$). Thus, the underlying distribution of the predictions is significantly different between the two label sets.

6. DISCUSSION

In light of our research question *"To what extent can crowdsourcing and machine learning effectively be employed to identify thematic and episodic framing in video news?"*, we first discuss the results of the framing annotation studies and classification experiments, and then we provide the implications and limitations of our approach.

6.1. Framing annotation by experts and crowd

Framing annotation for political crisis generates more disagreement than in other domains. Experts tend to have only moderate to substantial agreement (see Section 5.1) when annotating frames. Such values are on-par with existing research on annotating news articles in the political domain (Burscher et al., 2014), which shows Krippendorff's α values between 0.21 and 0.58. In the medical domain (Kang et al., 2017), however, the annotation of thematic and episodic framing shows an almost perfect agreement. We hypothesize that the medical topic in (Kang et al., 2017) is less challenging because many videos analyzed depict personal stories of people dealing with ADHD, making it easier to identify episodic frames. Our and Burscher et al. (2014) analyses could indicate that political news generates more disagreement even among experts. Existing research on the framing of international conflicts, in particular the Ukraine crisis, supports this: there are substantial differences in the interpretation of the crisis both between parties involved (Makhortykh and Sydorova, 2017) and the academic communities (Makhortykh and Bastian, 2020).

To the best of our knowledge, no crowdsourcing study looked at episodic and thematic framing, so a systematic comparison can not be performed. In the area of evaluating information veracity, such as credibility assessment (Bhuiyan et al., 2020) or bias (Lim et al., 2020), it is, however, common for crowd annotators to exhibit high disagreement. Finally, we hypothesize that the lack of background knowledge of crowd annotators on the Crimea Crisis could influence their perception regarding the dominant frame of the videos—they internalized the effect of the problems discussed either as having a large or restricted impact.

Episodic framing generates more disagreement among both experts and crowds than thematic framing. Our analysis showed that thematic videos are assigned framing scores with an average of 2, while episodic videos are assigned framing scores very close to the neutral value, namely 4.8. This could be a problem inherent to our case study and online news videos. The Crimea Crisis case study could potentially touch on various issues or events, but their impact on individuals or society is not easily understood. Moreover, many videos in our dataset contain interviews and analysis pieces, which might be more challenging to quantify in terms of framing.

So far, we found that social and political science scholars disagree with each other when annotating episodic and thematic framing, and the crowd agrees more with political science scholars. These observations are on-par with existing credibility assessment research (Bhuiyan et al., 2020). Although all three experts specialize in area studies in the region of Eastern Europe, the disagreement among them is substantial. Thus, there could be disciplinary differences in treating the concept of framing and using it for analytical purposes. Furthermore, the social science expert comes from Ukraine, whereas the two political experts do not. In the case of political crises, having an expert from the country where the crisis takes place can have an ambiguous effect. It can bring additional domain knowledge but also can make the judgments more biased or subjected to self-censorship.

Expert and crowd annotators disagree on channel and video characteristics, but they agree more on the videos' sentiment. Experts differentiate the video channels based on their trustworthiness and tend to characterize the reporting style of the video, e.g., "objective", "biased", "propaganda". The crowd, however, tends to evaluate all channels and videos as highly trustworthy, which is a fundamental problem of news videos made for propaganda purposes. We hypothesize that the crowd might not know about the Crimea Crisis and have little to no understanding of all actors involved and how the topic is depicted across news outlets.

Expert and crowd annotators are less likely to use visual news video aspects to decide on the framing type. This, however, could be a property of the dataset, as in news broadcasts there are usually one or multiple anchors in a studio without showing footage.

Crowd annotators found the task and instructions clear and easy to follow, in a proportion of 36% (cf question 12 in Table 2). However, a larger proportion (42%), considered the additional references and examples provided in the instructions necessary, to better understand the task.

6.2. Machine Learning

Our classification experiment suggests that the problem of identifying framing type is not only challenging for humans but also for machine learning. This, however, is expected, as the uncertainty of human annotators is embedded in the labels that we use to train our models with. These annotations also contain a considerable disagreement, which translates to a lower signal-to-noise ratio and a higher variance in the underlying distribution, making it more challenging to fit a model to. Despite these challenges, however, the polynomial SVM managed to achieve a statistically significant performance compared to the others. This might be the result of the polynomial kernel, which has a certain robustness to noise and outliers (Hoak, 2010).

Many tested classifiers might underperform due to the relatively low number of labelled samples in our dataset. This, in combination with the already complex learning problem of classifying natural language, may have provided an insufficiently strong signal to learn over. This effect may have been more profound for models with a relatively large number of learnable parameters, such as neural networks (Foody et al., 1995), which performed poorly irrespective of label set. However, the experiments on the expert label set are likely affected the most because of their size. Another possible reason might be found in the lack of agreement between the annotations provided by the experts and crowd, which leads to a different distribution of lower-quality crowd labels compared with the higher-quality expert labels. Thus, the model is required to learn a more complex function.

Preliminary experiments indicated that for our learning problem, shallow models outperform deep language models such as temporal CNNs, LSTMs, and transformers, despite this latter group of models being very popular for language classification. We hypothesize this is due to the relatively low number of samples in our dataset, which is insufficient to effectively train the large number of parameters of a deep model, especially given the difficulty of the task and the relatively low signal-to-noise ratio.

The reasons for which we abstain from using the videos as input for our machine learning pipeline are both empirical and practical: *(1)* our analysis showed that experts annotators never rely on images alone to decide on the framing type, *(2)* images and words are similarly used for both thematic and episodic framing, with little discrimination among the two, and finally, *(3)* the low number of sample videos in our experiments. Concretely, including the videos themselves would have required us to incorporate a video summarization component into our model that learns which frames are most relevant to convey framing type (Fajtl et al., 2018), e.g., by using CNNs. However, since our classifiers have no prior knowledge of how to identify framing type, all error correction will have to come from the labels. In our case, the number of samples is insufficient for this purpose, given the relatively large number of parameters of a CNN or similar video summarization model.

6.3. Implications

The results of our annotation studies for identifying episodic and thematic framing in online news videos have several implications for improving the annotation process, but also for raising awareness about the societal impact of these types of framing.

Video framing annotation should be performed at more granular levels. Existing literature on framing analysis and annotation focuses primarily on textual sources (Burscher et al., 2014),(Cremisini et al., 2019), and also on smaller granularity such as news headlines (Guo et al., 2021) or sentences (Opperhuizen and Schouten, 2020). Thus, given the high probability that news videos would contain both mentions of episodic and thematic framing, we argue that the identification of framing type in videos should be performed at more granular levels, such as video fragments. Furthermore, the annotation should focus more on understanding which particular aspects of the video are framed as thematic or episodic, instead of attaching a dominant frame to the entire video.

We need solutions for raising awareness on the content that people consume online. Previous research (Gross, 2008) showed that episodic frames are more prone to evoke emotional responses and thus influence more the opinion of the public. Conversely, thematic frames are more prone to decrease the emotional involvement of the public. In addition, news outlets make use of such framing effects to communicate their messages. Thus, the manner in which such messages are communicated further impacts the news consumers. Research is extensively focusing on helping or making people reflect on the credibility and veracity of the news they read (Kirchner and Reuter, 2020; Bhuiyan et al., 2018). However, such solutions have not focused yet on more subtle ways of framing political and societal issues, such as episodic and thematic framing. The fidelity with which crowd annotators rated all videos highly trustworthy is another argument supporting such tools, since framing aspects can have an even higher societal and individual impact.

As seen in recent literature Burscher et al. (2014), frame detection is difficult. We also showed that thematic and episodic framing annotation is difficult. Nevertheless, so is the annotation and detection of emotions or bias. This does not preclude the different kinds of frames influencing the viewer and their perception of how trustworthy the news is. Thus, working on detection and annotation is important to provide viewers with more tools to decipher what media aim to do. In our study, we provide preliminary means to study and capture news media bias and, more precisely, episodic and thematic framing. Accurately capturing the framing and its perception by the viewers constitutes the first step that could ultimately allow citizens, news consumers, to detect misrepresentations in the news videos they watch. Furthermore, our results and analyses can be used along methods that aim to measure and improve, for instance, the degree of (viewpoint) diversity in news recommendations (Mulder et al., 2021; Heitz et al., 2022) or search results (Draws et al., 2022).

6.4. Limitations

We identified several limitations of our study: *(1)* number of samples, *(2)* expert annotation procedure, *(3)* crowd annotators profiling, *(4)* study interface, and *(5)* case study generalizability.

Number of samples. We have limited the number of samples to 120 videos, 58 of which were annotated by both crowd workers and experts. These numbers are large enough to reliably perform statistical comparisons, but barely sufficient to train a classifier without endangering the external validity. The low number of samples prevented us from using deep language models (Barbedo, 2018), which, in recent years, have shown better performance on multiple language tasks compared to shallow models (Sundermeyer et al., 2012; Irie et al., 2019). Similarly for the absence of video features in our classifier, which are commonly based on CNNs or other deep architectures and therefore require a much higher number of samples to be effective.

Expert annotation procedure. The experts annotated all videos independently. Due to time constraints, we could not organize sessions for experts to discuss the videos on which they disagree. As shown in literature (Moretti et al., 2011), this can improve the IRR. Disagreement among our experts, however, was an important indicator of the task's difficulty.

Crowd annotators profiling. We neither accounted for annotators' potential biases nor measured or collected the impact of annotators' location, gender, knowledge, or stance regarding the Crimea Crisis. However, we acknowledge that crowd annotators from affected areas or with various degrees of prior knowledge could perceive episodic and thematic frames differently. Similarly, we did not test for the ability of the crowd annotators to differentiate between episodic and thematic framing.

User study interface. The crowd annotators watched the videos on an external page, but in comments, they suggested integrating the video into the actual task.

Case study generalizability. The Crimea Crisis is a rather distinct case study, characterized by intense use of opposing frames by major geopolitical powers (i.e., the West and Russia). Thus, the methodological approach proposed in the current research might seem to be harder to generalize. However, the geopolitical aspect of episodic and thematic framing is also common for many other news topics, such as, COVID-19 pandemic or US 2020 presidential elections. Furthermore, episodic and thematic framing are general frames that appear across issues, time, and space.

7. CONCLUSION AND FUTURE WORK

In this paper, we proposed an annotation study to identify episodic and thematic framing in news videos about the Crimea Crisis. We employed social and political science expert annotators to help us understand how episodic and thematic framing are depicted in videos and which video characteristics (i.e., sentiment, trustworthiness) can help identify the dominant framing type. We then use expert insights to conduct an annotation task on the same topic with crowd annotators. We found that, for both expert and crowd annotators, identifying thematic and episodic framing is a difficult task, highly prone to disagreement. In general, the crowd performs similarly to the expert annotators when identifying frames. However, crowd annotators can not capture particular aspects of the video that could help characterize the type of frame.

To investigate whether machine learning can scale up the framing annotation process, we trained various classifiers to distinguish between episodic and thematic framing based on audio transcriptions and video metadata. We used expert and crowd annotations as target labels in several configurations. Our results showed that, by combining expert and crowd annotations, and using these in combination with a polynomial SVM, we significantly improved the classification performance over using just the crowd labels. While the accuracy of this approach still fell short of a perfect score, this outcome nevertheless shows that supervised classification can scale up the number of annotations.

In future work, we plan to further experiment with crowdsourcing tasks for annotating frames in videos, i.e., to improve the user interface, simplify the annotation task, and better guide the annotators and help them understand the difference between episodic and thematic framing. A larger labeled dataset would also allow us to experiment with deep language models such as LSTMs and transformers. Including video features as input to the model might also help improve classification performance, e.g., by using CNNs to summarize raw video data. These predictions could also help (online) media platforms to automatically tag (video) content to create awareness of framing strategies, recommend same topics items with different interpretations, or help identify stories promoting highly biased interpretations. Finally, we plan to study the effect of annotators' location and prior knowledge concerning the Crimea Crisis in annotating and perceiving frames in videos.

ACKNOWLEDGMENTS

This research was supported by the Capture Bias[8] project, part of the VWData Research Programme funded by the Startimpuls programme of the Dutch National Research Agenda, route "Value Creation through Responsible Access to and Use of Big Data" (NWO 400.17.605/4174).

8. REFERENCES

Aarøe, L. (2011). Investigating frame strength: The case of episodic and thematic frames. *Political communication* 28, 2 (2011), 207–226.

Aydin, F. T and Sahin, F. K. (2019). The politics of recognition of Crimean Tatar collective rights in the post-Soviet period: With special attention to the Russian annexation of Crimea. *Communist and Post-Communist Studies* 52, 1 (2019), 39–50.

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and electronics in agriculture* 153 (2018), 46–53.

Baumer, E, Elovic, E, Qin, Y, Polletta, F, and Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1472–1482.

Bhuiyan, M. M, Zhang, A. X, Sehat, C. M, and Mitra, T. (2020). Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

Bhuiyan, M. M, Zhang, K, Vick, K, Horning, M. A, and Mitra, T. (2018). FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2018, Jersey City, NJ, USA, November 03-07, 2018*, Vanessa Evers, Mor Naaman, Geraldine Fitzpatrick, Karrie Karahalios, Airi Lampinen, and Andrés Monroy-Hernández (Eds.). ACM, New York, NY, USA, 205–208. DOI: http://dx.doi.org/10.1145/3272973.3274056

Biersack, J and O'lear, S. (2014). The geopolitics of Russia's annexation of Crimea: narratives, identity, silences, and energy. *Eurasian geography and economics* 55, 3 (2014), 247–269.

Bleiker, R. (2018). *Visual global politics*. Routledge.

Bock, M. A. (2016). Showing versus telling: Comparing online video from newspaper and television websites. *Journalism* 17, 4 (2016), 493–510.

Boräng, F, Eising, R, Klüver, H, Mahoney, C, Naurin, D, Rasch, D, and Rozbicka, P. (2014). Identifying frames: A comparison of research methods. *Interest Groups & Advocacy* 3, 2 (2014), 188–201.

Boyd-Barrett, O. (2017). Ukraine, mainstream media and conflict propaganda. *Journalism studies* 18, 8 (2017), 1016–1034.

Bryant, J and Miron, D. (2004). Theory and research in mass communication. *Journal of communication* (2004).

Burscher, B, Odijk, D, Vliegenthart, R, De Rijke, M, and De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures* 8, 3 (2014), 190–206.

Card, D, Boydstun, A, Gross, J. H, Resnik, P, and Smith, N. A. (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)*. 438–444.

Castillo, C, Mendoza, M, and Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Res.* 23, 5 (2013), 560–588. DOI: http://dx.doi.org/10.1108/IntR-05-2012-0095

Cremisini, A, Aguilar, D, and Finlayson, M. A. (2019). A Challenging Dataset for Bias Detection: The Case of the Crisis in the Ukraine. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 173–183.

[8] https://capturebias.wordpress.com/home/

Dai, P, Rzeszotarski, J. M, Paritosh, P, and Chi, E. H. (2015). And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 628–638.

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

Devlin, J, Chang, M.-W, Lee, K, and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10, 7 (1998), 1895–1923.

Dimitrova, A, Frear, M, Mazepus, H, Toshkov, D, Boroda, M, Chulitskaya, T, Grytsenko, O, Munteanu, I, Parvan, T, and Ramasheuskaya, I. (2017). The Elements of Russia's Soft Power: Channels, Tools, and Actors Promoting Russian Influence in the Eastern Partnership Countries. (2017).

Dimitrova, D. V. (2006). Episodic frames dominate early coverage of Iraq War in the NYTimes. com. *Newspaper Research Journal* 27, 4 (2006), 79–83.

Draws, T, Inel, O, Tintarev, N, Baden, C, and Timmermans, B. (2022). Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 135–145.

Drucker, H, Burges, C. J, Kaufman, L, Smola, A, Vapnik, V, and others, . (1997). Support vector regression machines. *Advances in neural information processing systems* 9 (1997), 155–161.

Dumitrache, A, Inel, O, Aroyo, L, Timmermans, B, and Welty, C. (2018). CrowdTruth 2.0: quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080* (2018).

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

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

Fajtl, J, Sokeh, H. S, Argyriou, V, Monekosso, D, and Remagnino, P. (2018). Summarizing videos with attention. In *Asian Conference on Computer Vision*. Springer, 39–54.

Fengler, S, Kreutler, M, Alku, M, Barlovac, B, Bastian, M, Bodrunova, S. S, Brinkmann, J, Dingerkus, F, Hájek, R, Knopper, S, and others, . (2020). The Ukraine conflict and the European media: A comparative study of newspapers in 13 European countries. *Journalism* 21, 3 (2020), 399–422.

Field, A, Kliger, D, Wintner, S, Pan, J, Jurafsky, D, and Tsvetkov, Y. (2018). Framing and Agenda-Setting in Russian News: a Computational Analysis of Intricate Political Strategies.. In *EMNLP*. Association for Computational Linguistics, 3570–3580.

Flintham, M, Karner, C, Bachour, K, Creswick, H, Gupta, N, and Moran, S. (2018). Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

Foody, G, McCulloch, M, and Yates, W. (1995). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing* 16, 9 (1995), 1707–1723.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Grant, T. D. (2015). Annexation of crimea. *American journal of international law* 109, 1 (2015), 68–95.

Graves, A and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

Gross, K. (2008). Framing persuasive appeals: Episodic and thematic framing, emotional response, and policy opinion. *Political Psychology* 29, 2 (2008), 169–192.

Guo, L, Mays, K, Zhang, Y, Wijaya, D, and Betke, M. (2021). What makes gun violence a (less) prominent issue? A computational analysis of compelling arguments and selective agenda setting. *Mass Communication and Society* (2021).

Hart, P. S. (2011). One or many? The influence of episodic and thematic climate change frames on policy preferences and individual behavior change. *Science Communication* 33, 1 (2011), 28–51.

Heitz, L, Lischka, J. A, Birrer, A, Paudel, B, Tolmeijer, S, Laugwitz, L, and Bernstein, A. (2022). Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism* (2022), 1–21.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

Hoak, J. (2010). The Effects of Outliers on Support Vector Machines. *Portland State University* (2010).

Hovy, D, Berg-Kirkpatrick, T, Vaswani, A, and Hovy, E. (2013). Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1120–1130.

Irie, K, Zeyer, A, Schlüter, R, and Ney, H. (2019). Language modeling with deep transformers. *arXiv preprint arXiv:1905.04226* (2019).

Iyengar, S. (1994). *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Iyengar, S. (1996). Framing responsibility for political issues. *The Annals of the American Academy of Political and Social Science* 546, 1 (1996), 59–70.

Iyyer, M, Enns, P, Boyd-Graber, J, and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1113–1122.

John, G. H and Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964* (2013).

Kang, S, Ha, J.-S, and Velasco, T. (2017). Attention deficit hyperactivity disorder on YouTube: Framing, anchoring, and objectification in social media. *Community mental health journal* 53, 4 (2017), 445–451.

Khaldarova, I and Pantti, M. (2016). Fake news: The narrative battle over the Ukrainian conflict. *Journalism practice* 10, 7 (2016), 891–901.

Kirchner, J and Reuter, C. (2020). Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

Kofman, M, Migacheva, K, Nichiporuk, B, Radin, A, Oberholtzer, J, and others, . (2017). *Lessons from Russia's Operations in Crimea and Eastern Ukraine*. Rand Corporation.

Korostelina, K. (2015). Crimean Tatars from mass deportation to hardships in occupied Crimea. *Genocide Studies and Prevention: An International Journal* 9, 1 (2015), 6.

Le, Q and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

Lim, S, Jatowt, A, Färber, M, and Yoshikawa, M. (2020). Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 1478–1484.

Lopatecki, J, Rose, A, Hughes, J, and Wilson, B. (2019). Passively monitoring online video viewing and viewer behavior. (April 23 2019). US Patent App. 10/270,870.

Makhortykh, M. (2018). # NoKievNazi: Social Media, Historical Memory and Securitization in the Ukraine Crisis. In *Memory and Securitization in Contemporary Europe*. Springer, 219–247.

Makhortykh, M and Bastian, M. (2020). Personalizing the war: Perspectives for the adoption of news recommendation algorithms in the media coverage of the conflict in Eastern Ukraine. *Media, war & conflict* (2020), 1750635220906254.

Makhortykh, M and González Aguilar, J. M. (2020). Memory, politics and emotions: internet memes and protests in Venezuela and Ukraine. *Continuum* 34, 3 (2020), 342–362.

Makhortykh, M and Sydorova, M. (2017). Social media and visual framing of the conflict in Eastern Ukraine. *Media, War & Conflict* 10, 3 (2017), 359–381.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (1947), 153–157.

Mejova, Y, Zhang, A. X, Diakopoulos, N, and Castillo, C. (2014). Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152* (2014).

Moretti, F, van Vliet, L, Bensing, J, Deledda, G, Mazzi, M, Rimondini, M, Zimmermann, C, and Fletcher, I. (2011). A standardized approach to qualitative content analysis of focus group discussions from different countries. *Patient education and counseling* 82, 3 (2011), 420–428.

Morstatter, F, Wu, L, Yavanoglu, U, Corman, S. R, and Liu, H. (2018). Identifying framing bias in online news. *ACM Transactions on Social Computing* 1, 2 (2018), 1–18.

Mulder, M, Inel, O, Oosterman, J, and Tintarev, N. (2021). Operationalizing framing to support multiperspective recommendations of opinion pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 478–488.

Muralidharan, S, Rasmussen, L, Patterson, D, and Shin, J.-H. (2011). Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts. *Public Relations Review* 37, 2 (2011), 175–177.

Neuman, W. R, Neuman, R. W, Just, M. R, and Crigler, A. N. (1992). *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

Nitsova, S. (2021). Why the Difference? Donbas, Kharkiv and Dnipropetrovsk After Ukraine's Euromaidan Revolution. *Europe-Asia Studies* 73, 10 (2021), 1832–1856.

Nitsova, S, Pop-Eleches, G, and Robertson, G. (2018). Revolution and Reform in Ukraine. *Evaluating Four Years of Reform'Institute for European, Russian and Eurasian Studies at the George Washington University's Elliott School of International Affairs* (2018), 1–70.

Onuch, O. (2015). EuroMaidan protests in Ukraine: Social media versus social networks. *Problems of Post-Communism* 62, 4 (2015), 217–235.

Opperhuizen, A and Schouten, K. (2020). Dynamics and tipping point of issue attention in newspapers: quantitative and qualitative content analysis at sentence level in a longitudinal study using supervised machine learning and big data. *Quality & Quantity* (2020), 1–19.

Ott, M, Choi, Y, Cardie, C, and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 309–319.

Pantti, M. (2016). Seeing and not seeing the Syrian crisis: New visibility and the visual framing of the Syrian conflict in seven newspapers and their online editions. *JOMEC journal* 4 (2016).

Papacharissi, Z and de Fatima Oliveira, M. (2008). News frames terrorism: A comparative analysis of frames employed in terrorism coverage in US and UK newspapers. *The international journal of press/politics* 13, 1 (2008), 52–74.

Park, C. S. (2012). How the media frame political corruption: Episodic and thematic frame stories found in Illinois newspapers. In *Paper Originally Prepared for the Ethics and Reform Symposium on Illinois Government (September 27-28, 2012).*

Park, S, Ko, M, Kim, J, Choi, H.-J, Song, J, and others, . (2011). NewsCube 2.0: an exploratory design of a social news website for media bias mitigation. In *Workshop on Social Recommender Systems*.

Pranckevičius, T and Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing* 5, 2 (2017), 221.

Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of. *Brain Mechanisms* (1962), 555–559.

Saez-Trumper, D, Castillo, C, and Lalmas, M. (2013). Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1679–1684.

Semetko, H. A and Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of communication* 50, 2 (2000), 93–109.

Sethi, R. J. (2017). Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. 315–316.

Shariff, S. M, Zhang, X, and Sanderson, M. (2014). User perception of information credibility of news on twitter. In *European conference on information retrieval*. Springer, 513–518.

Stefanone, M. A, Vollmer, M, and Covert, J. M. (2019). In news we trust? Examining credibility and sharing behaviors of fake news. In *Proceedings of the 10th International Conference on Social Media and Society*. 136–147.

Sundermeyer, M, Schlüter, R, and Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Suykens, J. A and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters* 9, 3 (1999), 293–300.

Van Aelst, P, Strömbäck, J, Aalberg, T, Esser, F, De Vreese, C, Matthes, J, Hopmann, D, Salgado, S, Hubé, N, Stępińska, A, and others, . (2017). Political communication in a high-choice media environment: a challenge for democracy? *Annals of the International Communication Association* 41, 1 (2017), 3–27.

Vaswani, A, Shazeer, N, Parmar, N, Uszkoreit, J, Jones, L, Gomez, A. N, Kaiser, L, and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

Wall, M. (2015). Citizen Journalism: A retrospective on what we know, an agenda for what we don't. *Digital journalism* 3, 6 (2015), 797–813.

Wang, Z, He, Y, and Jiang, M. (2006). A comparison among three neural networks for text classification. In *2006 8th international Conference on Signal Processing*, Vol. 3. IEEE.

Wijekoon, H, Schegolev, B, and Merunka, V. (2019). Fighting Biased Online News: Lessons from Online Participation and Crowdsourcing. In *International Conference on e-Democracy*. Springer, 209–220.

Wu, S, Rizoiu, M.-A, and Xie, L. (2018). Beyond views: Measuring and predicting engagement in online videos. In *Twelfth International*

AAAI Conference on Web and Social Media.

Zhang, T and Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information retrieval* 4, 1 (2001), 5–31.