Framing the News: From Human Perception to Large Language Model Inferences

David Alonso del Barrio
ddbarrio@idiap.ch
Idiap Research Institute
Switzerland

Daniel Gatica-Perez
gatica@idiap.ch
Idiap Research Institute and EPFL
Switzerland

ABSTRACT

Identifying the frames of news is important to understand the articles' vision, intention, message to be conveyed, and which aspects of the news are emphasized. Framing is a widely studied concept in journalism, and has emerged as a new topic in computing, with the potential to automate processes and facilitate the work of journalism professionals. In this paper, we study this issue with articles related to the Covid-19 anti-vaccine movement. First, to understand the perspectives used to treat this theme, we developed a protocol for human labeling of frames for 1786 headlines of No-Vax movement articles of European newspapers from 5 countries. Headlines are key units in the written press, and worth of analysis as many people only read headlines (or use them to guide their decision for further reading.) Second, considering advances in Natural Language Processing (NLP) with large language models, we investigated two approaches for frame inference of news headlines: first with a GPT-3.5 fine-tuning approach, and second with GPT-3.5 prompt-engineering. Our work contributes to the study and analysis of the performance that these models have to facilitate journalistic tasks like classification of frames, while understanding whether the models are able to replicate human perception in the identification of these frames.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Human-centered computing** → *Text input*.

KEYWORDS

Covid-19 no-vax, news framing, GPT-3, prompt-engineering, transformers, large language models

ACM Reference Format:
David Alonso del Barrio and Daniel Gatica-Perez. 2023. Framing the News: From Human Perception to Large Language Model Inferences . In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592278

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592278

1 INTRODUCTION

In recent years, there has been a proliferation in the use of concepts such as data journalism, computational journalism, and computer-assisted reporting [15] [29], which all share the vision of bridging journalism and technology. The progress made in NLP has been gradually integrated into the journalistic field [5][8][54]. More specifically, machine learning models based on transformers have been integrated in the media sector in different tasks [41] such as the creation of headlines with generative languages models [17], summarization of news articles [28][27], false news detection [49], and topic modeling and sentiment analysis [25]. The development of large language models such as GPT-3 [9], BLOOM [51] or ChatGPT show a clear trend towards human-machine interaction becoming easier and more intuitive, opening up a wide range of research possibilities. At the same time, the use of these models is also associated with a lack of transparency regarding how these models work, but efforts are being made to bring some transparency to these models, and to analyze use cases where they can be useful and where they cannot [35]. Based on the premises that these models open up a wide range of research directions [7], and that at the same time (and needless to say) they are not the solution to all problems, we are interested in identifying use cases and tasks where they can be potentially useful, while acknowledging and systematically documenting their limitations [56]. More specifically, the aim of this work is to analyze the performance of GPT-3.5 for a specific use case, namely the analysis of frames in news, from an empirical point of view, with the objective of shedding light on a potential use of generative models in journalistic tasks.

Frame analysis is a concept from journalism, which consists of studying the way in which news stories are presented on an issue, and what aspects are emphasized: Is a merely informative vision given in an article? Or is it intended to leave a moral lesson? Is a news article being presented from an economic point of view? Or from a more human, emotional angle? The examples above correspond to different frames with which an article can be written.

The concept of news framing has been studied in computing as a step beyond topic modeling and sentiment analysis, and for this purpose, in recent years, pre-trained language models have been used for fine-tuning the classification process of these frames [60] [10], but the emergence of generative models opens the possibility of doing prompt-engineering of these classification tasks, instead of the fine-tuning approach investigated so far.

Our work aims to address this research gap by posing the following research questions:

RQ1: What are the main frames in the news headlines about the anti-vaccine movement, as reported in newspapers across 5 European countries?

RQ2: Can prompt engineering be used for classification of headlines according to frames?

By addressing the above research questions, our work makes the following contributions:

Contribution 1. We implemented a process to do human annotation of the main frame of 1786 headlines of articles about the Covid-19 no-vax movement, as reported in 19 newspapers from 5 European countries (France, Italy, Spain, Switzerland and United Kingdom.) At the headline level, we found that the predominant frame was human interest, where this frame corresponds to a personification of an event, either through a statement by a person, or the explanation of a specific event that happened to a person. Furthermore, we found a large number of headlines annotated as containing no frame, as they simply present information without entering into evaluations. We also found that for all the countries involved, the distribution of frame types was very similar, i.e., human interest and no frame are the two predominant frames. Finally, the generated annotations allowed to subsequently study the performance of a large language model.

Contribution 2. We studied the performance of GPT-3.5 on the task of frame classification of headlines. In addition to using the fine-tuning approach from previous literature, we propose an alternative approach for frame classification that requires no labeled data for training, namely prompt-engineering using GPT-3.5. The results show that fine-tuning with GPT-3.5 produces 72% accuracy (slightly higher than other smaller models), and that the prompt-engineering approach results in lower performance (49% accuracy.) Our analysis also shows that the subjectivity of the human labeling task has an effect on the obtained accufracy.

The paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe the news dataset. In Section 4, we describe the methodology for both human labeling and machine classification of news frames. We present and discuss results for RQ1 and RQ2 in Sections 5 and 6, respectively. Finally, we provide conclusions in Section 7.

2 RELATED WORK

Framing has been a concept widely studied in journalism, with a definition that is rooted in the study of this domain [23]: "To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."

For frame recognition, there are two main approaches: the inductive approach [16], where one can extract the frames after reading the article, and the deductive approach [38], where a predefined list of frames exists and the goal is to interpret if any of them appears in the article. In the deductive case, there are generic frames and subject-specific frames, and the way to detect them typically involves reading and identifying one frame at a time, or through answers to yes/no questions that represent the frames. Semetko et al. [52] used 5 types of generic frames (attribution of responsibility, human interest, conflict, morality, and economic consequences) based on previous literature, and they defined a list of 20 yes/no questions to detect frames in articles. For instance, the questions about morality are the following: "Does the story contain any moral message? Does the story make reference to morality, God, and other religious tenets? Does the story offer specific social prescriptions about how to behave?", and so on for each of the frame types. This categorization of frames has been used in various topics such as climate change [18] [19], vaccine hesitance [13], or immigration [34].

We now compare the two approaches on a common topic, such as Covid-19. Ebrahim et al. [21] followed an inductive approach in which the frames were not predefined but emerged from the text (e.g., deadly spread, stay home, what if, the cost of Covid-19) using headlines as the unit of analysis. In contrast, the deductive approach has studied very different labels. El-Behary et al. [22] followed the method of yes/no questions, but in addition to the 5 generic frames presented before, they also used blame frame and fear frame. Adiprasetio et al. [1] and Rodelo [50] used the 5 generic frames with yes/no questions, while Catalán-Matamoros et al. [14] used the 5 frames and read the headline and subheadline to decide the main frame. Table 1 summarizes some of the the existing approaches. This previous work showed how frame labels can be different, and also that frame analysis has been done at both headline and article levels. These two approaches (inductive and deductive) that originated in journalism have since been replicated in the computing literature.

We decided to follow the deductive approach because a predefined list of frames allows to compare among topics, countries, previous literature, and also because they represent a fixed list of labels for machine classification models. Furthermore, the inductive approach tends to be more specific to a topic, and from the computing viewpoint, past work has tried to justify topic modeling as a technique to extract frames from articles.

Ylä-Antitila et al. [60] proposed topic modeling as a frame extraction technique. They argued that topics can be interpreted as frames if three requirements are met: frames are operationalized as connections between concepts; subject-specific data is selected; and topics are adequately validated as frames, for which they suggested a practical procedure. This approach was based on the choice of a specific topic (e.g., climate change) and the use of Latent Dirichlet Allocation (LDA) as a technique to extract a number of subtopics. In a second phase, a qualitative study of the top 10 words of each subtopic was performed, and the different subtopics were eliminated or grouped, reducing the number and establishing a tentative description. In a third phase, the top 10 articles belonging to that frame/topic were taken, and if the description of the topic fitted at least 8 of the 10 articles, that topic/frame remained. The frames found in this article were: green growth, emission cuts, negotiations and treaties, environmental risk, cost of carbon emissions, Chinese emissions, economics of energy production, climate change, environmental activism, North-South burden sharing, state leaders negotiating, and citizen participation.

From Entman's definition of frame [23], it seems that the deductive approach is more refined than the inductive approach (which seems to resemble the detection of sub-themes.) For example, with regard to climate change, there are stories on how people have been affected by climate change from an emotional point of view, thus personalizing the problem. In this case, we could categorize the corresponding frame as human interest, as the writer of the article is selecting "some aspects of a perceived reality and make them

more salient". The language subtleties with which news articles are presented cannot be captured with basic topic modeling.

Isoaho et al.[30] held the position that while the benefits of scale and scope in topic modeling were clear, there were also a number of problems, namely that topic outputs do not correspond to the methodological definition of frames, and thus topic modeling remained an incomplete method for frame analysis. Topic modeling, in the practice of journalistic research, is a useful technique to deal with the large datasets that are available, yet is often not enough to do more thorough analyses [31]. In our work, we clearly notice that frame analysis is not topic modeling. For example, two documents could be about the same topic, say Covid-19 vaccination, but one article could emphasize the number of deaths after vaccination, while the other emphasized the role of the vaccine as a solution to the epidemic.

We also consider that the larger the number of possible frame types, the more likely it is to end up doing topic modeling instead of frame analysis. Using a deductive approach, Dallas et al. [12] created a dataset with articles about polemic topics such as immigration, same sex marriage, or smoking, and they defined 15 types of frames: "economic, capacity and resources, morality, fairness and equality, legality, constitutionality and jurisprudence, policy prescription and evaluation, crime and punishment, security and defense, health and safety, quality of life, cultural identity, political, external regulation and reputation, other". In this case, they authors did not use a list of questions. Instead, for each article, annotators were asked to identify any of the 15 framing dimensions present in the article and to label text blurbs that cued them (based on the definitions of each of the frame dimensions) and decide the main frame of each article. In our case, we followed the idea of detecting the main frame by reading the text instead of answering questions, but instead of using the 15 frames proposed in [12] , we used the 5 generic frames proposed in [52].

A final decision in our work was the type of text to analyze, whether headlines or whole article. For this decision, the chosen classification method was also going to be important. For example, Khanehzar et al. [33] used traditional approaches such as SVMs as baseline, and demonstrated the improvement in frame classification with the use of pre-trained languages models such as BERT, RoBERTa and XLNet, following a fine-tuning approach, setting as input text a maximum of 256 tokens (although the maximum number of input tokens in these models is 512 tokens.) Liu et al. [37] classified news headlines about the gun problem in the United States, arguing for the choice of headlines as a unit of analysis based on previous journalism literature [6], [44], that advocated for the importance and influence of headlines on readers and the subsequent perception of articles. From a computational viewpoint, using headlines is also an advantage, since you avoid the 512 token limitation in BERT-based models. Therefore, we decided to work with headlines about a controversial issue, namely the Covid-19 no-vax movement.

Continuing with the question of the methods used for classification, much work has been developed in prompt engineering, especially since the release of GPT-3. Liu et al.[36] presented a good overview of the work done on this new NLP paradigm, not only explaining the concept of prompt engineering, but also the different strategies that can be followed both in the design of prompts, the potential applications, and the challenges to face when using this approach. Prompt engineering applications include knowledge probing [46], information extraction [53], NLP reasoning [57], question answering [32], text generation [20], multi-modal learning [58], and text classification [24], the latter being the prompt-engineering use case in our work. Puri et al.[45] presented a very interesting idea that we apply to our classification task. This consists of providing the language model with natural language descriptions of classification tasks as input, and training it to generate the correct answer in natural language via a language modeling objective. It is a zero-shot learning approach, in which no examples are used to explain the task to the model. Radford et al. [48] demonstrated that language models can learn tasks without any explicit supervision. We have followed this approach to find an alternative way to do frame analysis.

As mentioned before, the emergence of giant models like GPT-3, BLOOM, and ChatGPT are a very active research topic. To the best of our knowledge, on one hand our work extends the computational analysis of news related to the covid-19 no-vax movement, which illustrates the influence of the press on the ways societies think about relevant issues [40], [59], and on the other hand it adds to the literature of human-machine interaction, regarding the design of GPT-3 prompts for classification tasks [39], [2].

Table 1: Summary of deductive approaches for frame analysis

Ref	Frames	Goal	Technique	Number of samples
[12]	15 generic frames: "Economic", "Capacity and resources", "Morality", "Fairness and equality", "Legality, constitutionality and jurisprudence", "Policy prescription and evaluation", "Crime and punishment", "Security and defense", "Health and safety", "Quality of life", "Cultural identity", "Public opinion", "Political", "External regulation and reputation", "Other".	To label frames of full articles	Reading the full article, the annotator defines the main frame	20000 articles
[33]	15 generic frames	Classification	BERT based models	12000 articles
[52]	5 generic frames: "human interest", "conflict", "morality", "attribution of responsibility", and "economic consequences".	To label frames of full articles	Yes/No questions.	2600 articles and 1522 tv news stories
[37]	9 specific frames:"Politics", "Public opinion", "Society/Culture", and "Economic consequences" , "2nd Amendment" (Gun Rights), "Gun control/regulation", "Mental health", "School/Public space safety", and "Race/Ethnicity".	To label frames of full articles/ Classification	Reading the full article, the annotator defines the main frame. BERT based models	2990 headlines
[22]	5 generic frames + blame frame and fear frame	To label frames of full articles	Yes/No questions.	1170 articles
[1]	5 generic frames	To label frames of full articles	Reading the full article, the annotator defines the main frame.	6713 articles
[50]	5 generic frames + pandemic frames	To label frames of full articles	Yes/No questions.	2742 articles
[14]	5 generic frames, journalistic role and pandemic frames	To label frames of full articles	Reading the headline and subheadline, the annotator defines the main frame.	131 headlines + subheadlines

3 DATA: EUROPEAN COVID-19 NEWS DATASET

We used part of the European Covid-19 News dataset collected in our recent work [3]. This dataset contains 51320 articles on Covid-19 vaccination from 19 newspapers from 5 different countries: Italy,

France, Spain, Switzerland and UK. The articles cover a time period of 22 months, from January 2020 to October 2021. All content was translated into English to be able to work in a common language. The dataset was used for various analyses, such as name entity recognition, sentiment analysis, and subtopic modeling, to understand how Covid-19 vaccination was reported in Europe through the print media (in digital format.) The subtopic modeling analysis revealed a subsample of articles on the no-vax movement, which is the one we have used in this paper. We took the headlines of the articles associated with the no-vax movement, selecting all articles containing any of the keywords in Table 2 in the headline or in the main text. This corresponds to a total of 1786 headlines.

Table 2: Keywords used to identify no-vax articles

	Keywords
NO VAX TOPIC	"anti-vaxxers", "anti-vaccine", "anti-vaxx", "anti-corona", "no-vax", "no vax","anti-vaccin"

In Table 3, we show the number of headlines per country and newspaper. France is the country with the most no-vax articles in the corpus, with 523 articles, followed by Italy with 508. However, note that there are 6 newspapers from France, while only 2 from Italy. Corriere della Sera is the newspaper that dealt most frequently with the subject (429 articles), while The Telegraph is the second one (206 articles). The total number of articles normalized by the number of newspapers per country is also shown in the last column of the Table. Using these normalized values, the ranking is Italy, UK, France, Switzerland, and Spain.

Table 3: Number of headlines by newspaper and country

COUNTRY	NEWSPAPER	HEADLINES	TOTAL (NORM. TOTAL)
FRANCE	La Croix	94	523 (87.1)
	Le Monde	125	
	Les Echos	49	
	Liberation	97	
	Lyon Capitale	8	
	Ouest France	150	
ITALY	Corriere della Sera	429	508 (254.0)
	Il Sole 24 Ore	79	
SPAIN	20 minutos	27	303 (50.5)
	ABC	50	
	El Diario	32	
	El Mundo	77	
	El Español	22	
	La Vanguardia	95	
SWITZERLAND	24 heures	97	230 (76.6)
	La Liberté	22	
	Le Temps	111	
UNITED KINGDOM	The Irish News	16	222 (111.0)
	The Telegraph	206	
			1786

4 METHODOLOGY

4.1 Human labeling of news frames

To carry out the labeling of the frames in our corpus of headlines, we first designed a codebook, which contained the definitions of each of the frame types and a couple of examples of each type, as well as a definition of the corpus subject matter and definitions of the concept of frame analysis, so that the annotators could understand the task to be performed. The codebook follows the proposed by [52] with 5 generic frames (attribution of responsibility, human interest, conflict, morality, and economic consequences) plus one additional 'no-frame' category. Two researchers were engaged to annotate a sample of the collected newspaper articles following a three-phase training procedure.

In the first phase, annotators had to read the codebook and get familiar with the task. In the second phase, they were asked to identify the main frame in the same subset of 50 headlines. At the end of the second phase, the intercoder reliability (ICR) was 0.58 between the 2 annotators. We analyzed those cases where there were discrepancies, and observed that in some cases, there was not a unique main frame, because both annotators had valid arguments to select one of the frames. In other cases, the discrepancies were due to slight misunderstanding of the definitions. In the third phase, the annotators coded again 50 headlines, and the ICR increased to was 0.66. We realized that the possibility of having two frames remained. They discussed the cases in which they had disagreed, and if the other person's arguments were considered valid, it could be said that there were two frames. After this three-phase training procedure, annotators were ready to annotate the dataset independently. We divided the dataset into two equal parts, and each person annotated 893 headlines.

4.2 Fine-tuning GPT-3.5 and BERT-based models

With the annotated dataset, we investigated two NLP approaches: the first one involves fine-tuning a pre-trained model; the second one is prompt engineering. Pre-trained language models have been

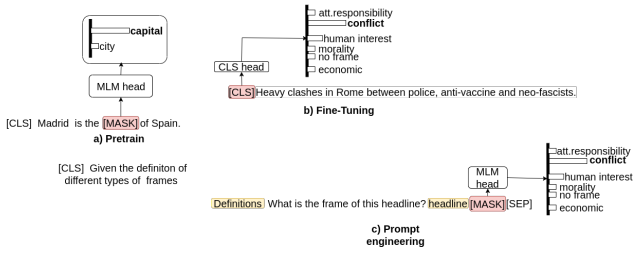

Figure 1: Pre-train, fine-tune, prompt

trained with large text strings based on two unsupervised tasks, next sentence prediction and masked language model. Figure 1 summarizes these techniques.

In the first approach, a model with a fixed architecture is pre-trained as a language model (LM), predicting the likelihood of the observed textual data. This can be done due to the availability of large, raw text data needed to train LMs. This learning process can produce general purpose features of the modeled language. The learning process produces robust, general-purpose features of the language being modeled. The above pre-trained LM is then adapted to different downstream tasks, by introducing additional parameters and adjusting them using task-specific objective functions. In this approach, the focus was primarily on goal engineering, designing the training targets used in both the pre-training and the fine-tuning stages [36].

We present an example to illustrate the idea. Imagine that the task is sentiment analysis, and we have a dataset with sentences and their associated sentiment, and a pre-trained model, which is a saved neural network trained with a much larger dataset. For that pre-trained model to address the target task, we unfreeze a few of the top layers of the saved model base and jointly train both the newly-added classifier layers and the last layers of the base model. This allows to "fine-tune" the higher-order feature representations in the base model to make them more relevant for the sentiment analysis task. In this way, instead of having to obtain a very large dataset with target labels to train a model, we can reuse the pre-trained model and use a much smaller train dataset. We use a part of our dataset as examples for the model to learn the task, while the other part of the dataset is used to evaluate model performance.

Previous works related to frame classification in the computing literature have used fine-tuning, BERT-based models. In our work, we have done the same as a baseline, but we aimed to go one step further and also produce results using fine-tuning of GPT-3.5.

4.3 Prompt-engineering with GPT-3.5

Model fine-tuning has been widely used, but with the emergence of generative models such as GPT-3, another way to approach classification tasks has appeared. The idea is to use the pre-trained model directly and convert the task to be performed into a format as close as possible to the tasks for which it has been pre-trained. That is, if the model has been pre-trained from next word prediction as in the case of GPT-3, classification can be done by defining a prompt, where the input to the model is an incomplete sentence, and the model must complete it with a word or several words, just as it has been trained. This avoids having to use part of the already labeled dataset to teach the task to be performed to the model, and a previous labeling is not needed [36].

In this approach, instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are reformulated to look more like those solved during the original LM training with the help of a textual prompt. For example, when recognizing the emotion of a social media post, "I missed the bus today.", we may continue with a prompt "I felt so _", and ask the LM to fill the blank with an emotion-bearing word. Or if we choose the prompt "English: I missed the bus today. French: _"), an LM may be able to fill in the blank with a French translation. In this way, by selecting the appropriate prompts, we can influence the model behavior so that the pre-trained LM itself can be used to predict the desired output, even without any additional task-specific training [36].

We use this emerging NLP approach to classify frames at headline level. We are not aware of previous uses of this strategy to classify frames as we propose here. The idea is the following. Prompt engineering consists of giving a prompt to the model, and understands that prompt as an incomplete sentence. To do prompt engineering with our dataset, we needed to define an appropriate prompt that would produce the headline frames as output. We defined several experiments with the Playground of GPT-3, in order to find the best prompt for our task. In our initial experiments, we followed existing approaches in prompt engineering to do sentiment analysis, where the individual answer was an adjective, and this adjective was matched with a sentiment. In a similar fashion, we decided to build a thesaurus of adjectives that define each of the frames. For instance, the human interest frame could be 'interesting', 'emotional', 'personal', 'human'. The conflict frame could be: 'conflictive', 'bellicose', 'troublesome', 'rowdy', 'quarrelsome', 'troublemaker', 'agitator', etc. After the list of adjectives was defined, we needed to define the prompt in order to get, as an answer, one of the adjectives in our thesaurus to match them with the frame. We used the GPT-3 playground using the headline as input and asking for the frame as output, but the strategy did not work. In our final experiment, instead of giving the headline as input, we gave the definitions of each type of frame plus the headline, and we asked the model to choose between the different types of frames as output. In this way, the output of the model was directly one of the frames, and we avoided the step of matching adjectives with frames. An example is shown in Figure 2.

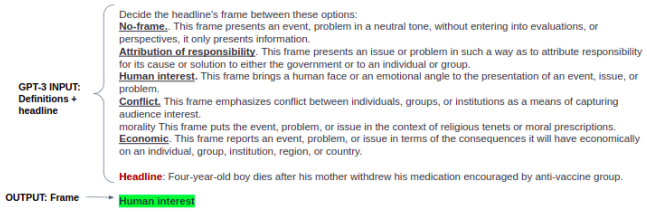

Figure 2: GPT-3.5 for frame inference: input and output

For the GPT-3 configuration [1], there are 3 main concepts:

- TEMPERATURE [0-1]. This parameter controls randomness, lowering it results in less random completions.
- TOP_P [0-1]. This parameter controls diversity via nucleus sampling.
- MAX_TOKENS[1-4000]. This parameter indicates the maximum number of tokens to generate,
- MODEL. GPT-3 offer four main models with different levels of power, suitable for different tasks. Davinci is the most capable model, and Ada is the fastest.

After testing with the GPT-3 playground and varying different hyper-parameters to assess performance, we set the temperature to 0, since the higher the temperature the more random the response. Furthermore, the Top-p parameter was set to 1, as it would likely get a set of the most likely words for the model to choose from. The maximum number of tokens was set to 2; in this way, the model is asked to choose between one of the responses. As a model, we used the one with the best performance at the time of experimental design, which was TEXT-DAVINCI-003, recognized as GPT 3.5.

5 RESULTS: HUMAN LABELING OF FRAMES IN NO-VAX NEWS HEADLINES (RQ1)

In this section, we present and discuss the results of the analysis related to our first RQ.

Figure 3 shows the distribution of frames per country at headline level, with human interest and no-frame being the predominant

[1] https://beta.openai.com/docs/introduction

ones. Attribution of responsibility is the third one except in Switzerland, where the corresponding frame is conflict. Finally, morality and economic are the least represented in the dataset for every country.

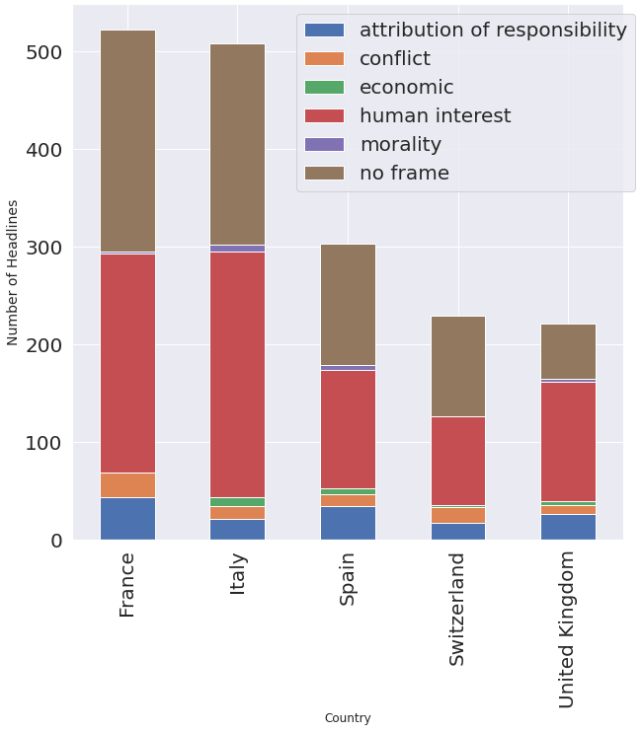

Figure 3: Non-normalized distribution of frames per country

The monthly distribution of frames aggregated for all countries is shown in Fig. 4. We can see two big peaks, the first one in January 2021 and the second one in August 2021. In all countries, the vaccination process started at the end of December 2020, so it makes sense that the no-vax movement started to be more predominant in the news in January 2021. Human interest is the most predominant frame. Manual inspection shows that this is because the headlines are about personal cases of people who are pro- or anti- vaccine. Attribution of responsibility is also present. Manual inspection indicates that local politicians and health authorities had to make decisions about who could be vaccinated at the beginning of the process. The second peak at the end of summer 2021 coincided with the health pass (also called Covid passport in some countries), and we can observe a peak in the curve corresponding to the conflict frame, reflecting the demonstrations against the measure of mandatory health passes taken by country governments.

In Figure 5, we compare the sentiment per frame and per country, to understand if there were any major differences. The sentiment analysis labels were obtained using BERT-sent from the Hugging Face package [47], used in our previous work (please refer to our original analysis in [3] for details.) We normalized the results between 0 and 1 to compare frames between countries. We see that the sentiment is predominantly neutral (in blue). Examining in more

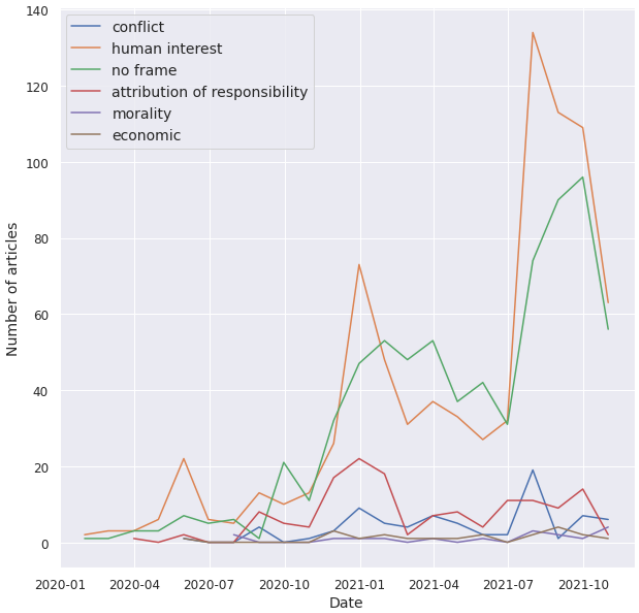

Figure 4: Non-normalized monthly distribution of frames.

detail the negative and positive sentiment of each frame category, we observed a few trends:

- Attribution of responsibility: Negative sentiment represents 30-40% of the cases, while positive tone is only found in residual form in Italy, Switzerland, and the United Kingdom.
- Conflict: Negative sentiment represents 20-35% of the cases.
- Economic: Predominantly neutral, with only negative tone in Italy and UK (in the latter case, all headlines with this frame were considered negative.)
- Human interest: Negative sentiment represents 30-40% of the cases, while positive tone is only found in residual form in Italy, Spain, and Switzerland.
- Morality: Predominantly neutral, with negative tone in Italy, Switzerland, and the United Kingdom,
- No frame: 20-30% of negative content.

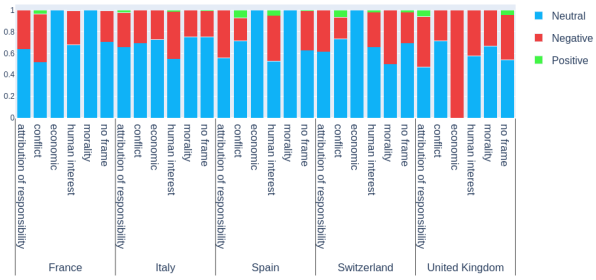

Figure 5: Sentiment of headline by frame and by country

Regarding the results of the annotation process, the fact that the distribution of the 6 frame types is relatively similar between countries suggests that the anti-vaccine movement issue was treated in a similar way in these countries. The fact that human interest is the most dominant frame indicates that this issue was treated from a more human and emotional approach, with headlines about personal experiences, celebrities giving their opinion about vaccination, and politicians defending vaccine policies. Moreover, the reason for many headlines being classified as no-frame is partly due to how data was selected. We chose articles that contained words related to no-vax, either in the headline or in the article. This resulted in many headlines not containing anything specific related to no-vax, while the no-vax content was actually included in the main text of the corresponding articles.

It is worth mentioning that prior to obtaining the results, we had expected that attribution of responsibility would be among the most prominent frames, since governments took many measures such as mandatory health pass requirements to access certain sites; we had also expected that the conflict frame would be prominent, since there were many demonstrations in Europe. In reality, however, these frames categories were not reflected as frequently at the headline level.

Regarding the analysis at the temporal level, it is clear that certain events were captured by the press, such as the start of vaccination or the mandatory vaccination passport.

Finally, the sentiment analysis of the different frames shows that the predominant tone in all of them is neutral or negative, with very similar trends between countries. This association between sentiment analysis and frames has been discussed in previous literature [11] [43].

6 RESULTS: GPT-3.5 FOR FRAME CLASSIFICATION OF HEADLINES (RQ2)

Here, we present and discuss the results related to our second RQ.

6.1 Fine-tuning GPT-3.5

Table 4 shows the results of the 6-class classification task using 5-cross validation. Three models were used: GPT-3.5 and two BERT-based models. We observe that, on average, GPT-3.5 performs better than the BERT-based models. This is somehow expected as GPT-3.5 is a much larger model. Overall, in the case of fine-tuning, the best performance for the six-class frame classification task is 72% accuracy, which is promising, with an improvement over previous models based on BERT. Yet, it should be noted that the performance differences are modest (2% improvement between GPT-3.5 and RoBERTa).

Table 4: Classification results for six-class frame classification and 5-fold cross validation

ACCURACY	0	1	2	3	4	AVERAGE
BERT	0.68	0.69	0.72	0.64	0.70	0.67
RoBERTa	0.70	0.72	0.72	0.67	0.71	0.70
GPT3	0.75	0.70	0.72	0.71	0.71	0.72

On the other hand, BERT is open-source, while GPT-3 has an economic cost as the use of the model is not free, which monetarily limits the number of experiments that can be performed with it, as well as the different configurations one can explore to improve performance. This is important because much of the improvement in performance requires empirical explorations of model parameters More specifically, the cost of an experiment for each of the folds has a cost of 4 dollars (at the time of writing this paper.) This represents a limitation in practice.

Furthermore, GPT-3 has a significant carbon footprint. Similarly, for prompt engineering (discussed in the next subsection), choosing the right prompt (i.e., the words that best define the task so that the model is able to perform adequately) is also based on trial and error. This also has an impact on carbon footprint. In connection with this topic, Strubell et al.[55] argue that improvements in the accuracy of models depend on the availability of large computational resources, which involve large economic and environmental costs. A criticism has been made as 'the rich get richer', in the sense that not all research groups have sufficient infrastructure resources and access to funding needed to use these models and improve their performance. Also in relation to this analysis, the work of Bender et al. [4] evaluates the costs and risks of the use of large language models, stating that researchers should be aware of the impact that these models have on the environment, and assess whether the benefits outweigh the risks. The work in [4] provides a very telling example, where people living in the Maldives or Sudan are affected by floods and pay the environmental price of training English LLMs, when similar models have not been produced for languages like Dhivehi or Sudanese Arab. In short, there is a need to establish ways to use this technological development responsibly, and it all starts with being aware of the risks it presents.

6.2 Prompt-engineering with GPT-3.5

For each headline, we got the frame that the model considered the most likely, and we compared these GPT-3.5 inferences with the frames labeled by the annotators. The agreement between model and annotator was of 49%. Analyzing the results, and specifically looking at the cases where the annotator and GPT-3.5 disagreed, we discovered that according to the frame definitions, the model in some cases proposed a frame that indeed made sense. This observation, together with our previous experience in the annotation process, where headlines could have more than one valid frame, led us to design a second post-hoc experiment. We took all the headlines where each of the two annotators had disagreed with GPT-3.5, and we asked the annotators to state whether they would agree (or not) with each GPT-inferred label for a given headline. It is important to emphasize that the annotators did not know the origin of that label, i.e., they did not know if it was the label they had originally assigned, or if it was a random one. In this way, we could quantify how GPT-3.5 worked according to valid arguments provided by the annotators. In this post-hoc experiment, the model agreed in 76% of cases with the annotators.

Looking at the results of the classification models, the 49% accuracy of the prompt-engineering approach can be considered low, yet we consider that it is a valid avenue for further investigation, as in the second post-hoc analysis, we found that the model agrees

with human annotators in 76% of the cases. Clearly, framing involves aspects of subjectivity [42]. Much of what we do as people has a subjective component, influenced by how we feel or how we express opinions.

News reading is never fully objective, and the annotators engaged in the frame classification task, influenced by their personal state of mind, experience, and culture, may perceive information differently. Monarch affirms that "for simple tasks, like binary labels on objective tasks, the statistics are fairly straightforward to decide which is the 'correct' label when different annotators disagree. But for subjective tasks, or even objective tasks with continuous data, there are no simple heuristics for deciding what the correct label should be" [42].

Subjectivity is involved in both the generation and perception of information: the assumption that there is only one frame is complicated by the point of view of the reader. In the case of news, the information sender (the journalist) has an intention, but the receiver (the reader) plays a role and is influenced by it. In psychology, this is known as the lens model of interpersonal communication, where the sender has certain objectives, but the receiver can interpret or re-interpret what the sender wants to say, with more or less accuracy [26].

Following this discussion on subjectivity, the question arose as to what would happen if, instead of headlines, we used the complete article as a source of analysis. We wondered if longer text could make the frame labeling task clearer than when using headlines. Yet another possible hypothesis is that having to read longer texts could lead to the same subject being presented from different angles. Please recall that in the existing literature discussed in Section 2, both headlines and full articles have been used from frame analysis (see Table 1.) This remains as an issue for future work.

7 CONCLUSIONS

In this paper, we first presented an analysis of human-generated news frames on the covid-19 no-vax movement in Europe, and then studied different approaches using large language models for automatic inference of frames. We conclude by answering the two research questions we posed:

RQ1: What are the main frames in the news headlines about the covid-19 anti-vaccine movement in 5 European countries? After annotating the headlines, we found that of the 1786 headlines, the predominant frame is human interest (45.3% of cases), which presents a news item with an emotional angle, putting a face to a problem or situation. We also found that a substantial proportion of headlines were annotated as not presenting any frame (40.2% of cases). Finally, the other frame types are found more infrequently.

RQ2: Can prompt engineering be used for classification of headlines according to frames? We first used fine-tuning of a number of language models, and found that GPT-3.5 produced classification accuracy of 72% on a six-frame classification task. This represented a modest 2% improvement over BERT-based models, at a significantly larger environmental cost. We then presented a new way of classifying frames using prompts. At the headline level, inferences made with GPT-3.5 reached 49% of agreement with human-generated frame labels. In many cases, the GPT-3.5 model inferred frame types that were considered as valid choices by human annotators, and in an post-doc experiment, the human-machine agreement reached 76%. These results have opened several new directions for future work.

ACKNOWLEDGMENTS

This work was supported by the AI4Media project, funded by the European Commission (Grant 951911) under the H2020 Programme ICT-48-2020. We also thank the newspapers for sharing their online articles. Finally, we thank our colleagues Haeeun Kim and Emma Bouton-Bessac for their support with annotations, and Victor Bros and Oleksii Polegkyi for discussions.

REFERENCES

[1] Justito Adiprasetio and Annissa Winda Larasati. 2020. Pandemic crisis in online media: Quantitative framing analysis on Detik. com's coverage of Covid-19. *Jurnal Ilmu Sosial Dan Ilmu Politik* 24, 2 (2020), 153–170.

[2] Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, et al. 2021. RAFT: A real-world few-shot text classification benchmark. *arXiv preprint arXiv:2109.14076* (2021).

[3] David Alonso del Barrio and Daniel Gatica-Perez. 2022. How Did Europe's Press Cover Covid-19 Vaccination News? A Five-Country Analysis. (2022), 35–43. https://doi.org/10.1145/3512732.3533588

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (2021), 610–623.

[5] Santosh Kumar Biswal and Nikhil Kumar Gouda. 2020. Artificial intelligence in journalism: A boon or bane? In *Optimization in machine learning and applications*. Springer, 155–167.

[6] Erik Bleich, Hannah Stonebraker, Hasher Nisar, and Rana Abdelhamid. 2015. Media portrayals of minorities: Muslims in British newspaper headlines, 2001–2012. *Journal of Ethnic and Migration Studies* 41, 6 (2015), 942–962.

[7] Michael Bommarito and Daniel Martin Katz. 2022. GPT Takes the Bar Exam. https://doi.org/10.48550/ARXIV.2212.14402

[8] Meredith Broussard, Nicholas Diakopoulos, Andrea L Guzman, Rediet Abebe, Michel Dupagne, and Ching-Hua Chuan. 2019. Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly* 96, 3 (2019), 673–695.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten De Rijke, and Claes H De Vreese. 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.

[11] Bjorn Burscher, Rens Vliegenthart, and Claes H de Vreese. 2016. Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review* 34, 5 (2016), 530–545.

[12] Dallas Card, Amber Boydstun, Justin Gross, Philip Resnik, and Noah Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. 2 (01 2015), 438–444. https://doi.org/10.3115/v1/P15-2072

[13] Daniel Catalan-Matamoros and Carlos Elías. 2020. Vaccine hesitancy in the age of coronavirus and fake news: analysis of journalistic sources in the Spanish quality press. *International Journal of Environmental Research and Public Health* 17, 21 (2020), 8136.

[14] Daniel Catalán-Matamoros and Carmen Peñafiel-Saiz. 2019. Media and mistrust of vaccines: a content analysis of press headlines. *Revista latina de comunicación social* 74 (2019), 786–802.

[15] Mark Coddington. 2015. Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital journalism* 3, 3 (2015), 331–348.

[16] Stephen D Cooper. 2010. The oppositional framing of bloggers. In *Doing News Framing Analysis*. Routledge, 151–172.

[17] Robert Dale. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27, 1 (2021), 113–118.

[18] Astrid Dirikx and Dave Gelders. 2010. To frame is to explain: A deductive frame-analysis of Dutch and French climate change coverage during the annual UN Conferences of the Parties. *Public Understanding of Science* 19, 6 (2010), 732–742. https://doi.org/10.1177/0963662509352044 arXiv:https://doi.org/10.1177/0963662509352044 PMID: 21560546.

[19] Astrid Dirikx and Dave Gelders. 2010. To frame is to explain: A deductive frame-analysis of Dutch and French climate change coverage during the annual UN Conferences of the Parties. *Public understanding of science* 19, 6 (2010), 732–742.

[20] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014* (2020).
[21] Sumayya Ebrahim. 2022. The corona chronicles: Framing analysis of online news headlines of the COVID-19 pandemic in Italy, USA and South Africa. *Health SA Gesondheid (Online)* 27 (2022), 1–8.
[22] Hend Abdelgaber Ahmed El-Behary. 2021. A Feverish Spring: A Comparative Analysis of COVID-19 News Framing in Sweden, the UK, and Egypt. (2021).
[23] Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory* 390 (1993), 397.
[24] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).
[25] Piyush Ghasiya and Koji Okamura. 2021. Investigating COVID-19 news across four nations: a topic modeling and sentiment analysis approach. *Ieee Access* 9 (2021), 36645–36656.
[26] Robert Gifford. 1994. A Lens-Mapping Framework for Understanding the Encoding and Decoding of Interpersonal Dispositions in Nonverbal Behavior. *Journal of Personality and Social Psychology* 66 (02 1994), 398–412. https://doi.org/10.1037//0022-3514.66.2.398
[27] Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1792–1810.
[28] Anushka Gupta, Diksha Chugh, Rahul Katarya, et al. 2022. Automated news summarization using transformers. In *Sustainable Advanced Computing*. Springer, 249–259.
[29] Alfred Hermida and Mary Lynn Young. 2017. Finding the data unicorn: A hierarchy of hybridity in data and computational journalism. *Digital Journalism* 5, 2 (2017), 159–176.
[30] Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä. 2021. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal* 49, 1 (2021), 300–324.
[31] Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers. 2016. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital journalism* 4, 1 (2016), 89–106.
[32] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.
[33] Shima Khanehzar, Andrew Turpin, and Gosia Mikołajczak. 2019. Modeling Political Framing Across Policy Issues and Contexts. In *ALTA*.
[34] Jeesun Kim and Wayne Wanta. 2018. News framing of the US immigration debate during election years: Focus on generic frames. *The Communication Review* 21, 2 (2018), 89–115.
[35] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
[36] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. (2021). https://doi.org/10.48550/ARXIV.2107.13586
[37] Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*.
[38] Jörg Matthes and Matthias Kohring. 2008. The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication* 58 (06 2008). https://doi.org/10.1111/j.1460-2466.2008.00384.x
[39] Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E Losada. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–6.
[40] Stuart E Middleton, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Social computing for verifying social media content in breaking news. *IEEE Internet Computing* 22, 2 (2018), 83–89.
[41] Marko Milosavljević and Igor Vobič. 2021. 'Our task is to demystify fears': Analysing newsroom management of automation in journalism. *Journalism* 22, 9 (2021), 2203–2221.
[42] R. Monarch. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Manning. https://books.google.ch/books?id=LCh0zQEACAAJ
[43] Tom Nicholls and Pepper D Culpepper. 2021. Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication* 38, 1-2 (2021), 159–181.
[44] Zhongdang Pan and Gerald M Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication* 10, 1 (1993), 55–75.
[45] Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165* (2019).
[46] Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599* (2021).
[47] Rabindra Lamsal. 2021. Sentiment Analysis of English Tweets with BERTsent. https://huggingface.co/rabindralamsal/finetuned-bertweet-sentiment-analysis.
[48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[49] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. 2022. Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering* 3 (2022), 98–105. https://doi.org/10.1016/j.ijcce.2022.03.003
[50] Frida V Rodelo. 2021. Framing of the Covid-19 pandemic and its organizational predictors. *Cuadernos. info* 50 (2021), 91–112.
[51] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).
[52] Holli Semetko and Patti Valkenburg. 2000. Framing European Politics: A Content Analysis of Press and Television News. *Journal of Communication* 50 (06 2000), 93 – 109. https://doi.org/10.1111/j.1460-2466.2000.tb02843.x
[53] Richard Shin, Christopher H Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. *arXiv preprint arXiv:2104.08768* (2021).
[54] Efstathios Sidiropoulos and Andreas Veglis. 2017. Computer Supported Collaborative Work trends on Media Organizations: Mixing Qualitative and Quantitative Approaches. *Studies in Media and Communication* 5 (04 2017), 63. https://doi.org/10.11114/smc.v5i1.2279
[55] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).
[56] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503* (2021).
[57] Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847* (2018).
[58] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.
[59] Sandra A Vannoy and Prashant Palvia. 2010. The social influence model of technology adoption. *Commun. ACM* 53, 6 (2010), 149–153.
[60] Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2022. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication* 18, 1 (2022), 91–112.