Human Interest or Conflict? Leveraging LLMs for Automated Framing Analysis in TV Shows

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

Max Tiel
mtiel@beeldengeluid.nl
Nederlands Instituut voor Beeld en
Geluid
The Netherlands

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

ABSTRACT

[1] In the current media landscape, understanding the framing of information is crucial for critical consumption and informed decision making. Framing analysis is a valuable tool for identifying the underlying perspectives used to present information, and has been applied to a variety of media formats, including television programs. However, manual analysis of framing can be time-consuming and labor-intensive. This is where large language models (LLMs) can play a key role. In this paper, we propose a novel approach to use prompt-engineering to identify the framing of spoken content in television programs. Our findings indicate that prompt-engineering LLMs can be used as a support tool to identify frames, with agreement rates between human and machine reaching up to 43%. As LLMs are still under development, we believe that our approach has the potential to be refined and further improved. The potential of this technology for interactive media applications is vast, including the development of support tools for journalists, educational resources for students of journalism learning about framing and related concepts, and interactive media experiences for audiences.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

media, TV, LLMs, prompt-engineering, framing analysis

---

1 INTRODUCTION

Framing analysis is a technique for understanding how media content shapes our perception of the world, as it helps media audiences identify the underlying perspectives that are used to present information [14]. Identifying frames in news stories allows to understand what aspects of the story are being emphasized. For example, given two news items about the same topic such as the war between Russia and Ukraine, one may emphasize the economic repercussions in Europe due to this event (economic frame), while the other may show the personal case of a family displaced by the war and left homeless (human interest frame). Understanding these frames can support media readers and viewers to act as more critical consumers of media and to make informed decisions about the consumed information [1].

In recent years there have been great advances in the combination of journalism and computing. Concepts closely related to interactivity such as Automated Journalism are emerging [15, 41], further emphasised by the latest developments in artificial intelligence [10, 30, 39].

The development of Large Language Models (LLMs) has opened up new possibilities for media content analysis [11]. LLMs are large neural networks trained on large amounts of text data. Recently, they have proven to be effective in a variety of domains, including journalism [27, 33] and natural language processing tasks, such as text classification, machine translation or summarization [19, 46, 47]. The use of Large Language Models (LLMs) for frame analysis is a relatively new field of research, but it is attracting increasing interest as LLMs are effective in identifying frames in text and can be applied to a variety of media formats, such as headlines, tweets, or news articles [3, 4, 24]. This has ignited a growing interest in their potential application to spoken content, such as television transcripts, which is the focal point of our work. Framing analysis plays a pivotal role in comprehending how news programs shape our perception of reality. By identifying the underlying frames employed in newscasts, we can gain insights into the perspectives presented and the potential influence on public opinion [18]. However, conventional framing analysis methods are often tedious and labor-intensive [7], thus restricting their applicability to large-scale analysis. LLMs offer a viable solution to this challenge by automating the frame detection process, enabling more efficient and comprehensive analysis of media content.

To address this research gap, we pose three research questions:

RQ1: How accurately can a GPT model identify frames in television programs using a prompt-engineering approach?

RQ2: In comparison to human annotation, what are the limitations of a GPT-3.5 model in classifying frames of TV show transcripts?

RQ3: How can GPT models and prompt engineering be used in potential interactive applications of framing analysis?

Our research addresses these questions by making the following contributions:

1. We developed and validated a framework that leverages LLMs for automated frame detection in TV transcripts. A dataset has been created with transcripts of two current affairs programs on Dutch television. These program extracts have been classified by GPT-3.5 using a prompt composed by the definition of different frame types, the transcription item and finally we asked the model to choose the predominant frame based on the definitions.

2. We created an annotation system based on Google Forms where an expert read the transcription to be annotated as well as the frame definitions, and answered a series of questions that justified their answer. After that annotation, we analyzed, together with experts in media and frame analysis, the results obtained from the agreement/disagreement between human and machine.

3. We explored potential integration strategies to incorporate LLMs into existing media workflows, empowering journalists, media consumers, and researchers towards improved and better informed media analyses. Our research could support journalists to expeditiously verify (or identify) the framing employed in their news articles as conveyed by their text. This capability can significantly enhance accuracy and objectivity in their reporting. Additionally, our research empowers media consumers to gain a deeper understanding of the perspectives presented in TV news, by fostering critical consumption and informed decision-making. Our research represents a step forward in automating frame detection and integrating LLMs into media workflows. This has the potential for more comprehensive and insightful media analysis, ultimately enriching the media landscape for all stakeholders.

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 the results and discuss them in Section 5. Finally, we provide conclusions in Section 6.

2 RELATED WORK

As proposed by Entman, "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 definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described" [17]. This definition was the starting point for a long history of research on framing analysis in journalism and media studies [14]. There are two main approaches to framing analysis: an inductive one, in which frames emerge from the text [43], and a deductive one, in which, given already defined types of frames, they are tried to be identified in the text [16]. There is also a second subclassification. There are generic frames [38] [9] such as economic, conflict, etc., and issue-specific frames such as frames about Covid [45] or climate change [5]. At the same time, frames have been studied in different types of content, whether in the written press or on television [13].

In terms of computing research in this field [2], including attempts to automate the task, much of the previous work was done in the domain of Computational Journalism [12] [2]. Dallas et al. [9] created a dataset with articles about polemic topics such us immigration, same-sex marriage, or smoking, and they defined 15 types of frames; for each article, annotators were asked to identify any of the 15 framing dimensions present in the article, and to label spans of text which cued them, based on the definitions of each of the frame dimensions, while stating the main frame of each article. This dataset has been used in the work of Khanehzar et al. [29], where they used conventional classification techniques like Support Vector Machines (SVM) as a baseline method and showcased the enhancement in frame categorization by employing pre-trained language models like Bert, ROBERTa, and XLNet, through a fine-tuning methodology. This dataset was recently used for a task about detection of frames in SemEval-23, an International Workshop on Semantic Evaluation [36].

With the various advances in Natural Language Processing, the previous literature shows how new technologies have been adapted to the way frames are identified. This spans the use of corpus linguistics software to extract frames by Touri et al. [42]; the use of topic modeling techniques by Walter et al. [44], that describes computational methods to inductively analyse framing in texts; and the use of fine-tuned BERT based models [31]. Recently, there is a growing number of studies on the use of LLMs in Data Journalism and Media [28]. Fatemi et al. [19] explored the potential of GPT in a zero-shot setting for multi-class classification of news articles. Bianchini et al. [8] attempt to classify various dimensions of freedom (which are defined in the prompt) in several interviews with Peruvian political leaders using OpenAI models to do the transcription and then the classification of abstract concepts like freedom, which often poses challenges to computational analysis. More specific to our paper (using LLMs for framing detection), we can highlighted the works of Gilardi et al. [24] and Alizadeh et al. [3], where they classified the 15 frame types of Dallas et al. [9] in tweets and news articles, using a prompt-engineering approach that first gives the definitions of the frames and the text to classify, and then asks the model about the most predominant frame.

Our work follows a deductive approach, i.e., we use a set of frames already defined in the literature. These frames are generic and our content are transcripts of television programs. Specifically, the generic frames are those presented by Semetko et al. [38] which defines 5 types of frames: human interest, conflict, economic, morality, and attribution of responsibility. In their original work, they classified both print and TV content, but only manually. This classification system was recently used by Alonso del Barrio et al. [4], who also used LLMs, but in this case their dataset was headlines about Covid vaccination.

Our research has the potential to support journalists in their daily tasks and help the public become more critical consumers of information [25]. There are several works on the use of LLMs to create interactive tools [23] in different domains [6]. Our paper moves in the direction of supporting interactivity and the use of LLMs in journalism, not just as a text generator but as an analytical tool.

3 DATA

In this section, we describe the dataset and pre-processing steps.

3.1 Dutch TV news dataset

The dataset we used is a selection of 2000 news media items from broadcasts of the public Dutch television news programs EenVandaag (1000 items) and Nieuwsuur (1000 items).

EenVandaag [2] (OneToday) is a daily evening program broadcasted on Dutch public television channel NPO1. EenVandaag has the format of a news program with current issues and background information behind the news. The program is about 30 minutes long and deals with various news topics during an episode. The program has multiple presenters introducing various news items, and also interview experts live in the studio.

Nieuwsuur [3] (News Hour) is also an evening program and is broadcasted on NPO2, a Dutch public television channel. The broadcasts are between 30 and 45 minutes long, and also have the format of a news program with current issues and background information behind the news. This program also has multiple presenters and live interviews with experts.

We chose these two current newscasts as they provide a good overview of Dutch news items on a daily basis. These shows also provided a large corpus of items over the years, with little change in the show format. Due to this, the data is very consistent.

For analysis, the spoken words in the video recordings of these programs are transcribed. This was done with the open-source, Kaldi automatic transcriber [37]. This software can automatically transcribe Dutch spoken language into text. This pre-processing step resulted in a dataset of 2000 texts covering news between 2014 and 2018, varying in length, with an average number of 499 words for Nieuwsuur and 664 words for EenVandaag. As these texts are automatically transcribed and are thus a literal transcription of the spoken words, they contain errors at word level and also regarding sentence construction. The noisy transcriptions caused some issues for the human annotator (Section 4.1), as it involved more time to analyze some of the texts.

3.2 Translation of content

To be able to use GPT-3.5 in an optimal way, and to obtain results in which the language would not be detrimental to model performance, we translated the content into English using the Deepl API [4]. We created a script that called the API; for each entry in Dutch, it translated the text; the English translation is the text used for further analuysis. As mentioned earlier in this section, the original Dutch content contained some noise, in the sense that automatically transcribed text is not entirely clean on some occasions; this noise has been transferred to the translation. The average number of words in the English content corresponds to 496 for Nieuwsuur and 650 for EenVandaag.

4 METHODOLOGY

In this section, we explain our methodology for framing analysis, first describing the annotation process, and then summarizing the prompt-engineering technique.

4.1 Annotation

For this task, we engaged a person with university-level education and a basic knowledge in the field of framing. This was needed to make sure that the annotator was aware of the potential for human bias and was thus able to look at the texts in the most objective way possible. More specifically, the annotator had a background in language and cultural studies.

To carry out the labeling of the frames in the data, we first designed a codebook with several definitions of the frame analysis concept. This was then used to train the annotator on the task. In addition, we created an interactive environment through Google Forms to do the annotation. Using Google Apps Script [5], we created a script that allowed us to generate forms automatically.

In the annotation form, we first showed the piece of text to annotate, followed by the definitions of the 5 types of frame proposed by [38]:

- Attribution of responsibility. This frame presents an issue or problem in such a way as to attribute responsibility for its cause or solution to either the government or to an individual or group
- Human interest. This frame brings a human face or an emotional angle to the presentation of an event, issue, or problem.
- Conflict. This frame emphasizes conflict between individuals, groups, or institutions as a means of capturing audience interest.
- Morality. This frame puts the event, problem, or issue in the context of religious tenets or moral prescriptions.
- Economic. This frame reports an event, problem, or issue in terms of the consequences it will have economically on an individual, group, institution, region, or country.

After these definitions, we asked the annotator (1) to define the main frame; (2) to define an alternative frame if there was one; (3) to copy-paste sentences that helped the annotator chose the main frame; and (4) to add free text in a section called comments, in case that the annotator had something to explain. Figure 1 shows an example of a form.

We created 20 forms for each program, where each form had 50 pieces of texts to annotate. Subsequently, the answers to these forms were saved for later analysis.

4.2 Classification with GPT-3.5

The generative language model is used as follows. This type of model, given an input that we pass (called prompt) is able to generate text that continues that prompt (called output.) As a simple example, if we ask a generative model "how many sides does a triangle has?", it will generate an output through a series of tokens (a token can be a word, or a smaller unit, so a word can be formed by more than one token), and those tokens have a probability, which reflects how confident the model is of the answer, based on the text
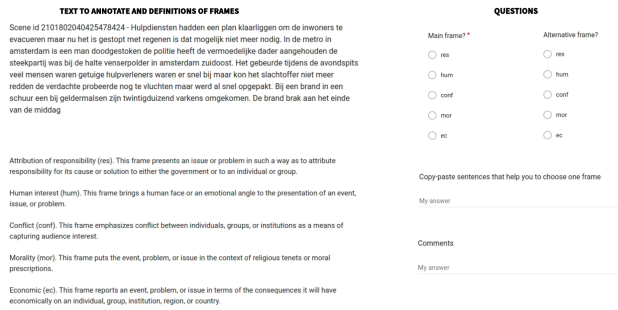
Figure 1: Example of the structure (text to annotate + definition of frames + questions) of the form to do the annotation.

it has been trained with. In this simple example, the answer would be "three" with a 100% probability of that token. Based on this idea (i.e., that the generative model produces an answer and gives that answer a probability), we defined a prompt in which we first pass the definitions of the different frame types, then we pass the text to classify, and finally we ask the model, among the 5 frame type options that we gave, which was the most likely frame. Figure 2 shows an example of the prompt used. The model gives a probability to each of the 5 frame options, so the frame with the highest probability is the predominant frame identified by the model. Furthermore, the fact of being able to access the probabilities given to the other frame types, allows us to study cases where more than one frame was possible, because the second or third options had a high probability.
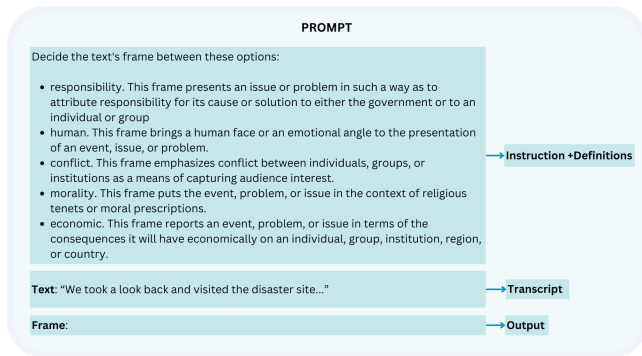

Figure 2: Example of the prompt used for frame classification, given a transcript.

To use GPT-3.5 as a frame classifier, we use the OpenAI API [6]. As we mentioned before, the model outputs the most probable token, which in this case corresponded to one of the frames, and also outputs the next most probable tokens, which correspond to other frames. In this way, we could save the probability associated with each type of frame for each text, being the frame with the highest probability the one that GPT-3.5 considered most predominant.

---
[6]https://platform.openai.com/docs/introduction/overview

GPT3 has different parameters that adjust the randomness and creativity of the answer. We set the temperature to 0, since the higher the temperature the more random the answer. Moreover, 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). As a model, we used the one with the best performance at the time of experimental design, which was TEXT-DAVINCI-003. Once we defined the parameters and the prompt to use, we made a call to the API for each of the pieces of text, obtaining the most likely frame associated with each text and the probabilities associated with each frame, for further analysis.

4.3 Understanding human-machine disagreement

Once the human annotation and machine inference were produced as described earlier in this section, we evaluated the results by quantifying the cases of agreement and disagreement between human and machine. In addition, we had access to the probabilities given by GPT3.5 to the label given by the human, which allowed us to analyze how high the probabilities were in the cases of agreement between the two, and at the same time how high the probabilities given by the machine to the human label were in cases of disagreement.

This analysis was done with the support of the annotator, as well as academic experts, through semi-structured interviews. This qualitative analysis enriches our understanding of how LLMs are used in this task.

More specifically, evaluation sessions with three framing analysis experts were held to obtain their input on the results of the research, and on the best possible ways to utilize the outcomes for future research. We held three evaluation sessions with 2 Full Professors and one Assistant Professor. We invited these three scholars due to their knowledge on framing analysis, computational methods, media and television heritage materials, and the use of audiovisual datasets. The interviews were held in June and July 2023. The interviews were semi-structured, and were centered around getting feedback on our research setup and results. To elicit responses, we first shared the project goals and the outcomes of the first phase of the research. In addition, the annotator was interviewed about her personal experiences labeling the items. The interviews with experts and annotator were not recorded, but notes were taken. Additional input was provided by email.

5 RESULTS AND DISCUSSION

5.1 Frame labeling: agreement between human and machine (RQ1)

5.1.1 EENVANDAAG. In the case of EenVandaag, the agreement between annotator and GPT-3.5 is 483 of 1000 items. This corresponds to an accuracy of 48.3%. Of those cases of agreement, 303 are human interest, 162 are conflict, 16 are economic, 1 is morality and 1 is attribution responsibility. The confusion matrix is shown in Figure 3. We see that:

- Conflict (211) is identified by the machine in 162 cases. Then is confused with human interest (37) and economic (10).

Human Interest or Conflict? Leveraging LLMs for Automated Framing Analysis in TV Shows
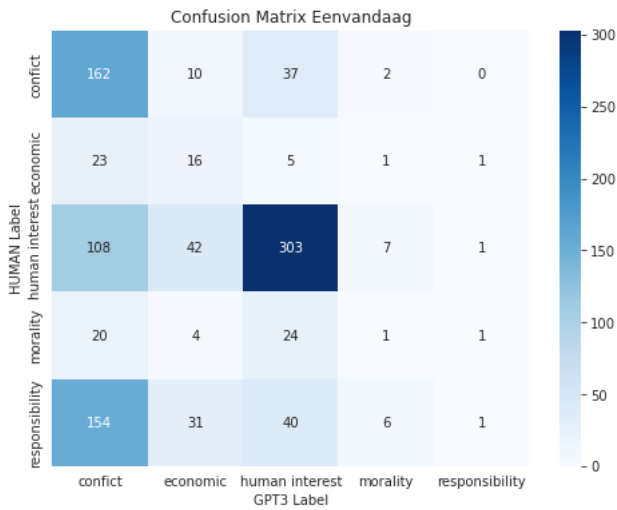
IMX '24, June 12–14, 2024, Stockholm, Sweden

Figure 3: Agreement between human annotator and GPT-3.5 on classification of EenVandaag transcripts into five categories: Conflict, Economic, Human Interest, Morality and Responsibility.

Figure 4: Agreement between human annotator and GPT-3.5 on classification of Niewwsuur transcripts into five categories: Conflict, Economic, Human Interest, Morality and Responsibility.

- Economic label (46) is well detected in 16 items, and it is confused with conflict (23), human interest (5), morality (1), and attribution of responsibility (1).
- Human interest (461) is correctly identified by the machine in 303 cases, and it is confused with conflict (108), economic (42), morality (7) and attribution of responsibility (1).
- Morality (50) is not identified at all. The machine inferred human interest (24) and conflict (20) in most cases.
- Responsibility (231) is not identified at all. The machine most commonly inferred conflict (154), human interest (40) and economic (31).

5.1.2 NIEUWSUUR. In the case of Niewwsuur, the agreement between annotator and GPT-3.5 is 387 of 1000 cases. This corresponds to an accuracy of 38.7%. Of these 387 cases, 197 are classified as human interest, 173 as conflict, and 17 as economic. Figure 4 shows the confusion matrix. We observe that:

- Conflict (256) is identified by the machine in 173 cases. Then is confused with human interest (50), economic (26), attribution of responsibility (5) and morality (2).
- Economic label (48) is identified in 17 cases, and is confused with conflict (25), human interest (6).
- Human interest (449) is identified by the machine in 197 cases. Confused with conflict (175), economic(68), and attribution of responsibility (1).
- Morality (3) is not identified at all. The machine inferred conflict in all the cases(3).
- Attribution of responsibility (244) is not identified at all. The machine most often inferred conflict(160), human interest (38), economic (44) and morality (2).
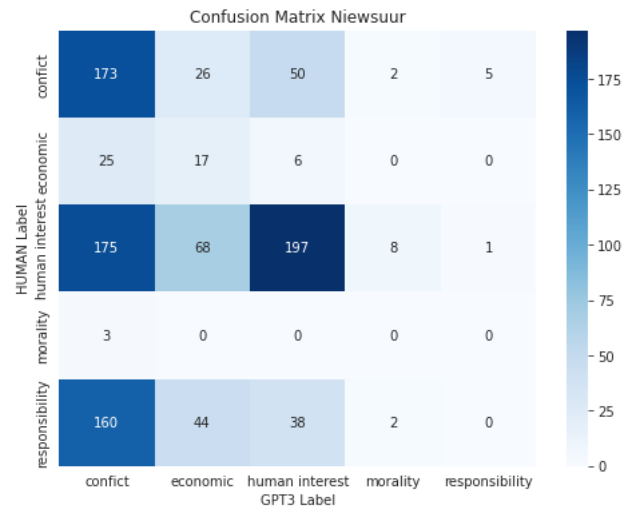
5.2 Effect of text length (RQ1)

We investigated whether text length had an influence on the cases of agreement and disagreement between human and machine. We hypothesized that longer texts might be more likely to have more than one frame, and that this may in turn lead to disagreement, while shorter texts might be more likely to have only one frame.

It is important to consider the differences in the number of articles between the two categories (human-machine agreement and human-machine disagreement) when comparing their distributions of article lengths, because the differences in sample sizes can affect the results. For example, in our case we have more data points of disagreement and the percentages of disagreement would be higher simply due to the larger number of data points, which could give a misleading impression of the data distribution

To limit this issue, we normalized the distributions by calculating relative frequencies. Instead of looking at the absolute number of articles in each length category, we have calculate the percentage of articles within each length category for both agreement and disagreement. This allows us to compare the distributions while accounting for the differences in sample sizes. Similarly, the length of the articles is not equally distributed, therefore we have divided the articles into bins based on word count. We have created bins of 100-199 words, 200-299 words, and so on, up to 800 words. This allowed as to categorize articles into different word count ranges. Within each word count range, we counted the number of cases of disagreement between human and machine and the number of cases of agreement. Later we have normalized the counts within each word count range by dividing the counts by the total number of articles in that range. This gave us the percentage of cases of disagreement and agreement within each word count range.

5.2.1 EENVANDAAG. In Figure 5a, the x-axis represents the word count ranges (e.g., 100-199, 200-299, etc.), and the y-axis represents
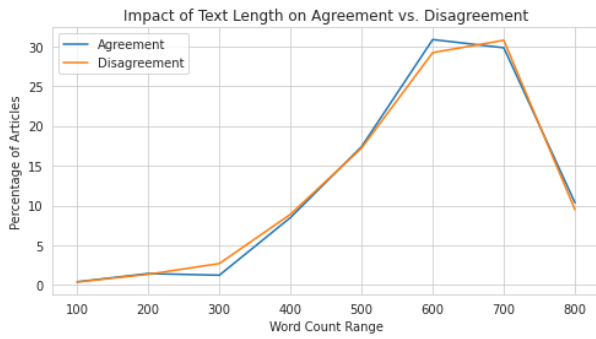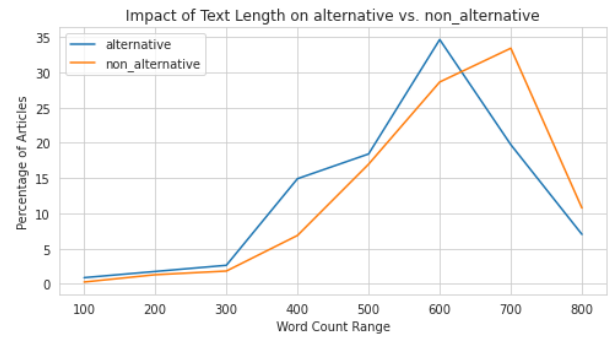
(a) EenVandaag
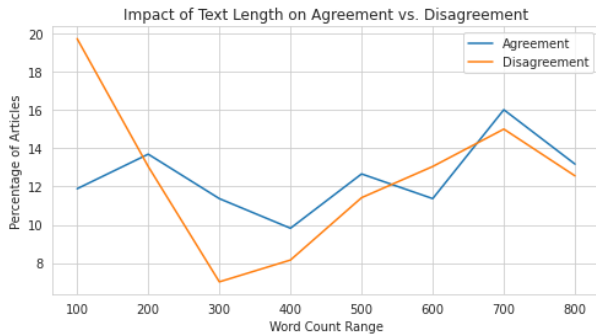
(b) Nieuwsuur

Figure 5: Relative frequencies of agreement and disagreement between human and GPT-3.5 annotations for EenVandaag/Nieuwsuur transcripts across text length bins.

(a) EenVandaag

(b) Nieuwsuur

Figure 6: Relative frequencies of alternative/non alternative frame between human and GPT-3.5 annotations for EenVandaag/Nieuwsuur transcripts across text length bins.
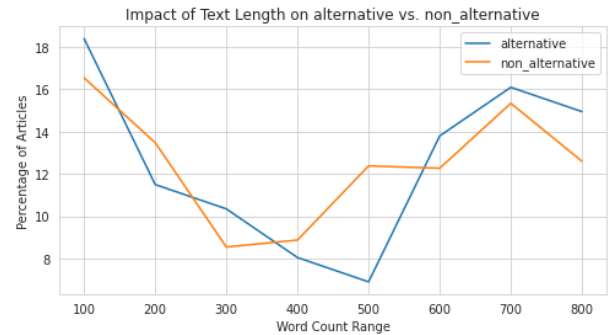
the percentage of cases of disagreement and agreement. As we can see, the figure reveals similar patterns for agreement and disagreement across text lengths. Between 100 and 700 words, there's a general increase in both agreement and disagreement with more words. However, the 800-word bin deviates from this growth trend, but in both cases agreement and disagreement are declining in a similar way (in the last bin there were not many samples). Overall, these findings suggest that text length might not be a significant factor influencing GPT-3.5's performance.

In addition, as we have cases with more than one frame (the annotator had the option to choose an alternative frame to the first choice in cases of doubt), we wanted to check if long texts are more likely to have more than one frame (alternative frame), or if, on the contrary, a longer text defines more clearly a single frame (non alternative frame).

In Figure 6a, we can observe the percentage of number of words in 2 cases, in blue the cases with an alternative frame, and in orange the cases without an alternative frame. Both lines follow a similar pattern, remaining close to each other across most word count ranges. This indicates that text length has minimal impact on the presence of alternative frames. There might be slight variations between the lines in some bins, but they are not consistent enough to suggest a significant relationship.

5.2.2 NIEUWSUUR. In Figure 5b we see a slightly more noticeable difference than in the previous case between the cases of agreement and disagreement. In the case of 100 words bin, we see 8% more of disagreement while in the case of 300 words there is around 10% more of agreement articles. For the range between 400 and 800 words we see a fairly similar distribution.

In the case of the relation between alternative frame and length of the piece of texts, as it happened in the previous case, we do not appreciate any correlation between them analyzing the results in Figure 6b, because the trends are quite similar. We only see a relevant difference between the articles with and without alternative frame of 10% in the 500 words bin.

5.3 Analyzing GPT output probabilities (RQ1)

5.3.1 EENVANDAAG. In Figure 7, we can study the distribution of probabilities given by GPT-3.5 to the human label, in the cases of agreement and disagreement.

In the case of human-machine agreement, we see that the probabilities are high, more than 50% in most cases, mean while in the case of disagreement, we see that the probabilities are very low, practically null i.e., the machine does not detect the frame chosen by the human.

5.3.2 NIEUWSUUR. Regarding the probabilities given by the machine to the human label in Figure 8, as it happened previously we
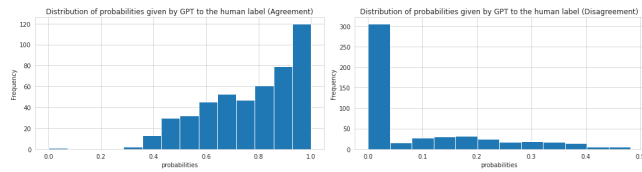
Figure 7: Distribution of the probabilities given by GPT-3.5 to the annotator label in the cases of agreement and disagreement.

observe a high probability in the cases of agreement, but really low in most of the cases of disagreement, that means that the human level is not identified at all in many cases by the GPT-3.5. In the right part of the figure we see how the probability associated with the human label in the cases of agreement is higher than 40% in most cases, while in the figure on the left we see how the probability given by the machine to the human label in the cases of disagreement is zero in most cases, so there are two distinct cases, cases where the frame is detected with a high probability, because having more than 40% probability in the choice of the main frame, having 5 possible options is quite high, but at the same time there are many cases where the frame has not been identified by the machine at all giving probabilities of zero.
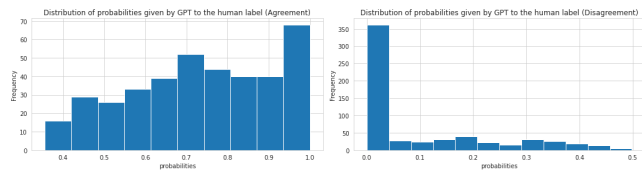

Figure 8: Distribution of the probabilities given by GPT-3.5 to the annotator label in the cases of agreement and disagreement.

5.4 Contextualizing the results: Comparison with related work (RQ1)

In this section, we put in context the results obtained, making a comparison with previous literature.

Research on identifying frames through the use of LLM is in an incipient phase, and emerging literature is gradually addressing this issue by using different types of texts to be classified, as well as different models or number of labels in the classification. As a result, we cannot make a direct comparison, since to the best of our knowledge, this is the first work where frames are automatically identified in TV news show transcripts. In any case, an indirect comparison helps to better understand the complexity of the task and the different approaches used to tackle it. Table 1 presents a summary of the main characteristics of each work.

In the work of Gilardi et al. [24] they classified the frames of tweets with a system of 15 different types of frames using ChatGPT. They work with tweets and a frame classification system of 15 different types, although in this case they use ChatGPT as a model with different parameter settings. In a dataset of tweets from 2020

and 2021, the accuracy of the model is around 40%, while a dataset of tweets from 2017-2022 is a little over 50%. Despite using another model, a system with a higher number of tags, and different type of media content, the human-machine agreement is very similar to our case.

Table 1: Summary of related work using LLMs for frame classification.

Related work	Type of data (avg.# words)	# labels	LLM	acc(%)
Gilardi et al.[24]	tweets (30)	15	ChatGPT	40-50
Alizadeh et al. [3]	tweets (30)	15	HuggingChat & FLAN	30-40
Alonso del Barrio et al.[4]	headlines (8)	6	GPT3.5	49
Ours	transcripts (600)	5	GPT3.5	43

In the work of Alizadeh et al. [3], which is related to the one mentioned above (the same data and the same frame classification system, but this time open-source models are used), the performance of open-source for that classification system is a bit worse than ChatGPT, yet in other annotation tasks the open-source version seems to work better.

In our previous work [4], we followed a similar approach, with two main differences, on one hand we used headlines from newspaper articles as text to classify, and on the other hand we used the same types of frames as we do here, adding an extra type of frame (called no-frame), since for such short sentences, there could be cases where indeed there was no frame. The results showed a 49% agreement between human and machine, which is not very far from our results, and reflects the complexity of the task as it entails significant subjectivity. In that work, we performed a post-hoc experiment where annotators were asked if they agreed or disagreed with the label provided by GPT-3.5 (without knowing the origin of this label), and they agreed in 75% of the cases. It is this type of result that illustrates the potential that LLMs can have as an annotator support tool, maintaining human-in-the-loop validation during the annotation process to ensure the reliability and accuracy of the annotations [40].

Given the discussed results, we see that our results are within an acceptable range, given that we have used much longer texts than those used in previous literature, thus making the task more difficult. Nevertheless, we still see potential and much room for improvement in the field of frame analysis with LLMs.

5.5 Expert and annotator feedback on human-machine agreeement (RQ2)

General remarks. Both the scholars and the annotator concluded that the human labeling process is time-consuming and exhausting, and thus influences the results over time. Whereas the computational labeling remains consistent over time, the human labeling does not. Because of this, the annotator stated that over time, she was more likely to choose labels that were better known to her, because those were the labels that she used most.

Due to limited time and resources, we worked with a single annotator. Although she received an extensive briefing on how the labels should be attributed, the results are nevertheless influenced by personal preferences and experiences. As all scholars pointed out, for this research to become more valuable and provide more factual information, there is a need for multiple human annotators.

This information accounts for the fact that some labels were chosen more often than others. Furthermore, in cases the annotator was uncertain about a specific label, she added an alternative label. All these cases should, in future research, be analyzed in relation to other human annotators.

Finally, the automatic transcription of the labels proved hard to read for the annotator. Some of the articles were scrambled because the spoken texts in the news item were unclear.

Experts' in-depth views. In response to the examination of the results of agreement/disagreement between human and machine, the experts discussed their hypotheses regarding the disagreement of labels. One of them suggested that "*the confusion between human interest and attribution of responsibility with the label conflict might be due to the intertwined nature of these frames*". The expert also argued that these frames are essentially building blocks of the same narrative, as a conflict often entails identifying responsibility and putting a face on an issue. The expert also emphasized that the distinction between conflict, responsibility, and human interest is based on the taken perspective, i.e., "*with conflict focusing on format, responsibility on actors in the conflict, and human interest involving one of these actors*".

Another expert provided two potential reasons for the disagreement. First, it was noted that human interest and attribution of responsibility often appear together in news stories, as "*responsibility is more sentiment-based, while human interest revolves around thematic elements. This co-occurrence could lead to confusion.* Second, the expert suggested that "*the framing cues used in the input data might not be well-defined and clear-cut enough*". The lack of precise distinctions between these frames may result in ambiguity, causing GPT-3.5 to mix them up during inference.

In summary, both experts offered insights into why GPT-3.5 might confuse human interest and attribution of responsibility with the label conflict. They highlighted the interconnected nature of these frames and the potential co-occurrence of elements in news stories as contributing factors to the disagreement. Additionally, they mentioned the importance of clear framing cues to avoid such confusion.

Annotator's in-depth views. In the case of the annotator, she suggested that the disagreement may be due to nuances in the content, emphasizing that these nuances are often quite subtle. For instance, when distinguishing responsibility from conflict, the annotator typically considered "*situations where one party is accusing another of causing a problem or demanding government intervention to address an issue. In their view, these distinctions might be too nuanced for a machine to grasp. The AI might simplify the situation and classify it simply as a conflict between two parties, even if one party is claiming responsibility.*"

Regarding human interest, the annotator explained that this frame was typically chosen when the content focused on individual cases within a broader context. She provided an example of the refugee crisis, where she would select the human interest frame when the narrative zooms in on a specific refugee, like a woman in a refugee camp, and explores her life. However, she acknowledged that "*there is a larger conflict underlying these human interest stories because people wouldn't become refugees if there were no conflicts*". The annotator suggested that "*these subtleties might not be readily detectable by the machine, leading to the misclassification of frames.*"

In summary, the annotator's views were centered on the idea that the nuances within the content might be challenging for the machine to distinguish accurately, resulting in the misclassification of frames like responsibility, conflict, and human interest.

5.6 Expanding and updating framing categories (RQ2)

We also asked the experts and the annotator if they could provide with alternatives frames they considered appropriate. The the answers were very varied, and are summarized below.

Experts' views. One of the experts pointed out to be supportive of the typology by Semetko and Valkenburg. "*Their view was that there was not a limited set of frames that journalists habitually choose from, which they could then apply to all possible topics.* While acknowledging the existence of these five frames within the studied set, the expert believed that "*the frames chosen in newsrooms were more issue-specific and played a vital role in giving meaning to social issues.*" [38].

In the expert's opinion, when thinking about a more generic frame that could apply to many issues, they suggested something like the "*politics-are-responsible frame*." This frame focused on "*how administrators, governments, and politicians are responsible for many issues, either as a cause or as a solution.* Additionally, the expert noted that many social issues were often defined in terms of "*left-wing and right-wing perspectives, which went beyond the conflict frame*". The expert saw that the conflict frame was as omnipresent as it was elusive, emphasizing that it would be more relevant to consider whether left or right forces were being examined in the process.

Furthermore, the expert identified the '*system frame*' as relevant, highlighting the existence of '*structures*' responsible for various issues. The expert viewed the '*system*' as something less tangible and partly non-human, citing examples like the invisible hand in economics. The expert believed that this frame was often used to evade individual responsibilities, such as during the banking crisis when banks referred to a systemic problem to deflect blame. The expert raised questions to explain this concept: "*w*ho controls the system?, how is it maintained?, and what do the underlying structures look like?

Additionally, the expert mentioned the concept of "*the people*" as a frame, where a vague notion of public opinion or desires was referenced without further scrutiny. The expert highlighted that it was important to recognize that "*the people are not a entity*" and that various perspectives existed within the population.

Another expert offered an alternative perspective on the frames used in the experiment, particularly the conflict, human interest, economic, morality, and attribution of responsibility frames from the Semetko typology. This expert suggested the possibility of incorporating more positive frames, such as "*reconciliation as an alternative to conflict.*" They also proposed rephrasing human interest to a more "*sentimental frame, like vulnerability, to align it better with the sentimental frames of morality and responsibility.*"

The expert noted that *"human interest," "economic," and "conflict" could be seen as overarching themes of a broadcast or news story, while "morality" and "responsibility" were more implicit or sentimental in nature.* This suggested that the first three frames could be easily

Human Interest or Conflict? Leveraging LLMs for Automated Framing Analysis in TV Shows

IMX '24, June 12–14, 2024, Stockholm, Sweden

distinguished from each other, while the last two may be more challenging to differentiate.

Annotator's views. As for the annotator, she expressed that the frame of morality was the one she used the least and the one she found most difficult to identify, perhaps because of its more religiously oriented definition. As for the economic frame, she did not use it much either, but it was an easy frame to identify. She commented that "*there were cases where the news apparently did not seem to have any frame, but was very neutral*" and perhaps an alternative frame that was "no frame" would have been useful, because sometimes she was forced to choose at least one, when, in her opinion, none of them fitted correctly in reality.

5.7 Moving forward: from current work to future applications (RQ3)

Regarding issues to be improved in LLM classification, there are several paths to explore. First, at prompt level, we wanted to start with a zero-shot learning approach; we consider it the logical initial step, since there is no need for annotated data. Providing some examples (few shot learning) could likely improve the classification of frames; at the same time, we recognize the subjectivity involved in this process, and that there will likely be cases where more than one option is correct.

Furthermore, we are aware of the lack of transparency of closed-source models, yet the fast development of these LLMs allowed us to use them and try to identify their strengths and limitations. An open-source model trained directly on Dutch would be ideal, but the reality is that large open-source models (which can be compared to OpenAI models) require computational resources that few people can access. Based on these points, and understanding the reality of the situation and the tools available to us (at the time of the experiment, GPT3.5 was the most stable model on the market, since ChatGPT was a conversational model, not an instruction-based model, and GPT4 had just been released), we decided to use GPT3.5. This paper shows both the limitations and the potential that LLMs have, as a complementary tool, to facilitate the work of media professionals in different scenarios.

This type of models could even be used to, once an article has been generated by a journalist, ask which is the frame identified by the LLM with arguments. A journalist can potentially take some advantage of the bias of the models, because these models have a bias according to the text with which they have been trained [20, 26], but audiences also have bias, so the model can reflect part of the audience. We believe it is a useful experience to do this type of exercise, and that the two options are valid: if there was agreement between human and machine, perhaps to show that the frame is very evident; if there was disagreement, it could be suggested that the article may have more than one valid frame, as long as the arguments given by the machine are convincing.

In other recent research, the work of Petridis et al. [32] is an example of an interactive tool for journalists using LLMs to explore angles for reporting a press release [35]. This work is an improvement of the performance of a previous work that uses traditional natural language processing techniques instead of LLMs [32]

Looking forward to the future use of LLMs in media, the work of Naoain et al. [34] argued that a change of mindset in the media

ecosystem is needed, and training on the use of these AI tools must be a priority given the lack of existing knowledge.

Media students need to develop an understanding of framing theory in order to critically analyze media messages. One potential direction of our work is the creation of interactive learning modules that allow students to practice identifying frames in media content or to explore the different ways that framing can influence audience perceptions. A potential application we can see is that just as annotators read a text and annotated the frame they considered giving arguments as to why they chose that frame, the frame chosen by the LLM could be shown with the machine's "arguments". In this way, there could be cases where the perception is the same but there could also be cases where it is different, and thus elicit the realisation that there are many ways of thinking and that one personal perception in many cases does not mean that it is the only possible valid one.

We also envision that the use of these tools can benefit and facilitate the work of students, and there is a lot of research being done on how LLMs can be used in education [21]. Gan et al. [22] define practical areas of the use of LLMs in Education such as instructional support tools, as assistants to teachers, providing intelligent instructional support tools and platforms, and educational assessment and feedback with learning data to provide assessment and feedback on their learning progress.

In summary, we believe that the future directions outlined here are logical implications of the work presented in this paper, and clearly need to be deepened and tested as part of future work.

6 CONCLUSIONS

In this article, we study the use of prompt-engineering for automatic frame inference in TV content. We conclude by answering the three research questions we posed:

RQ1: How accurately can a GPT model identify frames in television programs using a prompt-engineering approach? Taking advantage of text generation models for prompt-engineering classification tasks is a very promising option, as no annotated data are needed for their training phase. There is much room for improvement, both in the definition of the prompt and in the performance of such models but in any case, we see potential in the use of this tool to support journalists.

RQ2: In comparison to human annotation, what are the limitations of a GPT-3.5 model in classifying frames of TV show transcripts? We have seen that we have obtained results below 50% agreement between human and machine, but this is within the results obtained to date in previous literature. We also consider that the quality of the transcripts has influenced the understanding of both the annotator and the LLMs, as well as the possibility that, being a task involving some subjectivity, there is more than one correct answer and in future work we must have more annotators, as well as explore new types of frames that deal with the subtleties of language that can provoke human-machine disagreement.

RQ3: How can GPT models and prompt engineering be used in potential interactive applications of framing analysis? We have discussed the potential uses of this type of tools in the professional environment where there is a strong tendency to the immersion of information technology in the newsrooms, but also from the

academic point of view, where the professionals of tomorrow in the journalistic world have to adapt to technological advances.

ACKNOWLEDGMENTS

REFERENCES

[1] Toril Aalberg and James Curran. 2012. *How media inform democracy: A comparative approach.* Routledge.
[2] Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* 9335–9348.
[3] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. arXiv:2307.02179 [cs.CL]
[4] David Alonso del Barrio and Daniel Gatica-Perez. 2023. Framing the News: From Human Perception to Large Language Model Inferences. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (Thessaloniki, Greece) *(ICMR '23).* Association for Computing Machinery, New York, NY, USA, 627–635. https://doi.org/10.1145/3591106.3592278
[5] Nic Badullovich, Will J Grant, and Rebecca M Colvin. 2020. Framing climate change for effective communication: a systematic map. *Environmental research letters* 15, 12 (2020), 123002.
[6] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).
[7] Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, et al. 2021. OpenFraming: open-sourced tool for computational framing analysis of multilingual data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* 242–250.
[8] Gonzalo Espinoza Bianchini, Lisa Zanotti, and Carlos Meléndez. 2023. Using OpenAI models as a new tool for text analysis in political leaders' unstructured discourse. (2023).
[9] 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
[10] Marcelo Carvalho Afonso, Pedro Almeida, Pedro Beça, Telmo Silva, and Iulia Covalenco. 2022. Usability Of Text-To-Speech Technology in Creating News Podcasts using Portuguese Of Portugal. In *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences* (Aveiro, JB, Portugal) *(IMX '22).* Association for Computing Machinery, New York, NY, USA, 363–368. https://doi.org/10.1145/3505284.3532968
[11] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924* (2023).
[12] Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational journalism. *Commun. ACM* 54, 10 (oct 2011), 66–71. https://doi.org/10.1145/2001269.2001288
[13] Claes H. de Vreese. 2004. The Effects of Frames in Political Television News on Issue Interpretation and Frame Salience. *Journalism & Mass Communication Quarterly* 81, 1 (2004), 36–52. https://doi.org/10.1177/107769900408100104 arXiv:https://doi.org/10.1177/107769900408100104
[14] Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal+ document design* 13, 1 (2005), 51–62.
[15] Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media.* Harvard University Press.
[16] 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.
[17] Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory* 390 (1993), 397.
[18] Robert M Entman, Jörg Matthes, and Lynn Pellicano. 2009. Nature, sources, and effects of news framing. *The handbook of journalism studies* (2009), 175–190.
[19] Bahareh Fatemi, Fazle Rabbi, and Andreas L. Opdahl. 2023. Evaluating the Effectiveness of GPT Large Language Model for News Classification in the IPTC

[20] News Ontology. *IEEE Access* 11 (2023), 145386–145394. https://doi.org/10.1109/ACCESS.2023.3345414
[20] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *arXiv preprint arXiv:2305.08283* (2023).
[21] Fernando Filgueiras. 2023. Artificial intelligence and education governance. *Education, Citizenship and Social Justice* (2023), 17461979231160674.
[22] Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData).* IEEE, 4776–4785.
[23] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (, Virtual Event, Australia,) *(DIS '22).* Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533
[24] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (July 2023). https://doi.org/10.1073/pnas.2305016120
[25] Katerina Gorkovenko and Nick Taylor. 2019. Audience and Expert Perspectives on Second Screen Engagement with Political Debates. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video* (Salford (Manchester), United Kingdom) *(TVX '19).* Association for Computing Machinery, New York, NY, USA, 70–82. https://doi.org/10.1145/3317697.3323352
[26] Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. AI and the transformation of social science research. *Science* 380, 6650 (2023), 1108–1109.
[27] Muhammad Usman Hadi, qasem al tashi, Rizwan Qureshi, Abbas Shah, amgad muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. 2023. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. (Nov. 2023). https://doi.org/10.36227/techrxiv.23589741.v4
[28] Bahareh R. Heravi. 2019. 3Ws of Data Journalism Education. *Journalism Practice* 13, 3 (2019), 349–366. https://doi.org/10.1080/17512786.2018.1463167 arXiv:https://doi.org/10.1080/17512786.2018.1463167
[29] Shima Khanehzar, Andrew Turpin, and Gosia Mikołajczak. 2019. Modeling Political Framing Across Policy Issues and Contexts. In *ALTA.*
[30] Soomin Kim, JongHwan Oh, and Joonhwan Lee. 2016. Automated News Generation for TV Program Ratings. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (Chicago, Illinois, USA) *(TVX '16).* Association for Computing Machinery, New York, NY, USA, 141–145. https://doi.org/10.1145/2932206.2933561
[31] 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).*
[32] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the news: Digital creativity support for journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–11.
[33] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research* 25, 3 (2023), 277–304. https://doi.org/10.1080/15228053.2023.2233814 arXiv:https://doi.org/10.1080/15228053.2023.2233814
[34] Amaya Sánchez Naoaín. 2022. Addressing the Impact of Artificial Intelligence on Journalism: The perception of experts, journalists and academics. (2022).
[35] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–16.
[36] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023).* 2343–2361.
[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Hilton Waikoloa Village, Big Island, Hawaii, US). IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
[38] 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
[39] Than Htut Soe, Frode Guribye, and Marija Slavkovik. 2021. Evaluating AI assisted subtitling. In *Proceedings of the 2021 ACM International Conference on Interactive*

Media Experiences (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 96–107. https://doi.org/10.1145/3452918.3458792

[40] Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can ChatGPT-like LLMs effectively replace human annotators in NLP tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

[41] Sina Thäsler-Kordonouri and Kurt Barling. 2023. Automated Journalism in UK Local Newsrooms: Attitudes, Integration, Impact. *Journalism Practice* 0, 0 (2023), 1–18. https://doi.org/10.1080/17512786.2023.2184413 arXiv:https://doi.org/10.1080/17512786.2023.2184413

[42] Maria Touri and Nelya Koteyko. 2015. Using corpus linguistic software in the extraction of news frames: towards a dynamic process of frame analysis in journalistic texts. *International Journal of Social Research Methodology* 18, 6 (2015), 601–616.

[43] Baldwin Van Gorp and Tom Vercruysse. 2012. Frames and counter-frames giving meaning to dementia: A framing analysis of media content. *Social science & medicine* 74, 8 (2012), 1274–1281.

[44] Dror Walter and Yotam Ophir. 2019. News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures* 13, 4 (2019), 248–266.

[45] Philipp Wicke and Marianna M Bolognesi. 2020. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS one* 15, 9 (2020), e0240010.

[46] Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering LLM-based Machine Translation with Cultural Awareness. *arXiv preprint arXiv:2305.14328* (2023).

[47] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848* (2023).