Capturing the Viewpoint Dynamics in the News Domain

Enrico Motta[1,2] [0000-0003-0015-1952], Francesco Osborne[1,3] [0000-0001-6557-3131], Martino M. L. Pulici[1] [0009-0009-4533-7495], Angelo Salatino[1] [0000-0002-4763-3943] and Iman Naja[1] [0000-0001-6634-3266]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
[2] MediaFutures Centre, University of Bergen, Bergen, Norway
[3] Department of Business and Law, University of Milano Bicocca, Milan, IT
enrico.motta@open.ac.uk

Abstract. Despite the seismic changes brought about by the web and social media, mainstream news sources still play a crucial role in democratic societies. In particular, a healthy democracy requires a balanced and diverse media landscape, able to provide an arena in which the various *topics* and *viewpoints* relevant to the political discourse of the day are presented and discussed. Unfortunately, there is currently little effective computational support available to the various classes of users, who are interested in monitoring the topic and viewpoint dynamics in the news — e.g., for regulatory or research purposes. As a result, current analyses by researchers and practitioners tend to be small scale and, by and large, rely on manual investigations of topic and viewpoint coverage. To address this issue, we have developed a hybrid human-machine approach, which uses a Large Language Model (LLM) first to help analysts to identify the range of viewpoints relevant to the debate around a given topic, and then to classify the claims expressed in the news corpus of interest with respect to the identified viewpoints. We tested a variety of LLMs on a benchmark corpus of news items drawn from British media sources. Our results indicate that GPT4o outperforms the other alternatives and can already provide effective support for this classification task, even when run in a zero-shot learning modality.

Keywords: News Analytics, News Classification, Large Language Models, Viewpoints.

1 Introduction

Despite the seismic changes brought about by the web and social media, mainstream news sources still play a crucial role in democratic societies, informing the public about the important issues of the day and providing a platform for democratic debate. In this context, assuming an appropriate regulatory regime is in place to guarantee a diverse media ownership, the acid test for a democratic media landscape is related to *viewpoint diversity* [1], namely the extent by which media sources provide citizens with a robust range of alternative interpretations on a given topic. Here we use the term 'topic' to refer to a particular issue being discussed in the media — e.g., "the appropriate level of income tax in the UK", while a viewpoint refers to a position relevant to this issue —

e.g., a position advocating a low taxation regime. In particular, as explained later in the paper, the concept of 'viewpoint' does not indicate an individual statement reported in the news but defines a category that groups together all statements that subscribe to a common position — e.g., all the statements in favour of a low taxation regime.

Assessing viewpoint diversity accurately requires effective content analysis methods and tools, which constitutes a major challenge, given i) the sheer scale of news and information in the contemporary online environment[1] — an information overload that facilitates disinformation [2], and ii) the limited take-up of automated content analysis solutions in media research and related fields [3]. Arguably, this limited take-up reflects a disconnect between the needs of media scholars and the state of the art in computational techniques for news media analysis [4]. As a result, current studies by both researchers and practitioners — e.g., see [5] — tend to be small scale and, by and large, rely on a manual analysis of topic and viewpoint coverage.

To address this issue, we have developed a hybrid human-machine approach, which relies on a Large Language Model (LLM) to help analysts to identify the various viewpoints relevant to the debate around a given topic, and then classifies the claims expressed in a news corpus of interest with respect to the identified categories. For instance, if we take immigration as our topic of interest, through an analysis of a corpus of news items we can identify a number of relevant viewpoints, covering different perspectives on the debate. These may include the view of immigration as beneficial to the economy, the characterization of immigrants as a threat to a nation's culture and social fabric, a humanitarian perspective on refugees as vulnerable people, and several others. Once these viewpoints have been identified, we can then use a LLM to classify statements expressed in the media by political and other actors with respect to these viewpoints, thus generating a multi-dimensional representation of the debate on the immigration topic over time. Here, we use the term 'claim' to refer to a statement about a topic, which has been expressed by an actor and reported in a news item. Our approach has been evaluated on a corpus of news items drawn from British media sources, showing promising results. In particular, as discussed in Section 3, our evaluation shows that GPT4o outperforms the other alternatives and can provide good support for this classification task, even when run in a zero-shot learning modality.

The rest of the paper is structured as follows. In the next section we discuss the notion of viewpoint and provide a brief overview of computational approaches to viewpoint extraction. In Section 3 we present our approach, illustrating our methodology in a concrete scenario, where we model the viewpoint dynamics associated with the UK's immigration debate, as discussed in mainstream news media over a three month period. In Section 4 we briefly discuss the application-level perspective on this work, providing an example of the kind of user-centric visualizations that we can generate to provide insights to media scholars and practitioners. Finally, in Section 5 we reflect on the key results presented in this paper and discuss the outstanding issues that will be the focus of future research.

[1] Even if we restrict our analysis to UK's mainstream news outlets, we already ought to consider thousands of news items being published on a daily basis – see https://tinyurl.com/ycxwprjm.

2 Viewpoints in the media and computer science literature

2.1 Viewpoints as collections of coherent claims

The notions of viewpoint and viewpoint diversity in media sciences are discussed by Baden and Springer [1], who emphasise that differences in the ways arguments are *framed* [6] do not necessarily indicate different viewpoints. In particular, they point out that a specific position on an issue can be articulated through different framing devices — e.g., by presenting it in abstract terms through a thematic frame or alternatively by illustrating concrete examples, therefore making use of an episodic frame. According to Baden and Springer, in order to be distinct, viewpoints need instead to "construct different meaning in a consequential sense" — i.e., they need to open up perspectives that are semantically diverse. An important consequence of this approach is that viewpoints are therefore not atomic concepts but are instead constructed out of a number of individual positions expressed in the media. While these positions may be articulated using different terms and framing devices, they can be grouped around "commensurable interpretations" [1] — i.e., they subscribe to the same viewpoint. Building on these ideas, in a previous paper [7] we formalised the notion of viewpoint as a coherent collection of *claims* — i.e., one that only includes claims that satisfy the membership criterion associated with the viewpoint in question. In turn, a claim is defined in our model as a statement reported in a news item, which is associated to an agent and concerns a topic. This formalization has been realised in a formal OWL model, the News Classification Ontology (NCO), which can be found at http://data.open.ac.uk/ontology/news-classification.

An example of instantiating the notion of viewpoint as a device for bringing together commensurable claims can be found in [5], where the authors analyse the immigration debate in mainstream media and identify four different viewpoints (*"Negative", "Administrative burden", "Positive", "Victimization"*), which abstract from the variety of individual positions about immigration that are reported in the news.

2.2 Computational approaches to viewpoint extraction

The survey by Doan and Gulla [8] reviews the state of the art concerning automated approaches to identifying political viewpoints and concludes that it "falls somewhat short of our goal with automatic political viewpoint identification". In their survey, Doan and Gulla adopt a definition of political viewpoint that is broadly consistent with the one used in this paper. Essentially, they emphasise that, for a given topic, T, it is possible to identify a set of viewpoints, {p1, ….pn}, which summarise the contrastive positions that can be expressed about T[2]. The task of a system for political viewpoint identification is then to classify a piece of text (what we refer to as a claim in our terminology) in terms of the given set of viewpoints.

The approach by Trabelsi and Zaïane [9] proposes a generative Topic-Viewpoint model, which uses unigrams to characterise probabilistically the vocabulary that is used

[2] Actually, the definition given in [8] does not explicitly includes the notion of topic, however we assume that this is simply an oversight, given that the notion of topic is informally discussed elsewhere in the paper and implicitly assumed throughout this work.

by an author expressing a particular viewpoint about a topic. A particular strength of this approach is that it is unsupervised and is generic with respect to specific topics or viewpoints. However, from the examples shown in the paper, it appears that the approach is limited to a binary classification, in favour or against a particular position related to a topic. This limits the approach's flexibility, in particular its ability to identify the variety of viewpoints relevant to the debate about a topic and then to classify relevant statements in terms of these various viewpoints. Another key aspect of this approach is that it takes advantage of interactions between different post creators on social media to identify contrastive opinions. While this is of course an interesting and clever heuristic in the context of social media, it also means that this approach is not directly applicable to our news scenario.

The work by Quraishi et al. [10] also focuses on debates in social media to identify contrastive viewpoints on a topic. Like Trabelsi and Zaïane, they take advantage of interactions in social media to identify different communities that in turn can be associated with different viewpoints. This work uses a graph partitioning method and improves on the approach by Trabelsi and Zaïane because it is able to discover several viewpoints associated with a topic. They also focus on explaining the discovered viewpoints by means of Iterative Rank Difference, a technique that identifies the descriptive keywords associated with a viewpoint. However, judging from the examples shown in the paper, the resulting descriptions appear to be both opaque and rather noisy. Moreover, in contrast with the aim of our work, there is no attempt here to try and provide a comprehensive account of the debate around a topic by identifying all the relevant viewpoints and using the resulting classification to structure all relevant claims.

The paper by Hada et al. [11] focuses instead on measuring the viewpoint diversity which emerges from interactions on X (formerly known as Twitter). In particular, they use the Fragmentation metric [12], which aims to measure the degree of exposure to alternative viewpoints for each user. Their analysis of conversations about the immigration topic shows that nearly 70% of users have a very low Fragmentation score, indicating virtually no exposure to alternative viewpoints. However, in contrast with our work, there is no attempt in this paper to model the various relevant viewpoints and the focus is purely on measuring diversity. In addition, as is the case with research on stance detection [13], the various positions around a topic are clustered around two broad categories, in favour or against a particular position, in contrast with the more fine-grained differentiation proposed in this paper.

Finally, Chen et al. [14] test different solutions on a variety of tasks concerned with extracting and classifying perspectives. A combination of Information Retrieval and a BERT model obtains a 50.8% F1 measure in perspective extraction, while BERT alone obtains a 63.7% F1 measure on the task of clustering equivalent perspectives — in our terminology, positions that belong to same viewpoint. This work is relevant to our research, as it focuses on both identifying and grouping together individual positions about a topic. However, they extract their data from debating websites, such as idebate.com, which naturally provide more structure to the debate than what is available from the broader news domain, which is instead the focus of our research. In addition, in contrast with our methodology, there is no attempt at classifying individual claims with respect to the set of identified viewpoints. Instead, the authors focus on classifying

the stance of each perspective, which relies on a rather reductive assumption that any perspective on a topic can be classified in terms of binary support/oppose stance, which is not necessarily the case in the context of analysing complex issues[3].

3 Approach

3.1 Generating a corpus of news items covering the UK immigration debate

As already pointed out, our goal is to develop an approach that can identify the range of viewpoints relevant to a topic and then automatically classify all the claims in a news corpus with respect to the identified viewpoints. The starting point for our experiments was a news corpus comprising 603 news items, published between 1 June and 31 August 2023. The corpus was generated by accessing the Aylien news service (www.aylien.com), which provides an API to access information from about 80,000 news sources from around the world. In particular, we retrieved news items that had already been tagged with the keyword 'immigration' by the Aylien service and/or contained such a keyword (or related ones, such as 'immigrants') in the title or body. The search was restricted to the following mainstream news sources: The Guardian, BBC, ITV, Daily Mail, The Sun, The Daily Express, The Daily Telegraph, The Evening Standard, The Independent, Reuters and Associated Press.

Having done this, the next step was to extract the claims reported in the news corpus. A claim is formally defined as a sextuple *<utterance, actor, news item, news source, date, topic>*, where the utterance is a statement made by an actor, about a particular topic, which is reported in a news item on a certain date[4]. The field 'news source' indicates the outlet that published the news item in question and, in our case study, can only be one of the mainstream media sources listed above. In the experiments reported in this paper the field 'topic' is fixed, given that we are only concerned with statements made about the topic 'Immigration'. Finally, we should also point out that we only consider claims that consist of utterances that can be attributed to an actor explicitly mentioned in the associated news item. Such an association between utterance and actor may be expressed directly in the news item, by reporting a statement by the actor using quotation marks, or indirectly — e.g., by using an expression, such as, "[actor] said that…"). However, while both direct and indirect quotations are considered, in either case the actor must be explicitly mentioned in the news item. This means that opinions expressed implicitly — e.g., an opinion expressed by a journalist authoring a news item — are considered as out of scope for this study.

[3] For example, our analysis of the immigration debate in UK includes a viewpoint "Immigration as a Management Issue", which cannot be reduced to a stance pro or against immigration.

[4] Needless to say, a specific utterance expressed by a specific actor may be reported by multiple news sources, often on the same day, leading to a number of distinct claims that share a common utterance and actor. Analogously, it is also possible for multiple actors to repeat the same utterance. Indeed, this is not uncommon in the UK's political debate, where politicians belonging to the same party may be required to put forward a specific agreed line.

3.2 Extracting claims

To extract the set of claims we run GPT-4[5] on the news corpus, generating 3455 claims. However, once we analysed the extracted claims, we realised that they covered a variety of immigration contexts, not just in UK but also in other countries. Hence, we decided to create a more coherent corpus by restricting the analysis to UK immigration. This was achieved by considering only claims made by actors based in Britain[6] and, as a result, we ended up with 766 claims. These included both 'atomic' utterances, such as *"Britain has returned 1,800 migrants to Albania in just six months",* as well as more articulated positions, such as *"While the principle of reducing net migration was right, there was a shortage of care workers in the UK. It's not as simple as just putting the salary thresholds up as well, there's quite a lot of skilled but lower paid people that we need coming into this country".*

3.3 Identifying the range of viewpoints

A hybrid human-machine approach was used to generate a comprehensive set of viewpoints, using GPT-4 Turbo to produce an initial list that was then finalised by a human expert. In particular, we fed the utterances to the LLM in six batches and the model was asked to produce 5 viewpoints that could be used to classify the utterances. This produced a total of 30 viewpoints. Next, the LLM was asked to extract and synthesise the most significant and frequent viewpoints from all the ones generated by the various runs. The output was an initial set of viewpoints that was then checked by a human expert, to ensure that this set could provide a comprehensive range of dimensions to cover the UK immigration debate. In particular, the key role of the human expert was to provide clear distinct definitions for the various viewpoints and ensure complete coverage of the relevant perspectives. To this purpose, he also added an additional viewpoint, 'Immigration as a management issue', which had not been picked up by the LLM. This human-in-the-loop element of the methodology is in our view essential, given that our goal is to provide support to the various researchers and practitioners who engage in media analytics for regulatory or research purposes. Hence, it is crucial that the set of viewpoints used for classifying a debate is robust and consistent with the type of categories that expert analysts, such as media and political scientists, would be happy to consider.

The set of identified viewpoints is as follows:

1. Immigration as a management issue. This viewpoint characterises utterances that focus on the way immigration is managed, typically by the UK government. For example, criticisms of specific elements of immigration policy — e.g., the use of hotels to house immigrants — should be classified under this viewpoint, unless other factors, such as humanitarian considerations, are emphasised in the claim. A key aspect of this viewpoint is that it does not necessarily imply a stand in favour or against immigration.

[5] In the early phases of this work we used GPT-4 (for claim extraction) and GPT-4 Turbo (for viewpoint identification), while we later switched to GPT-4o, once this became available.

[6] These were identified by querying Wikidata [15] and, in the vast majority of cases, turned out to be British politicians.

2. Immigrants as victims/Humanitarian emphasis. This viewpoint is used to classify utterances that are sympathetic to the plight of immigrants — e.g., when a tragedy happens at sea.

3. Immigrants as potential criminals or threat/National security emphasis. This viewpoint classifies utterances that imply a view of migrants as criminals or the migration phenomenon as a threat to national security. This viewpoint also covers the rhetoric about "dodgy lawyers" — i.e., lawyers who instruct their refugee clients to lie in order to get asylum in UK. In this case immigrants are criminals by association. In addition, utterances that advocate the use of restraining measures that are normally used for criminals — e.g., security tags, also fall under this category.

4. Enhancing/Maintaining immigration pathways. This viewpoint is used to classify utterances which either advocate for measures that would make it easier to come to UK or alternatively criticise the introduction of new restrictions to immigrations. Interestingly, in the context of an immigration debate, statements that criticise a relaxation of immigration rules should not be classified under 'Maintaining immigration pathways'. In other words, 'Maintaining immigration pathways' is not a neutral category, but implies a (mild) positive attitude towards immigration. In addition, this viewpoint is not necessarily mutually exclusive with the 'Restricting immigration pathways' one, because an utterance may advocate for more legal migrants coming to UK while supporting stricter measures against illegal migrants.

5. Restricting immigration pathways. This viewpoint is used to characterise utterances that refer to measures that would make it more difficult to come to UK. It covers both legal and illegal immigration pathways. Furthermore, attempts to remove migrants from the UK and 'success stories' about sending migrants back to their country also fall under this category.

6. Economic benefits of immigration. This viewpoint is used to classify utterances that refer to the economic value of immigration. Note that this viewpoint is not mutually exclusive with the one labelled 'Economic cost of immigration', because an utterance may consider certain migrants as economically beneficial while maintaining that others introduce a financial burden for the country.

7. Economic cost of immigration. This viewpoint is used to classify utterances that refer to the economic cost of immigration — e.g., when talking about the cost of accommodation for illegal migrants.

8. Integration policies/Multiculturalism as a positive force. This viewpoint is used to classify utterances that propose practical measures for integrating migrants in UK society, emphasise the need to support the integration of migrants, or otherwise highlight the value of cultural diversity. This viewpoint is not mutually exclusive with the one labelled 'Anti-integration policies/Cultural identity preservation', because an utterance may express a favorable opinion about multiculturalism while advocating placing tracking tags on illegal migrants.

9. Anti-integration policies/Cultural identity preservation. This viewpoint is used to classify utterances that emphasise the cultural differences between UK people and foreign migrants as well statements that advocate separating migrants from

the rest of the UK population. For instance, the use of tracking tags on immigrants implies both a view of immigrants as criminals and also enforces an anti-integration policy.

3.4 Generating a gold standard for viewpoint classification

In order to construct a gold standard on which to benchmark a variety of LLMs, 402 claims were randomly selected from the corpus of claims relevant to the UK immigration debate and five human annotators were given the task to classify each of them in terms of the nine relevant viewpoints — as already mentioned, an individual claim can indeed instantiate more than one viewpoint. Following an initial standardization phase on a small subset of claims, which were classified by all five annotators, each claim was rated by exactly three annotators and we then used majority voting to generate the gold standard. To facilitate the annotation task, a customised spreadsheet was provided to each annotator, which, among other things, made it easy for them to quickly interpret and classify an utterance in the context of the associated news item — i.e., they could access a customised rendering of the news item in which the relevant utterance had been highlighted. To produce a baseline gold standard, comprising the entire corpus of 402 claims, a simple majority rule was used, where an utterance, say u_i, would be classified (or not) under a viewpoint, say v_j, if and only if at least two annotators agreed that u_i should go under v_j (or not). However, as shown in the second column of Table 1, only moderate agreement was achieved on average between the human annotators on the corpus of claims[7]. The reason for this is that the classification of political statement is a rather contested task, even for humans. Statements by politicians can be ambiguous and difficult to interpret, and therefore, despite putting significant effort in calibrating and harmonizing the scores from different annotators, only a moderate level of agreement could be achieved. For this reason, we also produced a restricted version of the gold standard, by only keeping claims for which, for a given viewpoint, say v_i, either at least two annotators agreed that v_i was relevant to the claim in question or alternatively all annotators agreed that it was not relevant. In other words, all utterances for which one and only one annotator flagged any viewpoint as relevant were discarded. This restricted version provided us with a more robust basis for evaluating the performance of different LLMs on the claim classification task.

Our analysis also showed that viewpoint 8 (Integration policies/Multiculturalism as a positive force) was particularly problematic, with its opposite, viewpoint 9 (Anti-integration policies/Cultural identity preservation) also exhibiting a low agreement score. Therefore we also produced agreement scores with only seven viewpoints, removing viewpoints 8 and 9. As shown in Table 1 and Table 2, if limit ourselves to only seven viewpoints, the level of agreement increases and in particular we reach a substantial level of agreement on the restricted dataset. Both datasets are freely available and can be accessed at https://doi.org/10.21954/ou.rd.26268025.

[7] Jacob Cohen himself has suggested that a score of 0.41 (moderate agreement) may be acceptable. However this position has been criticised [16] and in general a score denoting at least substantial agreement is expected (≥ 0.61), with almost perfect agreement (≥ 0.81) the recommended norm for critical domains, such as medical studies – see [16] for a discussion on this issue.

Table 1. Agreement between annotators on the different datasets.

Pair	Cohen's kappa on full dataset (402 claims) (9 viewpoints)	Cohen's kappa on full dataset (402 claims) (7 viewpoints)	Cohen's kappa on restricted dataset (219 claims) (7 viewpoints)
1–2	0.46	0.49	0.70
1–3	0.56	0.66	0.87
1–4	0.38	0.39	0.69
1–5	0.49	0.55	0.80
2–3	0.33	0.36	0.58
2–4	0.43	0.49	0.68
2–5	0.44	0.54	0.84
3–4	0.30	0.35	0.60
3–5	0.56	0.60	0.73
4–5	0.37	0.35	0.68
Average	***0.43***	***0.48***	***0.71***

Table 2. Average annotator agreement by viewpoint on the different datasets.

Viewpoint	Cohen's kappa on full dataset (402 claims)	Cohen's kappa on restricted dataset (219 claims) (7 viewpoints)
1	0.55	0.80
2	0.67	0.82
3	0.35	0.51
4	0.36	0.69
5	0.47	0.72
6	0.43	0.81
7	0.52	0.69
8	0.11	N/A
9	0.35	N/A

3.5 Using LLMs to classify claims in terms of the relevant viewpoints

Having produced both a baseline and a restricted gold standard, we then tested a variety of LLMs on both datasets, to assess to what extent they can effectively support the task of classifying claims with respect to the relevant viewpoints. All the LLMs were tested in a zero-shot learning modality. This approach was chosen to establish an initial baseline for future developments and also because zero-shot learning is well-suited for supporting discourse analysis across any domain, a feature that is very important for our

target users. Since most of the LLMs do not provide a 'clean' output — i.e., a yes/no binary classification result, regex patterns were employed to post-process the verbose output. For each of the experimental setups, standard binary classification metrics were computed for every model, including those of a random classifier, which provides a baseline reference for the performance of the other models. As is the norm for classification tasks, the performance of the LLMs was assessed in terms of precision, recall, and F1 scores.

In particular, we evaluated both open-source models (Llama 2, Llama 3, Mistral 8x7B) and commercial closed models accessible via API (GPT-4, GPT-4o, Titan Text Premier, Mistral Large). While the latter typically perform better in zero-shot settings, usually they do not disclose the number of parameters and other key characteristics, behaving as black boxes. Below, we briefly summarise the published characteristics of these models.

GPT-4 Turbo, developed by OpenAI, is a large commercial multimodal model featuring a context window of 128K tokens [17]. Originally, this was OpenAI's flagship model, however it has now been surpassed by GPT-4o. The training data extends up to December 2023.

GPT-4o is OpenAI's most advanced model[8], offering text generation that is twice as fast and 50% cheaper than GPT-4. It shares the same 128K-token context window as GPT-4 Turbo and uses training data up to October 2023.

Titan Text Premier is the latest addition to Amazon's Titan family of LLMs. It is designed for enterprise-grade text generation applications and was optimised for retrieval-augmented generation[9]. It has a context length of 32K tokens.

Mistral Large is Mistral's flagship language model [18], featuring a 32K-token context window. As with all the previous commercial models, the number of parameters has not been disclosed.

Mistral 8x7B is an open sparse mixture-of-experts network, consisting of 8 models, each with 7 billion parameters [18]. It uses a context length of 32K tokens. Specifically, we used the Mistral 8x7B Instruct, which has been fine-tuned through supervised learning and direct preference optimization for precise instruction following. It is regarded as one of the strongest open-weight models.

Llama 2 70B is the largest member of Meta's Llama 2 family, featuring a 4K-token context length [19]. For this study, we adopted Llama 2 70B Chat, a fine-tuned version of Llama 2 70B optimised for dialogue use cases. All Llama 2 models employ supervised fine-tuning and reinforcement learning with human feedback.

Llama 3 70B is the largest model in Meta's recently released Llama 3 family [20]. It offers a context window of 8.2K tokens and utilises Grouped-Query Attention to improve inference efficiency. It includes training data up to December 2023.

[8] https://openai.com/index/hello-gpt-4o/.

[9] https://aws.amazon.com/it/about-aws/whats-new/2024/05/amazon-titan-text-premier-amazon-bedrock/.

Table 3. Scores for each model on the full dataset (9 viewpoints).

Model	Precision	Recall	F1
Random	0.11	0.50	0.15
GPT-4 Turbo	0.36	0.69	0.46
GPT-4o	**0.49**	0.65	**0.52**
Llama 2 70B	0.20	0.66	0.30
Llama 3 70B	0.36	0.71	0.45
Titan Premier	0.42	0.48	0.43
Mixtral 8x7B	0.36	0.63	0.40
Mistral Large	0.36	**0.75**	0.46

Table 4. Scores for each model on the full dataset (7 viewpoints).

Model	Precision	Recall	F1
Random	0.12	0.50	0.17
GPT-4 Turbo	0.40	0.75	0.51
GPT-4o	**0.53**	0.80	**0.62**
Llama 2 70B	0.23	0.69	0.34
Llama 3 70B	0.41	0.84	0.52
Titan Premier	0.50	0.61	0.53
Mixtral 8x7B	0.39	0.75	0.48
Mistral Large	0.38	**0.86**	0.51

Table 5. Scores for each model on the restricted dataset (7 viewpoints).

Model	Precision	Recall	F1
Random	0.15	0.50	0.20
GPT-4 Turbo	0.51	0.75	0.60
GPT-4o	**0.71**	0.82	**0.73**
Llama 2 70B	0.27	0.65	0.37
Llama 3 70B	0.49	0.83	0.59
Titan Premier	0.64	0.58	0.58
Mixtral 8x7B	0.49	0.75	0.55
Mistral Large	0.47	**0.85**	0.58

As shown in Table 3, Table 4 and Table 5, GPT-4o outperforms all other models in all three test configurations. As expected, the performance of all LLMs improves monotonically in the three configurations, as we remove the most problematic viewpoints and claims. However, it is important to emphasise that GPT-4o exhibits a decent per-

formance even in the most challenging scenario (Table 3), with pretty good performance in both datasets once viewpoints 8 and 9 have been removed. In particular, its performance on the restricted dataset (Table 5) is arguably good enough to provide useful insights in a media analytics application scenario. In addition, regardless of the model in question, it can also be seen that all LLMs significantly outperform the random classifier. Given the strong dataset imbalance, this observation is crucial to paint a clearer picture of the performance of the models, whose ability may be underestimated without comparing it to a baseline. In Table 6 and Table 7 we provide more details about the performance of the best model (GPT-4o) with respect to the individual viewpoints on both the full and restricted datasets. Apart from the poor performance on the problematic viewpoints, 8 and 9, it also possible to see that there are a couple of viewpoints (1 and 7) with very low precision and very high recall, for both the full and restricted datasets. The reason for this behaviour becomes clear if we analyse Figure 1 and Figure 2, which report the frequency of positives across the viewpoints for the two datasets and GPT-4o. In particular, viewpoints 1 and 7 are the ones for which the gap between the gold standard positives and the LLM positives is the largest. In addition, viewpoint 1 was the only one that was not picked up by the LLM in the various runs to identify potential viewpoints but was instead added by the domain expert. Hence, it is not surprising that, in a zero-shot setting, this is the one where GPT-4o exhibits the weakest performance.

Table 6. GPT-4o scores on the full dataset — all viewpoints

Viewpoint	Precision	Recall	F1
1	0.28	0.94	0.44
2	0.68	0.66	0.67
3	0.51	0.65	0.57
4	0.50	0.90	0.64
5	**0.71**	0.67	0.69
6	0.60	0.86	**0.71**
7	0.44	**0.96**	0.60
8	0.00	0.00	0.00
9	0.67	0.23	0.34

Table 7. GPT-4o scores on restricted dataset — 7 viewpoints.

Viewpoint	Precision	Recall	F1
1	0.40	0.96	0.56
2	0.83	0.75	0.79
3	0.75	0.65	0.70
4	**0.88**	0.88	**0.88**
5	0.87	0.70	0.77
6	0.80	0.80	0.80
7	0.48	**1.00**	0.65

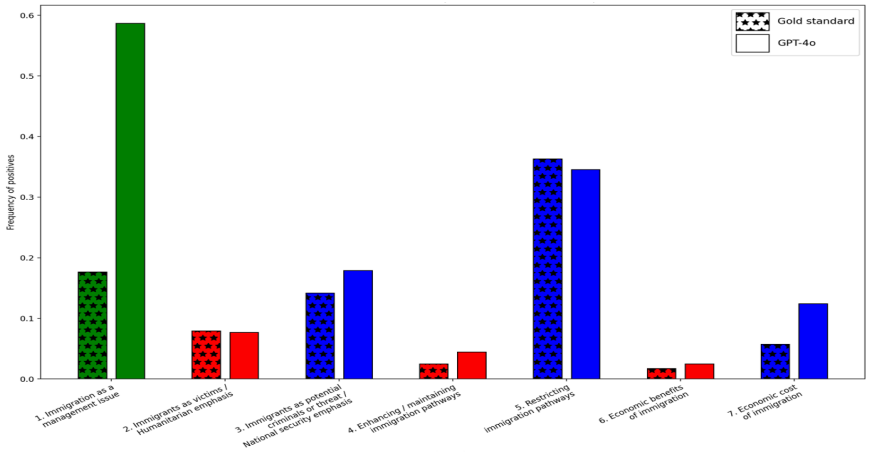

Fig. 1. Viewpoint percentages in full dataset and associated GPT-4o classification.

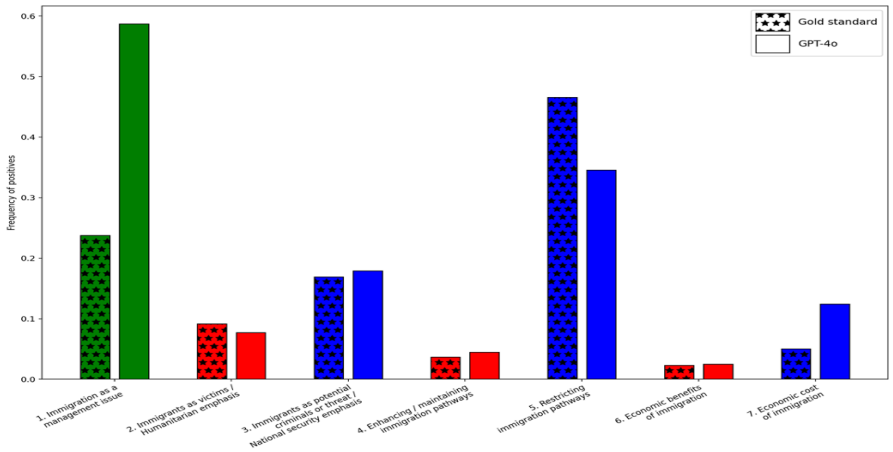

Fig. 2. Viewpoint percentages in restricted dataset and associated GPT-4o classification.

4 Application-level considerations

Despite the preliminary nature of the experiments reported in this paper, our results are already interesting from an application-level perspective and they appear to confirm the concerns expressed by scholars and commentators — see e.g., [21] and [22], who have highlighted distortions and lack of balance in UK's mainstream media coverage of key issues. In particular, Figure 1 and Figure 2 show that very little media coverage is allocated to pro-immigration viewpoints (shown in red in the figures), compared to anti-immigration ones (shown in blue in the figures). Indeed, we believe that a particular strength of our approach is the focus on modelling the news dynamics in a way that is consistent with the analyses carried out by media and political scientists — e.g., see [5],

who try to understand the dynamics of the debate on a particular issue. As shown in Figure 3, our approach can also support granular visualizations of the viewpoint dynamics over time, a feature that is particular useful when analysing the debate about a particular topic over a crucial time period — e.g., during the run up to an election.

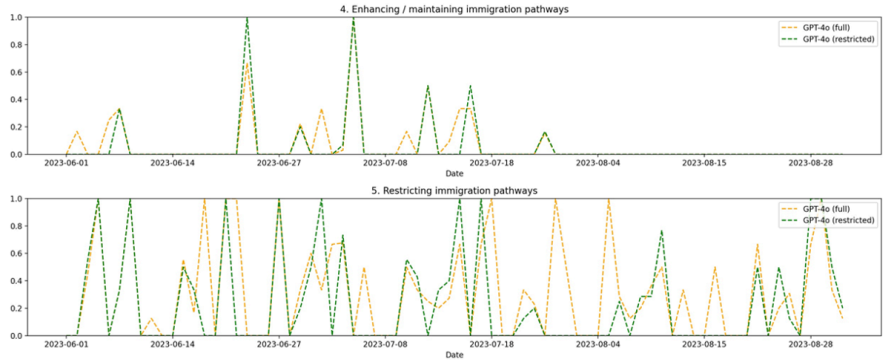

Fig. 3. Visualizing the viewpoint dynamics over time.

5 Conclusions

In this paper we have illustrated an approach and a set of initial experiments that use LLMs to model the viewpoint dynamics in the news. Our results indicate that, even in a zero-shot learning modality, the larger models, such as GPT-4o, already exhibit an acceptable level of performance. A key strength of our approach is that it goes beyond much computer science literature on capturing political opinions that, as discussed in Section 2.2, tends to adopt rather coarse-grained classifications — e.g., in favour or against a particular position — and fails to capture the variety of viewpoints that characterise the debate in the media. Having said so, we should also stress that this research is at a rather early stage and several issues still need to be tackled. These are discussed in what follows.

In the approach presented in this paper the viewpoints relevant to a debate are generated in advance of the claim classification task. Indeed, we believe that our hybrid human-machine approach worked well and the set of viewpoints presented in the paper provides a comprehensive and correct framework to analyse the immigration debate, as reported in the news during the June-August 2023 period. Nonetheless, this solution is structured in a rather waterfall fashion, where a corpus of news is considered, the viewpoints are generated and then the claims are classified. This works well for a retrospective analysis — e.g., a study on the election debate over the past few months, but would be unable to support the modelling of a live, ongoing debate, where new viewpoints may emerge dynamically. Hence, additional experiments will need to be carried out to test whether effective pipelines can be devised, which are able to handle the emergence of novel viewpoints in a debate. Such a solution would also make it possible to analyse the viewpoint dimensions themselves as an object of discourse — e.g., modelling the rate of change in viewpoint dimensions over time and across different topics.

Another issue concerns the typology of LLMs used in our study. Given the emphasis on measuring performance in a zero-shot setting, it is not surprising that the large, expensive and proprietary models produced the best results. Hence, more work is needed with smaller, cheaper and open models, first to test whether application-specific fine-tuning can significantly improve their performance in this task and secondly to develop practical pipelines where such fine-tuning can be effectively applied in user-centric solutions.

The claim extraction process also requires more work. While this aspect was not a priority in the context of the research reported here, more effort is needed to better characterise what is a claim and to develop highly performant techniques to extract them from a news corpus. Tackling these issues is essential to reduce the noise in the extracted corpus of claims, which ought to facilitate the manual construction of improved gold standards by human annotators and (most likely) lead to improvements in the performance of the LLMs on the claim classification task.

Another important challenge requires dealing with the contested nature of the domain. While we are happy with the robustness of the viewpoint generation process, it is clear that the claim classification task is a challenging one, given the inherent ambiguity of many claims expressed in the media. Indeed, as reported in Section 3.4, the level of agreement between human annotators was only moderate on the full dataset, thus affecting the robustness of our initial gold standard. While we were able to address the issue by generating a restricted version of the gold standard and by reducing the number of viewpoints, additional work is needed to give more robust foundations to the claim classification task, in particular, in terms of providing robust guidelines to human annotators. This aspect is closely related to the task discussed earlier: that is, it is our view that, if we are able to improve the claim extraction process and characterise more precisely what is a claim, then it will become easier to support human annotators in developing robust gold standards.

Finally, it is also important to emphasise that claim classification is a knowledge-intensive task. Statements by politicians can be ambiguous and in many cases can only be understood if the observer has enough background knowledge about their track record, political affiliation, key beliefs, etc. Hence, in the future we plan to extend our architecture by developing a *knowledge graph* [23] capturing key information about the actors expressing opinions in the media and by using it both to try and improve the performance of the system in the claim classification task and also to provide a wider range of domain analytics — e.g., by modelling the type and evolution of the positions expressed by individual actors or broader groups, such as political parties.

Acknowledgments. This work was partially supported by a grant from the Open Societal Challenges research programme of The Open University. The authors would also like to thank three anonymous EKAW reviewers for their insightful comments and criticisms.

Disclosure of Interests. The authors have no competing interests to declare, which are relevant to the content of this article.

References

1. Baden, C., Springer, N. Conceptualizing viewpoint diversity in news discourse. Journalism. 18, pp. 176–194 (2017). https://doi.org/10.1177/1464884915605028.
2. Bermes, A. Information overload and fake news sharing: A transactional stress perspective exploring the mitigating role of consumers' resilience during COVID-19. Journal of Retailing and Consumer Services. 61 (2021).
3. Boumans, J.W., Trilling, D. Taking Stock of the Toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. Digital Journalism. 4, pp. 8–23 (2016). https://doi.org/10.1080/21670811.2015.1096598.
4. Hamborg, F., Donnay, K., Gipp, B. Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal of Digital Libraries. 20, pp. 391–415 (2019). https://doi.org/10.1007/s00799-018-0261-y.
5. Masini, A., Van Aelst, P., Zerback, T., Reinemann, C., Mancini, P., Mazzoni, M., Damiani, M., Coen, S. Measuring and Explaining the Diversity of Voices and Viewpoints in the News: A comparative study on the determinants of content diversity of immigration news. Journalism Studies. 19, pp. 2324–2343 (2018).
6. Vreese, C.H. News framing: Theory and typology. Information Design Journal. 13, pp. 51–62 (2005). https://doi.org/10.1075/idjdd.13.1.06vre.
7. Motta, E., Daga, E., Gangemi, A., Gjelsvik, M.L., Osborne, F., Salatino, A. The Epistemology of Fine-Grained News Classification. Semantic Web Journal. (2024). https://www.semantic-web-journal.net/system/files/swj3659.pdf.
8. Doan, T.M., Gulla, J.A. A Survey on Political Viewpoints Identification. Online Social Networks and Media. 30, (2022). https://doi.org/10.1016/j.osnem.2022.100208.
9. Trabelsi, A., Zaiane, O. Unsupervised Model for Topic Viewpoint Discovery in Online Debates Leveraging Author Interactions. Proceedings of the International AAAI Conference on Web and Social Media. 12, (2018). https://doi.org/10.1609/icwsm.v12i1.15021.
10. Quraishi, M., Fafalios, P., Herder, E. Viewpoint Discovery and Understanding in Social Networks. In: Proceedings of the 10th ACM Conference on Web Science. pp. 47–56. Association for Computing Machinery, (2018). https://doi.org/10.1145/3201064.3201076.
11. Hada, R., Ebrahimi Fard, A., Shugars, S., Bianchi, F., Rossini, P., Hovy, D., Tromble, R., Tintarev, N.: Beyond Digital "Echo Chambers": The Role of Viewpoint Diversity in Political Discussion. Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. pp. 33–41. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539597.3570487.
12. Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., Helberger, N.: Recommenders with a Mission: Assessing Diversity in News Recommendations. Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. pp. 173–183. ACM (2021).
13. Küçük, D., Can, F.: Stance Detection: A Survey. ACM Computing Surveys 53, pp. 12:1–12:37 (2020). https://doi.org/10.1145/3369026.
14. Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., Roth, D.: Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In Burstein, J., Doran, C., and Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 542–557 (2019). https://doi.org/10.18653/v1/N19-1053.
15. Vrandečić, D., Krötzsch, M. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), pp. 78–85 (2014).
16. McHugh, M. L. Interrater reliability: the kappa statistic.Biochemia medica, 22(3), pp. 276–282 (2012).

17. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R. GPT-4 Technical Report. arXiv:2303.08774 (2023).
18. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.D.L., Hanna, E.B., Bressand, F. and Lengyel, G. Mixtral of Experts. arXiv:2401.04088 (2024).
19. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
20. Meta LLaMA Team. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. (2024)
21. Deacon D., Downey J., Harmer E., Stanyer, J., Wring, D. The narrow agenda: How the news media covered the Referendum. In Jackson D., Thorsen E. and Wring D. (eds), EU Referendum Analysis 2016, pp. 34–35 (2016).
22. Taylor, R. How well does the UK's media system support democratic politics and represent citizens' interests? Democratic Audit Blog. https://tinyurl.com/nz9rppfz. (2018).
23. Peng, C., Xia, F., Naseriparsa, M. and Osborne, F. Knowledge graphs: Opportunities and Challenges. Artificial Intelligence Review, 56(11), pp. 13071–13102 (2023).