

Integrating Large Language Models in Political Discourse Studies on Social Media: Challenges of Validating an LLMs-in-the-loop Pipeline

Giada Marino*  ^a

Fabio Giglietto  ^a

^a Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo (Italy)

Submitted: May 11, 2024 – Revised version: August 20, 2024
Accepted: August 21, 2024 – Published: October 30, 2024


Abstract

The integration of Large Language Models (LLMs) into research workflows has the potential to transform the study of political content on social media. This essay discusses a validation protocol addressing three key aspects of LLM-integrated research: the versatility of LLMs as general-purpose models, the granularity and nuance in LLM-uncovered narratives, and the limitations of human assessment capabilities. The protocol includes phases for fine-tuning and validating a binary political classifier, evaluating cluster coherence, and assessing machine-generated cluster label accuracy. We applied this protocol to validate an LLMs-in-the-loop research pipeline designed to analyze political content on Facebook during the Italian general elections of 2018 and 2022. Our approach classifies political links, clusters them by similarity, and generates descriptive labels for clusters. This methodology presents unique validation challenges, prompting a reevaluation of accuracy assessment strategies. By sharing our experiences, this essay aims to guide social scientists in employing LLM-based methodologies, highlighting challenges and advancing recommendations for colleagues intending to integrate these tools for political content analysis on social media.

Keywords: Large Language Models (LLMs); Political Discourse; Social Media; Natural Language Processing (NLP).

Acknowledgements

This work was partially supported by the vera.ai project, which is funded by the European Union through the Horizon Europe program (Grant Agreement ID 101070093).

*  giada.marino@uniurb.it

1 Introduction

Since ChatGPT's launch in November 2022, scholarly interest in Generative AI has grown significantly. A mini-review article published in August 2023 documented 156 Scopus-indexed publications referencing "ChatGPT" between November 2022 and April 2023 (Watters & Lemanski, 2023). As of April 2024, this number had surged to 4,642 publications for 2023 — with 1,303 in the social sciences — and 2,628 for 2024, with 622 in social sciences. This increase reflects widespread interest in generative AI's societal impacts and its integration into research practices, including those of social sciences (Rask & Shimizu, 2024), such as surveys, online experiments, and automated content analysis (Bail, 2024).

Large Language Models (LLMs), developed by organizations like OpenAI, Meta, Google, Anthropic, and Mistral AI, are versatile tools in natural language processing (NLP) workflows. These pre-trained models excel in general-purpose, prompt-based inferences and are widely used in chat-bot applications such as OpenAI's ChatGPT and Anthropic's Claude. Beyond chat-bots, LLMs' inferences are programmatically accessible and are known for their efficacy in zero-shot or few-shot learning tasks. They can be fine-tuned for specific needs across various domains. At their core, LLMs work by transforming text and multimedia content into numerical representations that capture core semantics. This process, referred to as embedding, is also performed by standalone embedding models and is currently used to enhance content retrieval in large datasets and support tasks like semantic search, clustering, topic modeling, and classification.

The potential of LLMs for text analysis and computational social science is widely recognized (Mu et al., 2024). However, concerns persist regarding their inherent limitations and biases (Grossmann et al., 2023), challenges with reproducibility (Balloccu et al., 2024; Chen et al., 2023), and the need for established best practices for their integration into research methodologies (Rask & Shimizu, 2024).

This essay contributes to the ongoing discourse by presenting a novel, fully LLM-integrated methodological pipeline, its text annotation, and analysis validation protocol. We focus on the unique challenges in validating such a pipeline, addressing a critical gap in current research on LLMs integration in social sciences. Our approach leverages state-of-the-art OpenAI models to uncover political narratives in Facebook-shared links during the 2018 and 2022 Italian general elections.

Our pipeline introduces LLMs in three ways: fine-tuning for binary classification of Italian political links, LLM-based embeddings for clustering similar political links, and direct API inferences for creating descriptive cluster labels.

Natural Language Processing (NLP) research has extensively employed transformer-based, fully fine-tuned models such as BERT, RoBERTa, DistilBERT, and XLNet to accomplish various tasks. However, despite the proliferation of domain-specific, language-specific, and task-specific versions, these models typically require fine-tuning before they can be effectively applied to specific tasks. Fine-tuning is both labor-intensive and computationally demanding (Bender et al., 2021). Once fine-tuned, the resulting model often performs well on the specific dataset, task, domain, or language, but its performance often degrades when any of these elements change. Traditional topic modeling algorithms, like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), face significant challenges, including limitations related to the granularity of topics and issues (Abdelrazek et al., 2023). They also require a delicate and cumbersome preliminary text-cleaning phase and produce clusters of words that can often be difficult for researchers to interpret and utilize effectively (Gillings & Hardie, 2022).

Implementing a fully LLM-based pipeline presents significant validation challenges, necessitating a reevaluation of established accuracy assessment strategies. Drawing on the experience gained while designing and validating the pipeline, we explore the specific choices made during the validation protocol, focusing on three key characteristics of LLM-integrated research that complicate accuracy evaluation: the versatility of LLMs as general-purpose models, offering numerous application options with varying degrees of supervision, from multilingual capabilities, including underrepresented languages in research, to diverse content types, tasks, and fields of study; the varying levels of granularity and nuance in LLM-uncovered narratives; and the limitations of human assessment capabilities when evaluating models pre-trained on extensive datasets.

Our tailored validation protocol addresses these issues in three phases: fine-tuning and validating a binary political classifier, evaluating cluster coherence, and assessing the accuracy of machine-generated cluster labels. By sharing our experiences, this essay aims to provide insights for social scientists considering LLM-based research designs, highlighting both challenges and potential solutions in employing these advanced technologies in NLP.

2 Pipeline and Research Question

2.1 The Pipeline

Our LLMs-in-the-loop pipeline has five steps (Figure 1), including the identification of political links (2), embedding/clustering (3/4), and the generation of cluster labels (5). All these steps leverage the advanced capabilities of models provided by OpenAI.

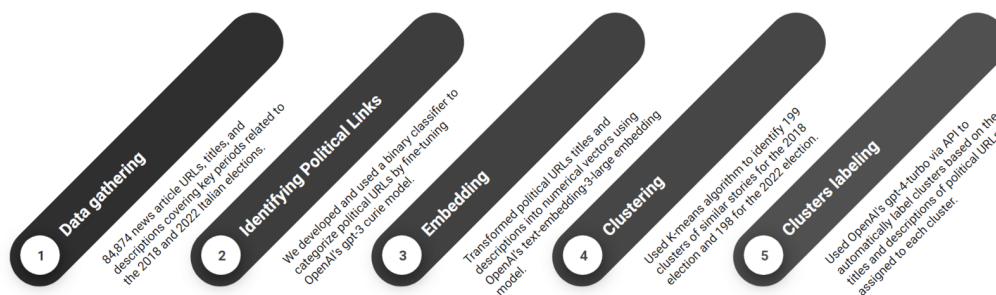


Figure 1: A graphical representation of the pipeline discussed in this article

2.1.1 Data Gathering

An initial dataset comprising 84,874 public news article URLs, titles, and descriptions was obtained by querying the Meta URL Shares Dataset for links first published on Facebook between December 24, 2017, and March 4, 2018, related to the 2018 election, and between July 21, 2022, and September 25, 2022, predominantly viewed by Italian users. Across the entire pipeline, only the title and description of these links (both are public content available at the respective URLs) have been used and thus fed to OpenAI's models.

2.1.2 Identifying Political Links

We developed a binary classifier to categorize political URLs by fine-tuning the GPT-3 *Curie* model, a now discontinued OpenAI model, suggested for this task. Seven Italian scholars with expertise in analyzing political news dissemination on social media supported the fine-tuning process. After standard training to ensure consistency (Krippendorff's alpha, Subjects = 200, Raters = 7, alpha = 0.812), they manually coded a proportional stratified (by-election and month) random sample of 4,190 URLs: 3,184 from 2018 and 1,006 from 2022. Excluding missing values and non-Italian URLs, the refined dataset for fine-tuning included 3,800 valid cases (1,801 political and 1,999 non-political). We concatenated titles and descriptions for each URL and filtered out non-Italian and empty titles and descriptions URLs, resulting in datasets of 59,838 URLs from 2018 and 17,690 from 2022. The classifier identified 54% of the 2018 posts and 53% of the 2022 posts as political, corresponding to 27,487 URLs in 2018 and 8,308 in 2022.

2.1.3 Grouping Together Similar Political Links

To identify clusters of similar links, we transformed our Italian political links into embeddings using a language model to convert the link's title and description into numerical vectors. After experimenting with various LLM-based embedding models (OpenAI's text-embedding-ada-002, Mistral AI's e5-mistral-7b-instruct (Wang et al., 2023), and OpenAI's text-embedding-3-large), we chose text-embedding-3-large based on clustering internal metrics. We preprocessed the text by removing HTML tags and hyperlinks before processing each URL's concatenated title and description.

Working with numerical vectors facilitates clustering-based topic modeling. Following OpenAI's recommendation, we used cosine distance to measure semantic similarity. We experimented with various clustering algorithms (k-means, DBSCAN, HDBSCAN, GenieClust (Gagolewski, 2021), and Kwikbucks (Silwal et al., 2023) and dimension reduction techniques (t-SNE and UMAP). Ultimately, we implemented cluster analysis using k-means with Lloyd's algorithm and retained all the initial 3,072 dimensions. Moreover, Giglietto's (2024) research on the same dataset demonstrated that LLMs outperform fully fine-tuned transformer models in NLP tasks. To determine the optimal number of clusters, we employed Bayesian optimization aimed at maximizing the Silhouette score and Hplus metric (Dyjack et al., 2023), ranging from 2 to 200 clusters, with 200 considered the maximum number for interpretability at the level of granularity requested by the scope of the study. This process identified 199 clusters as optimal for the 2018 election and 198 clusters for the 2022 election.

2.1.4 Clusters Labeling

We used GPT-4-turbo through the API to programmatically label clusters based on their content. The process involves feeding the model a sample of items from each cluster and requesting a short descriptive label. Table A3 (Appendix A) reports on the prompt specifics. The prompt includes a system prompt for context and output format and a user prompt supplying necessary documents.

The final prompt was crafted using strategies from OpenAI's prompt engineering guide (OpenAI, 2024) and tested for consistency across multiple runs and GPT models. The process of optimization was mainly aimed at instructing the model to output the specific label, only avoiding any further premise (e.g., "The label of the cluster is..") or comment. We prioritized a detailed prompt over cost optimization. Costs are computed per token, with different values for input and output tokens.

The model employed to generate the cluster labels includes training data up to December 2023, encompassing the 2022 election period. The total cost to label 199 clusters from 2018 and 198 from 2022 was \$30. Each label was requested using a prompt combining standard text with a density-based sample of cluster items. Despite GPT-4-turbo's 128,000 token capacity, we limited prompts to 8,000 tokens for a fair comparison with GPT-4. This approach achieved an average coverage of 84% of items per cluster.

2.2 Research Question

This essay focuses on how a full LLM-supported pipeline for social media political content annotation can be validated. This methodological approach poses several challenges, including issues with model reliability, data interpretation, integration of these models into existing research frameworks, and the relative newness of studies relying on these tools. For these reasons, this essay seeks to answer the following research question:

What are the main challenges researchers may face in validating an LLM-in-the-loop methodological approach, and how can they be addressed?

To answer this research question, each of the following sections of this essay is dedicated to a specific challenge we faced during our introduction of LLMs for text annotation tasks.

The next section discusses the general-purpose nature of LLMs, which are pre-trained on large datasets and perform general-purpose tasks based on given prompts. This means they can understand and generate text based on various inputs and handle different kinds of tasks, making them useful in many applications. Considering this adaptability, they may be employed, supervised or not, at different steps of a multiple research pipeline. This characteristic requires novel, tailored validation approaches. To address this, our protocol comprises three distinct phases, each corresponding to different LLM applications in our study.

The fourth section tackles the theoretical challenge of narrative definition. The possibility of unsupervised generation of embeddings with an almost unlimited number of dimensions and clustering them results in a cluster granularity that necessarily affects the validation preparation phase. This granularity ranges from general topics to specific journalistic stories. Our approach utilizes multi-level, detailed validation guidelines to ensure the accuracy and relevance of our findings.

The fifth and final section addresses the "knowledge" challenge. LLMs, having been trained on broad datasets, possess competencies that often surpass those of traditional coders. This necessitates a careful selection of the number and profiles of the coder team to ensure that the validation process is both thorough and effective.

3 Tailoring Validation Protocols for General-Purpose LLMs in Text Classification Tasks

While LLMs can be applied to a wide range of NLP tasks, their very complexity also means that their use in specific research contexts requires a significant degree of human guidance and decision-making. Unlike more narrowly defined machine learning models, LLMs are primarily designed for general-purpose, prompt-based inferences (Huang et al., 2023; Kuzman et al., 2023). Adapting them to meet the needs of a particular research objective or workflow involves a number of crucial human-led decisions at multiple stages.

In our case study, we needed to select a specific embedding model and clustering algorithm, as well as determine clustering parameters such as the number of clusters, the labeling model and prompt, and the sample size to input into the model. All these decisions require justification and must be evaluated against alternative options. Training a classifier, evaluating the performance of unsupervised cluster analysis, and extracting cluster narratives' require a training set or a ground truth to assess the LLMs' effectiveness in these tasks.

Researchers rely on different strategies and techniques to evaluate model fit. The prevailing methodologies for assessing the performance of LLMs typically involve a range of standardized tests covering areas from common sense reasoning and reading comprehension to arithmetic and coding. Although extensive, these benchmarks often fail to fully explore the nuanced capabilities afforded by a natural language interface. For instance, while they measure accuracy in specific tasks, they may not adequately assess the model's ability to handle ambiguous or contextually rich scenarios, nor do they always test for biases or the generation of novel content.

Similarly, embedding models are evaluated across diverse tasks such as classification, clustering, retrieval, and summarization (Muennighoff et al., 2022). However, these evaluations generally focus on optimizing straightforward metrics like accuracy or F1 scores, which might not capture more subjective qualities like the relevance or coherence of the content generated. Moreover, the language dependency of these tests presents another significant limitation. Most benchmarks are developed in English and subsequently translated for other languages, potentially skewing performance assessments due to translation inaccuracies or cultural nuances not being adequately represented. This approach can obscure the true versatility and effectiveness of LLMs and embedding models in non-English contexts, hence limiting our understanding of their global applicability and efficacy.

Given the limitations of automatic validation methods (Clinciu et al., 2021; Iskender et al., 2020), human evaluation has increasingly been recognized as a critical component in NLP research, either complementing or replacing these methods (Schuff et al., 2023). Validation protocols involving human teams require them to address specific research questions by following detailed guidelines, particularly when researchers test precise hypotheses (Schuff et al., 2023).

Given the general purpose nature of LLMs and the characteristics of our dataset, which includes a specific social media platform (Facebook), the domain (politics), and language-specific elements (Italian), the three-way LLMs are implemented in our workflow necessitated a distinct and tailored validation protocol to assess its efficacy. This implies that validating different LLM applications through existing standard processes employed for fully fine-tuned transformer models is challenging, and specific validation protocols are still under development. Furthermore, a validation workflow customized for our study might not be universally applicable or extendable to other datasets or domains.

In light of these considerations, we developed a three-step, ad hoc validation protocol. We opted for evaluation protocols involving human annotators.

The first round of validation pertained to the binary classifier of political vs. non-political URLs and was conducted employing standard validation approaches and measures. The fine-tuning dataset, manually labeled by seven human experts, was divided into training and validation sets. The training set was used to fine-tune the model, while the validation set assessed its performance, achieving an F1 score of 0.897, with a precision of 0.911 and a recall of 0.883.

The second round of validation regards a different task we accomplished through the use of LLMs, specifically cluster analysis. This phase involves assessing the coherence of clusters. Six expert coders, familiar with political content on social media and the Italian political landscape, evaluated a sample selected through systematic sequential pairing followed by a random subsampling, which comprised either 10% or at least five pairs from each cluster, totaling 2,754 pairs for the 2018 elections and 994 pairs for the 2022 elections. Coders were presented with pairs of links (Grimmer & King, 2011) from the same cluster and were required to assign a coherence level based on guidelines established during preliminary training.

Following the preliminary training, the coders were divided into three teams of two, each team comprising one experienced and one less experienced coder. Each team was assigned to a random subset of one-third of the items in the evaluation sample. Both team members independently coded the assigned pairs and held two meetings — a preliminary alignment meeting and a concluding meeting — to resolve any discrepancies in their evaluations with their teammates.

The guidelines provided to the coders (see Table A1 in Appendix A) use a scale ranging from 0, indicating a lack of coherence, to 4, indicating two items belonging to the same journalistic story, with an additional level for non-codable pair cases.

The last round of validation concerned the machine-generated labels. The goal is to evaluate how accurately each label represents the content it is intended to describe. The evaluation was carried out by the same six coders involved in the evaluation of the clusters' coherence. Following a phase of training performed on a pilot subsample of one item (and its respective label) for each cluster (199 for 2018 and 198 for 2022), the team agreed on a codebook consisting of four criteria (thematic alignment, implications, content coverage, and contextual alignment) and a three-level scale (misfit, partial fit, and good fit). Each coder is asked to rate the accuracy of a cluster label for one of the items assigned to that cluster. The evaluation employs a density-based sampling approach where each cluster contributes either a minimum of 10 items or 10% of its total, whichever is greater. This method ensures that each cluster is adequately represented in the sample. Specifically, the sampling technique is designed to represent proportionally the variety of centroid distances within each cluster. Focusing on density rather than a uniform distribution, the method ensures coverage across all regions of the distance distribution, from the closest to the furthest items from the centroid.

4 Validating LLMs-detected Political Narrative: Addressing Challenges in Theoretical Definition

Researchers have utilized various automatic classification methods to group similar social media political content and label them (Gupta et al., 2020). With the rise of social media as one of the primary news sources, narrative detection has become increasingly relevant. This is partly due to algorithmic indexing, which amplifies content based on its popularity, allowing certain narratives to gain more attention and thus be shown to more users.

However, defining the specific conditions that qualify a sequence of words or sentences as a narrative remains contentious in content annotation research. Despite its relevance, scholars

have struggled to reach a consensus on a definition. Generally, scholars agree that “narrative is a key concept for understanding human behavior and beliefs” (Piper et al., 2021, p. 298). Consistency in terminology is crucial for clearly defining the boundaries of the research object and setting the study’s objectives, particularly in NLP research, where a precise interpretation of linguistic phenomena is required.

In narratology, “narrative” refers to the structure of events involving a complex set of features such as time, context, participants, and the narrator’s perspective in organizing information (Genette, 1980; Pianzola, 2018; Piper et al., 2021).

In the analysis of political discourses, terms like “topics” or “issues” are more frequently used within the theoretical framework of the public agenda (Boydston, 2013). These terms serve as cognitive shortcuts that describe aspects of reality and vary in attention based on media coverage, thereby influencing public debate and political decisions (Scheufele, 2000).

A “political issue” is a subcategory of a topic describing an event or a series of events perceived as a significant problem by citizens (Wlezien, 2005). In political communication, various institutions and researchers label groups of content to study, for example, the main topics or issues of political parties and candidates’ campaigns (Illuminating, 2020) or disinformation during the elections in several European countries (EDMO, 2024).

The term “narrative” is less utilized in political communication studies because it is often conflated with storytelling or used in other scientific areas, such as linguistics. Groth (2019) refers to Eagleton (1979), who argued that narratives present closed stories with coherent logic, offering stringent explanations, causal relationships, and genealogies for socio-cultural and political realities. In this view, it is close to the definition of the more commonly used concept of media frames (Matthes & Kohring, 2008; McCombes et al., 2006; Reese, 2007)]. Also, Bradshaw et al. (2024) provide an insightful framework for examining “strategic narratives” in Russian discourse about the Ukrainian conflict. Drawing on prior literature, Popkova (2023) and Schmitt (2018) identify three key types of narrative manifestations: narratives related to international relations and global “world order,” identity narratives tied to a country’s culture and traditions, and issue-specific narratives focused on particular topics.

Also, Kotseva et al. (2023), employ a multidimensional hierarchical definition of narrative ranging from sub-narrative to super-narrative. Particularly interesting is the super-narrative definition. In comparison with the narrative, the super-narrative has a cross-temporal and cross-country nature, as a story-line that survives and evolves over time takes advantage every time and in different contexts of single events or local specificities.

In this fragmented scenario, the boundaries of a narrative are left to the discretion of researchers. When using supervised or semi-supervised methods, a tailored narrative definition is essential when setting a codebook for fine-tuning a transformer-based model for content annotation or cleaning datasets to achieve refined results (Groth, 2019; Kotseva et al., 2023). These approaches require the researcher to clearly delineate the scope and characteristics of the narratives upfront. In contrast, when using topic modeling techniques such as LDA, the dimensions of a narrative are left more open to interpretation based on the analysis outcomes.

Approaches that leverage LLMs for unsupervised or minimally supervised content annotation can produce results with varying levels of detail and granularity. As discussed earlier, this variability is not necessarily a weakness but rather a strength that allows for more nuanced and contextual findings.

In our own analysis, we took a theoretical holistic view, considering narratives as common story-lines that tap into collective memories, emotions, and historical analogies to achieve political objectives — aligning with the broader vision outlined by Bradshaw et al. (2024).

Challenges arise when researchers must validate these clustering outcomes. This process necessitates adaptable validation protocols that can assess different levels of coherence and accuracy. Clusters identified by k-means algorithms for our case tend to vary in both size and specificity. Some clusters are more generic, encompassing a range of closely related issues, while others are highly specific, tied to a single journalistic story or a particular media frame. This diversity in cluster characteristics underscores the need for flexible and robust validation methods.

To mitigate this issue during validation, we implemented some adaptation actions. Firstly, we evaluated cluster coherence by rating the coherence to random pairs of links extracted from each cluster (the guidelines are detailed in Table A1). We split the evaluation of coherence into three distinct levels. The basic level of coherence pertains to the topic as a broad area belonging to politics, such as economy, health, immigration, environment, safety, etc. A second, more specific level of coherence refers to stories with the same actor, event, place, or organization in common. Level three is the narrowest coherence estimation, and it regards only those pairs that refer to the same journalistic story, e.g., the murder case of Pamela Mastropietro in 2018. We also added a level 98 to indicate ambiguous cases or when the coder is uncertain. At the end of the coding phase, teammates discussed these specific cases to assign them another value in the scale.

We utilized a scale specifically designed to validate the accuracy levels of cluster labels generated by GPT-4-turbo. In contrast to assessing the coherence of the cluster — which relies on an established algorithm and innovative embeddings derived from Italian text — the application of an LLM to label the clusters is less conventional.

Moreover, these labels are critical for the subsequent phase of our research design, where exposure and engagement metrics will be calculated and analyzed based on the labels' meanings. Therefore, accurately assessing the labels' ability to represent the underlying content of each cluster is essential.

It is important to note that the two validation processes, though aimed at distinct tasks, are interconnected. A lack of coherence within a cluster would indeed hinder the creation of meaningful and representative short labels.

The rating scale adopted for evaluating the labels ranges from one (Misfit) to three (Good fit) (see Table A2 in Appendix A). The evaluation of label fit is based on four criteria established by the team of coders during the alignment meeting. More specifically, these criteria include:

- The thematic alignment criterion measures the extent to which the label corresponds to the central themes or subjects discussed in the item. Thematic alignment verifies that the label directly includes the primary topic addressed by the item.
- The implications or connotations suggested by the label. This criterion checks whether the label implies any outcomes, consequences, or broader trends consistent with the information or narrative provided in the item, ensuring that the label does not exaggerate, oversimplify, or misrepresent the content's potential impacts or significance.
- The content coverage standard assesses if the label encapsulates the key elements, facts, and details presented in the item. Content coverage ensures that the label addresses all significant points, leaving no major aspect of the content unrepresented or inaccurately portrayed. Additionally, a label should not encompass themes or details that extend beyond the scope of the item, which could mislead the understanding of the item's focus.
- The contextual alignment criterion evaluates the label's accuracy in reflecting the item's geographical, cultural, historical, or situational context. Contextual alignment confirms

that the label is suitable for the specific setting in which the content is placed, adhering to any particular nuances that influence content understanding.

The lowest value is attributed when a label completely fails to align with the item's content. The partial fit judgment is assigned when the label relates to the item in terms of theme and implications, but either the content covered by the label is too narrow or broad, or its context diverges from the item's context. This is the case, for example, with news discussing the rise in unemployment rates, specifically in rural parts of Italy due to local factory closures, and the label generated by the LLM for the cluster is "*Economic Challenges in the European Union.*" There is a good fit between a label and a piece of news when it accurately represents the item across all aspects. We consider a good fit case, for example, news related to rescue operations off the Sicilian coast highlighting the ongoing challenges faced by migrants and labeled as "*Migrant Crisis and Humanitarian Efforts in the Mediterranean.*"

5 Dealing with LLMs' High Levels of Knowledge

Assessing the reliability of LLM content annotation is a fundamental step in the process (Chiang & Lee, 2023; Gilardi et al., 2023), particularly challenging within complex research designs. Despite the enormous analytical opportunities and creative potential afforded by LLMs (Gilardi et al., 2023; Jahan et al., 2023), human evaluation remains essential.

Historically, human evaluation has been crucial to understanding the performance of natural language processing (NLP) models or algorithms (Gillick & Liu, 2010; Guzmán et al., 2015). We rely on human evaluators because certain textual aspects are difficult to assess with automatic evaluation metrics, necessitating human judgment either to train the model or to rate the quality of its outputs. However, human evaluation is known for its instability (Clark et al., 2021; Gillick & Liu, 2010), attributed to factors ranging from the quality of the workforce (Karpinska et al., 2021) to challenges in reproducing the same tasks or training human experts to provide consistent assessments (Chiang & Lee, 2023). Despite these limitations, human evaluation is prevalent and commonly considered indispensable in NLP, offering advantages over automatic metrics when carefully implemented.

In addition to the training phases, the most relevant task in models of human validation is recruiting the most appropriate team of annotators for the task. Primary strategies for recruiting annotators include hiring and training coders, such as students or research assistants, or utilizing crowdsourced work services like Amazon Mechanical Turk (MTurk) (Kasthuriarachchy et al., 2021). These strategies may be used individually or in combination, with trained coders annotating relatively small datasets considered gold standards and crowd workers increasing the volume of annotations (Gilardi et al., 2023). However, the limitations of these approaches increase when using LLMs for content annotation tasks, as they have been shown to outperform crowd workers, especially in complex tasks (Gilardi et al., 2023; Huang et al., 2023; Törnberg, 2023). Specifically, in the context of political communication research, LLMs possess significant knowledge of political and cultural contexts compared to a low-skilled workforce. Additionally, recruiting students or research assistants with deep knowledge of the political context is challenging, and this specific training is extremely time-consuming and resource-intensive. Furthermore, when conducting research in less commonly spoken languages, such as those other than English or Spanish, the recruitment process becomes complicated due to the scarcity of native-speaking crowd workers.

We conducted our validation rounds with these challenges in mind. Initially, we considered recruiting crowd workers from Fiverr, a platform that facilitates the hiring of Italian freelancers, and selected eight coders with expertise in copy-editing and data analysis.

However, after careful consideration, we decided against using crowd workers for validating our results. During the validation phase of our pipeline, we needed to thoughtfully select evaluators to assess the quality of clustering and labeling. As previously discussed, studies have shown that crowd workers may underperform compared to large language models in certain content annotation tasks (Gilardi et al., 2023; Huang et al., 2023; Törnberg, 2023; Zhang et al., 2023). This prompted us to reconsider whether crowd workers would be the most appropriate judges for an approach that surpasses their performance in the same tasks. For instance, when assessing the accuracy of the clustering, we encountered several cases that were challenging even for experts familiar with the national political context. This difficulty arises because it is unreasonable to expect humans to recall every specific political event and actor over the years. To illustrate with an example from our dataset, during one of the coder training sessions in the validation phase, we encountered the following story included in the cluster labeled as “*Corruption and Criminal Allegations in Italian Politics and Public Services*”:

Amedeo Matacena has died: struck down by a sudden illness. Matacena died at 59 years old in Abu Dhabi, where he had been living for years. The former Forza Italia deputy Amedeo Matacena, a well-known entrepreneur from Reggio Calabria, died at the age of 59. He was the son of the shipowner of the same name who passed away in 2003, and he was famous for initiating the ferry service across the Strait of Messina with Caronte [...] (translated from the original in Italian).

At a first look, this news seems to deal with the death of a secondary, former Italian politician. It was necessary to google the name Amedeo Matacena to discover that he had been convicted of involvement in a mafia association and had been a fugitive in Abu Dhabi until his death.

Given these challenges, we thus decided to rely on expert researchers in political communication to conduct the three validation rounds. As mentioned in the previous paragraph, we employed a team of seven coders for the fine-tuning and validation phase of the binary political classifier. This team consisted of all the authors of a paper we presented at the annual conference of the Italian Political Communication Association in 2023. Except for one PhD student, all co-authors are postdoctoral researchers and associate professors specializing in political communication and social media studies, and all are native Italian speakers.

In the second and third rounds of validation, we employed a team of six expert coders, four of whom had also participated in the first round. In this instance, the annotators were all PhD candidates, postdoctoral researchers, and associate professors focusing on political communication and social media research topics and all were native or proficient in Italian.

The less expert researchers were trailed and supervised by the more proficient ones, in particular concerning knowledge of the last ten years’ Italian political scenario. The processes of the second and third rounds of validation are described extensively in Section Three.

6 Conclusions

In this work, we pioneer the exploration of multiple validation protocols for different tasks in political discourse annotation using LLMs. Incorporating LLMs in natural language processing marks a significant paradigm shift within the field, offering a viable and adaptable method

for mostly unsupervised clustering analysis and narrative extraction. Indeed, they bring the potential of transformer language models like BERT to topic modeling methods (Mu et al., 2024). LLMs demonstrate their capability to handle specific domain, platform, and cultural context datasets with little to no fine-tuning required.

Thanks to their general-purpose nature, LLMs can manage extensive and complex tasks, enabling elaborate methodological pipelines. Specifically, we employed LLMs for three different tasks on two datasets of Facebook links related to the 2018 and 2022 Italian elections. We thus used LLMs in model fine-tuning to build a highly reliable binary classifier of political and non-political links, to generate LLM-based embeddings to cluster similar political content, and to make inferences via API to create short descriptive labels for the identified clusters.

However, using LLMs in all the steps of our NLP pipeline also introduces several new challenges, particularly in validating methodologies. We faced major challenges, particularly when we evaluated the outcomes of the unsupervised tasks, such as cluster analysis and label generation. At a general level, an LLM-in-the-loop pipeline necessitates distinct and tailored validation steps to assess the efficacy of each of the pipeline actions. In cluster analysis outcomes, we observed that LLMs can generate clusters/narratives with varying granularity levels, affecting how we consider the items within the same narrative group accurate or coherent and requiring highly detailed and adaptable codebooks. Moreover, LLMs' deep understanding of political and cultural contexts impacts the selection of the workforce for validation processes involving human participants, challenging traditional methods of recruiting content annotators and making the involvement of high-profile experts necessary. The versatility of LLMs encourages the phasing out of outdated annotation methods previously used in NLP studies. For instance, reliance on a low-skilled workforce annotation through crowdsourcing services like Amazon Mechanical Turk may become less necessary, as LLMs can efficiently process and understand large datasets with greater accuracy. This shift necessitates the development of new, robust validation protocols that keep pace with the rapid advancements in machine learning and artificial intelligence. These protocols must ensure that the models are not only effective but also free from bias and ethically compliant. Our validation protocol, for example, attempts to address potential biases by implementing a human-led task-by-task evaluation that relies fully on experts (Pangakis et al., 2023). Regarding the ethical concern of using models from proprietary providers for political content annotation, we mitigated this issue by choosing to provide the model with titles and brief descriptions of news stories that are already publicly available. Thus, we did not expose any proprietary, private, or sensitive information to the model.

The development of validation protocols for using LLMs to analyze the digital political discourse is a compelling issue. A timely implementation of LLMs in this field of studies may, in fact, be crucial in preventing their misuse.

Over the last two decades, each technological tool producing information flows has been susceptible to exploitation by malicious actors to spread problematic information and manipulate public opinion. In this context, LLMs can act as a double-edged sword. Prompt and competent adoption of these tools by political communication and science researchers may be pivotal in preventing or tackling such abuses and safeguarding the integrity of information while promoting responsible technology use in society.

Overall, the advancement of LLMs in NLP has opened new avenues for research and application. Although our research design is particularly complex, including various rounds of annotation that exploited LLMs for different tasks, the resources consumed in terms of time, costs, and researchers involved are, considering the scale of the project, limited.

Finding effective validation solutions that minimize the challenges of implementing an

LLMs-in-the-loop pipeline for content annotation may facilitate the introduction of LLMs into social science research. We wrote this essay to share our experience and expect it to serve as a guide to other researchers who would introduce LLMs in their studies. We hope that sharing our knowledge can contribute to the early adoption of similar methodological approaches using LLMs for digital political content annotation.

References

- Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic Modeling Algorithms and Applications: A Survey. *Information Systems*, 112, 1–17. <https://doi.org/10.1016/j.is.2022.102131>
- Bail, C.A. (2024). Can Generative AI Improve Social Science? *Proceedings of the National Academy of Sciences of the United States of America*, 121(21), 1–10. <https://doi.org/10.1073/pnas.2314021121>
- Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. *arXiv*. <https://doi.org/10.48550/arXiv.2402.03927>
- Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Boydston, A.E. (2013). *Making the News: Politics, the Media & Agenda Setting*. Chicago, IL: University of Chicago Press.
- Bradshaw, S., Elswah, M., Haque, M., & Quelle, D. (2024). Strategic Storytelling: Russian State-Backed Media Coverage of the Ukraine War. *International Journal of Public Opinion Research*, 36(3), edae028. <https://doi.org/10.1093/ijpor/edae028>
- Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's Behavior Changing over Time? *arXiv*. <https://doi.org/10.48550/arXiv.2307.09009>
- Chiang, C.-H., & Lee, H.-Y. (2023). Can Large Language Models Be an Alternative to Human Evaluations? *arXiv*. <https://doi.org/10.48550/arXiv.2305.01937>
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text. *arXiv*. <https://doi.org/10.48550/arXiv.2107.00061>
- Cliniciu, M., Eshghi, A., & Hastie, H. (2021). A Study of Automatic Metrics for the Evaluation of Natural Language Explanations. *arXiv*. <https://doi.org/10.48550/arXiv.2103.08545>
- Dyjack, N., Baker, D.N., Braverman, V., Langmead, B., & Hicks, S.C. (2023). A Scalable and Unbiased Discordance Metric with H. *Biostatistics*, 25(1), 188–202. <https://doi.org/10.1093/biostatistics/kxaco35>
- Eagleton, T. (1979). Ideology, Fiction, Narrative. *Social Text*, 2, 62–80. <https://doi.org/10.2307/466398>

- European Digital Media Observatory (EDMO). (2024). *Disinformation Narratives during the 2023 Elections in Europe*. <https://edmo.eu/publications/second-edition-march-2024-disinformation-narratives-during-the-2023-elections-in-europe/>
- Gagolewski, M. (2021). genieclust: Fast and Robust Hierarchical Clustering. *SoftwareX*, 15, 100722. <https://doi.org/10.1016/j.softx.2021.100722>
- Genette, G. (1980). *Narrative Discourse: An Essay in Method*. (J.E. Lewin, Trans.). Ithaca, NY: Cornell University Press. (Original work published 1972)
- Giglietto, F. (2024). Evaluating Embedding Models for Clustering Italian Political News: A Comparative Study of Text-Embedding-3-Large and UmBERTo. *OSF Preprints*. <https://doi.org/10.31219/osf.io/2j9ed>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd Workers for Text-annotation Tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(30), e2305016120. <http://doi.org/10.1073/pnas.2305016120>
- Gillick, D., & Liu, Y. (2010). Non-expert Evaluation of Summarization Systems is Risky. In C. Callison-Burch, & M. Dredze (Eds.), *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 148–151). Association for Computational Linguistics. <https://aclanthology.org/W10-0722>
- Gillings, M., & Hardie, A. (2022). The Interpretation of Topic Models for Scholarly Analysis: An Evaluation and Critique of Current Practice. *Digital Scholarship in the Humanities*, 38(2), 530–543. <https://doi.org/10.1093/lc/fqac075>
- Grimmer, J., & King, G. (2011). General Purpose Computer-assisted Clustering and Conceptualization. *PNAS*, 108(7), 2643–2650. <https://doi.org/10.1073/pnas.1018067108>
- Grossmann, I., Feinberg, M., Parker, D.C., Christakis, N.A., Tetlock, P.E., & Cunningham, W.A. (2023). AI and the Transformation of Social Science Research. *Science*, 380(6650), 1108–1109. <https://doi.org/10.1126/science.ad11778>
- Groth, S. (2019). Political Narratives / Narrations of the Political: An Introduction. *Narrative Culture*, 6(1), 1–18. <https://doi.org/10.13110/narrcult.6.1.0001>
- Gupta, S., Bolden, S., & Kachhadia, J. (2020). *PoliBERT: Classifying Political Social Media Messages with BERT* (Working paper SBP-BRIMS 2020 conference). Social, Cultural. <https://news.illuminating.ischool.syr.edu/2020/11/24/polibert-classifying-political-social-media>
- Guzmán, F., Abdelali, A., Temnikova, I., Sajjad, H., & Vogel, S. (2015). How Do Humans Evaluate Machine Translation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, & P. Pecina (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 457–466). Association for Computational Linguistics. <https://aclanthology.org/W15-3059/>
- Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In Y. Ding, J. Tang, J. Sequeda, L. Aroyo, C. Castillo, & G.-J. Houben (Eds.), *Companion Proceedings of the ACM Web Conference 2023* (pp. 294–297). ACM Digital Library. <https://doi.org/10.1145/3543873.3587368>

- Illuminating. (2020). *2020 Presidential Campaign Facebook and Instagram Ads*. https://illuminating.ischool.syr.edu/campaign_2020/
- Iskender, N., Polzehl, T., & Möller, S. (2020). Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation. In S. Eger, Y. Gao, M. Peyrard, W. Zhao, & E. Hovy (Eds.), *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 164–175). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.eval4nlp-1.16>
- Jahan, I., Laskar, M.T.R., Peng, C., & Huang, J. (2023). Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers. *arXiv*. <https://doi.org/10.48550/arXiv.2306.04504>
- Karpinska, M., Akoury, N., & Iyyer, M. (2021). The Perils of Using Mechanical Turk to Evaluate Open-ended Text Generation. In M.F. Moens, X. Huang, L. Specia, & S. Wen-tau Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1265–1285). Association for Computational Linguistics. <https://10.18653/v1/2021.emnlp-main.97>
- Kasthuriarachchy, B., Chetty, M., Shatte, A., & Walls, D. (2021). Cost Effective Annotation Framework Using Zero-shot Text Classification. *Proceedings of the 2021 International Joint Conference on Neural Networks* (pp. 1–8). IEEE. <https://doi.org/10.1109/IJCNN52387.2021.9534335>
- Kotseva, B., Vianini, I., Nikolaidis, N., Faggiani, N., Potapova, K., Gasparro, C., Steiner, Y., Scornavacche, J., Jacquet, G., Dragu, V., Della Rocca, L., Bucci, S., Podavini, A., Verile, M., Macmillan, C., & Linge, J. (2023). Trend Analysis of COVID-19 Mis/Disinformation Narratives: A 3-year Study. *PLOS ONE*, *18*(11), 1–26. <https://doi.org/10.1371/journal.pone.0291423>
- Kuzman, T., Mozetic, I., & Ljubešić, N. (2023). ChatGPT: Beginning of an End of Manual Linguistic Data Annotation. Use Case of Automatic Genre Identification. *arXiv*. <https://doi.org/10.48550/arXiv.2303.03953>
- Matthes, J., & Kohring, M. (2008). The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *The Journal of Communication*, *58*(2), 258–279. <https://doi.org/10.1111/j.1460-2466.2008.00384.x>
- McCombes, M., Lopez-Escobar, E., & Llamas, J.P. (2006). Setting the Agenda of Attributes in the 1996 Spanish General Election. *The Journal of Communication*, *50*(2), 77–92. <https://doi.org/10.1111/j.1460-2466.2000.tb02842.x>
- Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. *arXiv*. <https://doi.org/10.48550/ARXIV.2210.07316>
- Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. *arXiv*. <http://arxiv.org/abs/2403.16248>
- OpenAI. (2024). *Prompt Engineering*. <https://platform.openai.com/docs/guides/prompt-engineering/prompt-engineering>

- Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated Annotation with Generative AI Requires Validation. *arXiv*. <https://doi.org/10.48550/arXiv.2306.00176>
- Pianzola, F. (2018). Looking at Narrative as a Complex System: The Proteus Principle. In R. Walsh & S. Stepney (Eds.), *Narrating Complexity* (pp. 101–122). NY, New York: Springer International Publishing.
- Piper, A., So, R.J., & Bamman, D. (2021). Narrative Theory for Computational Narrative Understanding. In Moens, M.-F., Huang, X., Specia, L., & Yih, S. W.-T. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 298–311). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.26>
- Popkova, A. (2023). Strategic Narratives of Russiagate on Russian Mainstream and Alternative Television. In O. Boyd-Barrett & S. Marmura (Eds.), *Russiagate Revisited: The Aftermath of a Hoax* (pp. 203–223). NY, New York: Springer International Publishing.
- Rask, M., & Shimizu, K. (2024). Beyond the Average: Exploring the Potential and Challenges of Large Language Models in Social Science Research. *Proceedings of the 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications* (pp. 1–5). IEEE. <https://doi.org/10.1109/ACDSA59508.2024.10467341>
- Reese, S.D. (2007). The Framing Project: A Bridging Model for Media Research Revisited. *The Journal of Communication*, 57(1), 148–154. <https://doi.org/10.1111/j.1460-2466.2006.00334.x>
- Scheufele, D.A. (2000). Agenda-Setting, Priming, and Framing Revisited: Another Look at Cognitive Effects of Political Communication. *Mass Communication and Society*, 3(2–3), 297–316. https://doi.org/10.1207/S15327825MCS0323_07
- Schmitt, O. (2018). When Are Strategic Narratives Effective? The Shaping of Political Discourse through the Interaction between Political Myths and Strategic Narratives. *Contemporary Security Policy*, 39(4), 487–511. <https://doi.org/10.1080/13523260.2018.1448925>
- Schuff, H., Vanderlyn, L., Adel, H., & Vu, N.T. (2023). How to Do Human Evaluation: A Brief Introduction to User Studies in NLP. *Natural Language Engineering*, 29(5), 1199–1222. <https://doi.org/10.1017/S1351324922000535>
- Silwal, S., Ahmadian, S., Nystrom, A., McCallum, A., Ramachandran, D., & Kazemi, S.M. (2023). KwikBucks: Correlation Clustering with Cheap-weak and Expensive-strong Signals. In N.S. Moosavi, I. Gurevych, Y. Hou, G. Kim, Y.J. Kim, T. Schuster, & A. Agrawal (Eds.), *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing* (pp. 1–31). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.sustainlp-1.1>
- Törnberg, P. (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv*. <https://doi.org/10.48550/arXiv.2304.06588>
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2023). Improving Text Embeddings with Large Language Models. *arXiv*. <https://doi.org/10.48550/arXiv.2401.00368>

- Watters, C., & Lemanski, M.K. (2023). Universal Skepticism of ChatGPT: A Review of Early Literature on Chat Generative Pre-trained Transformer. *Frontiers in Big Data*, 6. <https://doi.org/10.3389/fdata.2023.1224976>
- Wlezien, C. (2005). On the Salience of Political Issues: The Problem with “Most Important Problem.” *Electoral Studies*, 24(4), 555–579. <https://doi.org/10.1016/j.electstud.2005.01.009>
- Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. (2023). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12, 39–57. https://doi.org/10.1162/tacl_a_00632

Appendix A

Table A1. Cluster Coherence Assessment Guidelines Scheme

Levels	Definitions and examples
Level 0: No Coherence	<p>Definition: The two links have nothing in common.</p> <p>Example: One link discusses an environmental policy regarding renewable energy, while the other covers a new education curriculum in schools. These stories do not share thematic elements.</p>
Level 1: Broad Thematic Coherence	<p>Definition: The two links pertain to the same broad area of politics (e.g., economy, health, taxes, immigration, environment, safety, ...) but refer to stories with different actors, events, places, or organizations.</p> <p>Example: Both links cover economic issues. One is about tax reforms affecting small businesses, and the other discusses federal spending on infrastructure. They share a broad theme of economic policy but focus on distinct topics.</p>
Level 2: Specific Thematic Coherence	<p>Definition: The two links have specific actors, events, places, or organizations in common but refer to different journalistic stories.</p> <p>Example: Both stories mention the World Health Organization's response to health crises but from different angles—one focuses on funding and resource allocation, while the other examines the impact of WHO guidelines on national health policies.</p>
Level 3: Same Journalistic Story	<p>Definition: The two links refer to the same journalistic story, covering the same actors, events, places, and organizations with closely related narratives.</p> <p>Example: Both links detail discussions and outcomes of a specific international climate summit, including the same participating countries, agreed-upon actions, and criticisms from environmental groups.</p>
Level 98: I do not know/I am not sure/one or both links contain multiple themes/stories	<p>Definition: The coder is unable to assess the coherence (either because of a lack of knowledge or because of the nature of the content).</p> <p>Example: At least one of the link titles or descriptions do not clearly convey its topic to the coder. </p>

Table A2. Labels accuracy assessment guidelines scheme

Levels	Definitions and examples
Misfit	<p>Definition: The label fails to align with the item's content, missing significant aspects or inaccurately representing its implications.</p> <p>Conditions: If the label fails to meet either the Thematic Alignment or Implications criteria, it is automatically categorized as a Misfit.</p> <p>Criteria Examples</p> <ol style="list-style-type: none"> 1. Thematic Alignment: The label introduces themes or subjects completely absent in the item. 2. Implications: The label implies a stance or narrative that contradicts the item's factual content or focus.
Partial Fit	<p>Definition: The label relates to the item in terms of theme and implications, but either the content covered by the label is too narrow or broad (include also other distinct themes not discussed by the item), or its context diverges from the context of the item.</p> <p>Conditions: The label meets the criteria for Thematic Alignment and Implications but fails to completely cover Content Coverage and/or Contextual Alignment.</p> <p>Criteria Examples</p> <ol style="list-style-type: none"> 1. Thematic Alignment: The label addresses the key theme of the item. 2. Implications: The label accurately reflects the item implications. 3. Content Coverage: The label encompasses themes or details that extend beyond the scope of the item or cover part of the item's content well but overlooks or inaccurately represents other significant parts. 4. Contextual Alignment: The label fails to reflect the item's specific geographical, cultural, or situational context accurately.
Good Fit	<p>Definition: The label fully and accurately represents the item across all aspects.</p> <p>Conditions: The label must completely satisfy all four criteria: Thematic Alignment, Content Coverage, Contextual Alignment, and Implications.</p> <p>Criteria Examples</p> <ol style="list-style-type: none"> 1. Thematic Alignment: Addresses the main themes or significant content of the item clearly. 2. Implications: Accurately reflects the implications or conclusions supported by the content. 3. Content Coverage: Captures all critical details, with only minor aspects possibly overlooked. 4. Contextual Alignment: Fits well within the context presented in the item with only slight inaccuracies.

Table A3. Prompts we used to feed gpt-4-turbo

Role	Message
System	“You are an assistant tasked with aiding a political scientist in analyzing social media content related to the [2018/2022] Italian elections. Your objective is to synthesize the core themes of groups of politically themed links shared on Facebook into succinct, descriptive labels in English. These labels should encapsulate the primary themes, issues, or narratives prevalent among the links in each group, providing a concise overview of their collective content.”
User	“Presented below is a selection of links from one such group. Each entry merges the title and description of a link, offering a glimpse into its thematic content. Based on these summaries, identify and articulate overarching themes or characteristics shared across these links. Your response should be a concise, descriptive phrase or label that accurately captures these shared elements. This label will be instrumental in cataloging and analyzing the political discourse related to the [year] Italian elections on Facebook. [text] (Placeholder where the actual items to be analyzed are inserted.) What descriptive English label best summarizes these shared characteristics?”

Giada Marino – Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo (Italy)

ORCID <https://orcid.org/0000-0002-9087-2608> | ✉ giada.marino@uniurb.it

📧 <https://aoir.social/@GiadaM>

Giada Marino holds a Ph.D. from the University of Urbino Carlo Bo (Italy), where she is currently a postdoctoral researcher in the vera.ai project. Her research focuses on the intersection of information disorder and political polarization, with a particular emphasis on how citizens engage with political content and discussions on social media platforms. She uses a mixed methods approach to analyze these dynamics.

Fabio Giglietto – Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo (Italy)

ORCID <https://orcid.org/0000-0001-8019-1035>

📧 <https://aoir.social/@fabiogiglietto>

Fabio Giglietto is an Associate Professor of Internet Studies at the University of Urbino Carlo Bo, holding a Ph.D. from the same institution. He explores the complex relationship between information theory, media, and digital technologies and their impact on society. His research delves into how these forces shape social systems and influence public opinion and political communication. Giglietto also employs advanced computational methods to analyze online information dissemination and media manipulation.