UKElectionNarratives: A Dataset of Misleading Narratives Surrounding Recent UK General Elections

Fatima Haouari[1], Carolina Scarton[1], Nicolò Faggiani[2], Nikolaos Nikolaidis[3], Bonka Kotseva[4], Ibrahim Abu Farha[1], Jens Linge[5], Kalina Bontcheva[1]

[1]Department of Computer Science, The University of Sheffield, Sheffield, UK
[2]Engineering S.p.A., Rome, Italy
[3]Athens University of Economics and Business, Athens, Greece
[4]Piksel S.r.l., Ispra, Italy
[5]European Commission, Joint Research Centre, Ispra, Italy
{f.haouari, c.scarton, k.bontcheva}@sheffield.ac.uk

Abstract

Misleading narratives play a crucial role in shaping public opinion during elections, as they can influence how voters perceive candidates and political parties. This entails the need to detect these narratives accurately. To address this, we introduce the first taxonomy of common misleading narratives that circulated during recent elections in Europe. Based on this taxonomy, we construct and analyse **UKElectionNarratives**: the first dataset of human-annotated misleading narratives which circulated during the UK General Elections in 2019 and 2024. We also benchmark Pre-trained and Large Language Models (focusing on GPT-4o), studying their effectiveness in detecting election-related misleading narratives. Finally, we discuss potential use cases and make recommendations for future research directions using the proposed codebook and dataset.

Introduction

Numerous misleading narratives tend to emerge during elections in Europe, often reflecting a mix of local and global concerns (Colliver et al. 2019; Panizio 2024). Specifically, this paper adopts the European Digital Media Observatory (EDMO) definition which describes misleading narratives as *"the clear message that emerges from a consistent set of contents that can be demonstrated as false using the fact-checking methodology"*(Panizio 2024).

While each European country may encounter unique narratives shaped by its national political context, some shared misleading narratives tend to emerge as well. Common divisive ones tend to focus on the integrity of the electoral process, economic uncertainty, foreign interference, and social issues, such as gender dynamics, religious tensions, and immigration. These narratives exploit societal divisions and amplify existing tensions to influence public opinion, and frequently serve as focal points for disinformation campaigns (Panizio 2024). Therefore, a dataset of common misleading election narratives can provide a valuable, unique opportunity for researchers to develop innovative techniques for detecting, analysing, and countering narratives and disinformation campaigns on a large scale (Colliver et al. 2019).

This paper addresses this crucial research need to study misleading narratives created and spread on social media during elections. The main contributions of this work are:

1. The first public multi-level taxonomy of misleading narratives that emerged during recent elections in Europe.[1]
2. The first human-annotated social media dataset **UKElectionNarratives**[2] of misleading narratives that were spreading during the UK General Elections in 2019 and 2024.
3. Benchmarking results on the new **UKElectionNarratives** dataset, which explore the effectiveness of Pre-trained (PLMs) and Large Language Models (LLMs) (particularly, GPT-4o) in detecting misleading narratives. The source code of these experiments is also made available for reproducibility.[3]

Related Work

This section discusses related work and gaps in the detection and discovery of misleading narratives, together with an examination of existing relevant datasets.

Misleading Narratives Studies

Research on misleading narratives spans various areas, employing diverse methodologies and datasets. In the health domain, Ganti et al. (2023) re-annotate health misinformation datasets to examine whether they contain identifiable narratives. Focusing on women's health, John et al. (2024) investigate misleading narratives about reproductive health across social media platforms and websites. Their study identifies 112 distinct narratives, encompassing issues such as contraception, abortion, and fertility.

Langguth et al. (2023) study stances toward 12 prevalent COVID-19 narratives on Twitter, emphasizing the role of social media in shaping public opinion. Pogorelov et al. (2021) delve into conspiracy theories, focusing on COVID-19 related 5G claims and evaluating whether tweets expressed

[1]https://doi.org/10.5281/zenodo.15228071
[2]https://doi.org/10.5281/zenodo.15228283
[3]https://github.com/GateNLP/UKElectionNarratives

Dataset	Source	# Instances	# Narratives	Topic	Lang
CARDS (Coan et al. 2021)	Articles	28,945	5/27	Climate Change	En
Climate (Piskorski et al. 2022)	Articles	1,118	5/27	Climate Change	En
COVID-19 (Kotseva et al. 2023)	Articles	58,625	12/51	COVID-19	En
COVID-19-Ger (Heinrich et al. 2024)	Telegram	1,099	14	COVID-19	De
DIPROMATS (Fraile-Hernández, Peñas, and Moral 2024)	Twitter	1,272/1,272	4/24	Narratives around COVID-19 by Authorities	En/Es
TweetIntent@Crisis (Ai et al. 2024)	Twitter	3,691	2	Russia-Ukraine Crisis	En
Climate_ads (Rowlands et al. 2024)	Facebook	1,330	7	Climate Change in US ads	En
UKElectionNarratives	Twitter	2,000	10/32	UK Elections 2019/2024	En

Table 1: Comparison between **UKElectionNarratives** and existing *public* datasets for detecting misleading narratives.

relevant allegations. Furthermore, Introne et al. (2020) explore anti-vaccination narratives in online discussion forums, offering insights into their evolution and influence. A comprehensive codebook for COVID-19 narratives for a period of three years was proposed by Kotseva et al. (2023).

Several studies focus on narratives about climate change. Zhou et al. (2024) explore the use of LLMs to extract and analyse the underlying narrative structures in English and Chinese news articles related to climate change. Additionally, Coan et al. (2021) introduce a multi-level taxonomy designed to categorise and understand climate change narratives.

In the political domain, Amanatullah et al. (2023) identify pro-Kremlin narratives related to the Russian-Ukraine war. Moral and Marco (2023) and Moral (2024) adopt an unsupervised approach to identify strategic narratives during the COVID-19 pandemic, disseminated by Chinese and Russian official government and diplomatic accounts respectively. Conversely, Fraile-Hernández, Peñas, and Moral (2024) argue that unsupervised approaches may lead to biases and, instead, propose an annotation methodology for creating narrative classification datasets. Recently, Sosnowski et al. (2024) propose a dataset for classifying narratives as either credible or disinformation adopting a human-in-the-loop approach. They study the effectiveness of different LLMs in detecting disinformation narratives. In contrast, this paper targets narratives surrounding recent elections in Europe, with UK Elections as a case study.

Existing Narratives Datasets

As presented in Table. 1, most of the existing datasets focus on COVID-19 (Kotseva et al. 2023; Heinrich et al. 2024) or climate change narratives (Coan et al. 2021; Piskorski et al. 2022; Rowlands et al. 2024). For instance, COVID-19 narratives posted by authorities in Europe, US, China and Russia are studied by Fraile-Hernández, Peñas, and Moral (2024). In contrast, Ai et al. (2024) create a dataset of narratives about the Russian-Ukraine war.

However, to the best of our knowledge, there is no dataset to date which contains human-annotated narratives that spread during elections in Europe or, specifically, during UK General Elections. To bridge this gap, we introduce *the first unified codebook* of misleading narratives that spread during Elections in Europe, and then validate this narrative taxonomy by creating *the first dataset* with misleading narratives spread during two UK General Elections.

Data Construction

Hereafter, we describe the details about our codebook and dataset. As shown in Figure 1, the dataset was constructed by annotating a set of tweets following three main steps (1) collecting relevant tweets; (2) data filtering; and (3) human annotation.

Election-based Misleading Narratives Codebook

We developed a comprehensive codebook that systematically documents and defines the key misleading narratives that spread during elections in Europe, both at national and European Union level. Our literature-based methodology is based on thorough reviews of numerous relevant reports and scholarly publications on disinformation and election narratives surrounding the European Parliamentary Elections (Colliver et al. 2019; Sawiris et al. 2019). These reports encompass a broad spectrum of European countries, including the United Kingdom, Spain, Germany, France, Italy, Poland, Hungary, Czechia, and Slovakia.

In addition to the European elections level, we also analysed studies that focused on specific national elections, including the Swedish (Colliver et al. 2018), Italian (Alaphilippe et al. 2018), German (Applebaum et al. 2017), and French (Ferrara 2017) ones. After analysing all these primary narratives, we engaged in discussions and brainstorming sessions with a diverse team of social scientists and NLP experts, aiming to identify and refine a set of unified core narratives. This iterative process involved multiple rounds of evaluation and refinement to ensure the creation of a comprehensive and robust codebook.

Drawing on insights from political science literature, we refined and enhanced several narrative definitions in our codebook, ensuring greater alignment with theoretical frameworks to define specifically the *Pro far-right* (Rooduijn et al. 2024; Mudde 2024; Sibley 2024; Pirro 2023), *Pro far-left* (Williams and Ishiyama 2018; Rooduijn et al. 2024), *Anti-liberal* (Coman and Volintiru 2023), and *Anti-woke* (Smith et al. 2023) narratives which are classed under the *Political hate and polarisation* super-narrative.

Similar to other narrative studies (Coan et al. 2021; Kotseva et al. 2023; Fraile-Hernández, Peñas, and Moral 2024), our codebook consists of general super-narratives and more

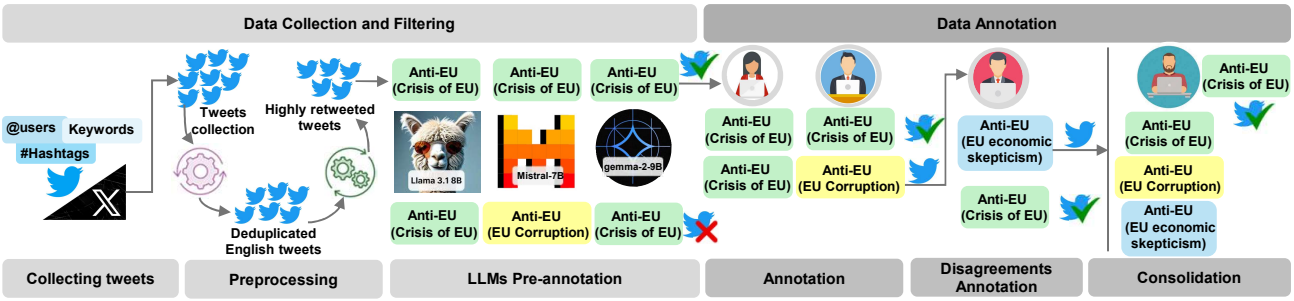

Figure 1: **UKElectionNarratives** construction approach. We first collect relevant tweets using a set of keywords, hashtags, and users. To select tweets with potential narratives, we select the highly retweeted tweets and those that 3 LLMs assigned identical narrative classes. The filtered tweets were then annotated by humans following a three-stage process.

specific narratives. For example, under the *Anti-EU* super-narrative, which covers narratives that express opposition, scepticism, or criticism towards the European Union (EU), there are four narratives namely *EU economic scepticism*, *Crisis of EU*, *EU political interference*, and *EU Corruption*.

The final codebook contains **10 super-narratives** and **32 narratives**. The full list of narratives and their descriptions is presented in the Appendix.

Tweets Collection

The dataset collection is focused specifically on narratives propagated on Twitter during the 2019 and 2024 UK General Elections, as a data-driven case study that validates the codebook. Thus, we collected tweets posted during both elections period as follows:

1. **UK General Election 2019 (UK_GE_2019):** Twitter data was collected in 2019 in real-time using the Twitter streaming API (V1.1). The data was collected using 33 hashtags (e.g., #generalelection2019, #borisjohnson) and 26 keywords (e.g., jeremy corbyn, caroline lucas). In addition we also included tweets from 18 political party accounts (e.g., @conservatives, @uklabour) where tweets posted by them or mentioning them were collected. Only tweets posted in November and December 2019 were included – the election was held on 12 December 2019 (Prosser 2021). This collected dataset contains more than 60M tweets.
2. **UK General Election 2024 (UK_GE_2024):** Historical tweets were collected from the period of May 1st, 2024 to July 4th, 2024 – the election was held on 4th July 2024 (Duggan, Milazzo, and Trumm 2024). Specifically tweets from 10 Labour user accounts (e.g., @SadiqKhan, @HackneyAbbott) and 6 Conservative user accounts (e.g., @RishiSunak, @SuellaBraverman) were collected.[4] Additionally, tweets mentioning these politicians by names are also included (e.g., Diane Abbott). This data consists of more than 500,000 tweets in total.

[4]We share all the hashtags, keywords, and user accounts used to collect our data.

Data Filtering

Given that data annotation is costly, time-consuming and ambiguous, mainly for our case with a large amount of narratives, we considered a filtering pipeline. This filtering was essential for eliminating tweets without potential misleading narratives, thereby making the process more efficient and focused for subsequent human annotation. The filtering pipeline comprises the following steps:

1. **Selecting highly retweeted tweets:** We selected the 20K top retweeted tweets from UK_GE_2019 and UK_GE_2024 (40K in total), and we excluded duplicates by keeping the tweet with the last timestamp.
2. **LLMs as pre-annotators:** Motivated by previous work that recently used LLMs for data filtering and annotation (Ai et al. 2024; Sosnowski et al. 2024), we adopted three LLMs namely gemma 2 (Team et al. 2024),[5] Llama-3.1 (Touvron et al. 2023),[6] and Mistral-7B (Jiang et al. 2023)[7] for pre-annotating the tweets into narratives, aiming to exclude tweets without potential narratives (see prompt in the Appendix). Specifically, for each tweet we prompted the LLMs to decide to which narrative class it belongs to. We then selected the tweets for which all three models agreed on their narrative labels, also excluding the tweets that were classified as not belonging to any narrative (i.e., labeled as *None* by the three LLMs). We ended up with 3,281 and 2,695 tweets from UK_GE_2019 and UK_GE_2024 collections, respectively. To evaluate the effectiveness of this approach, we prompted the LLMs to label a set of 150 tweets, all of which had achieved full agreement between two human annotators and were validated by us during our initial pilot studies. Among these, the three LLMs assigned identical labels to 49 tweets. The Cohen's kappa (Cohen 1960) between the LLMs label and the humans label is 0.70, which indicates substantial agreement.
3. **Selecting an annotation sample:** From the data filtered in the previous step, we sampled 1,000 tweets from each election by selecting the top retweeted tweets from each

[5]https://huggingface.co/google/gemma-2-9b-it
[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[7]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

narrative class. In total, we selected 2,000 tweets to be manually annotated.

Human Annotation

To annotate our data, we hired 15 students whose majors are political sciences or journalism, and we conducted the annotation process using the collaborative web-based annotation tool Teamware (Wilby et al. 2023).[8]

Due to the complexity of the task and the fine-grained nature of our narratives classes, we focused heavily on delivering clear instructions while training the annotators. To achieve this, we conducted multiple pilot studies prior to the official annotation task, that helped in identifying the main challenges and difficulties faced by annotators. As a result, our final annotation guidelines are the outcome of multiple rounds of refinement, each accompanied by internal discussions and analysis. Our pilot studies revealed that while the majority of tweets typically belong to a single narrative, a subset of tweets may encompass multiple narratives. To address this complexity, we introduced a confidence scoring system for annotators during the labeling process. Annotators were instructed to provide their confidence score for their assigned label on a scale of 1 to 5 (confidence scores descriptions are in the Appendix). If their confidence score fell below 4, they were required to assign a secondary label to the tweet. The inclusion of secondary labels and associated confidence scores have shown potential in improving the performance of some classification tasks as shown by Wu et al. (2023).

Before commencing the annotation process, all annotators attended an in-person training session. This training included a detailed explanation of the task, presentation of example tweets, and an opportunity for questions and discussion. Such preparatory training is needed to ensure annotators are capable to perform their task confidently and consistently. On average, annotators took an hour to label 50 tweets, earning an hourly wage of 17.2 GBP. To ensure the quality of our data, we designed an annotation process composed of *three stages*, similar to previous work (Da San Martino et al. 2019; Salman et al. 2023). The annotation process is as follows:

Stage 1 (annotation). We adopted a batch approach (Maarouf et al. 2023), i.e., a group of annotators annotate a batch of data to prevent exhaustion. Thus, we split our data (2,000 tweets) into 20 batches with 100 tweets each, and each batch was annotated by two annotators independently. To ensure the quality of our annotations, annotators have to pass a qualification test in order to be considered qualified for the task. The qualification test was done by injecting an extra 25 gold tweets (annotated during one of our pilot studies with full agreement of two annotators and checked by us) into the first batch annotated by each annotator. We then excluded the annotators who achieved Cohen's kappa less than 0.4 on the gold set, and consequently, we re-annotated the batches annotated by those annotators. The overall Cohen's kappa in *Stage 1* is 0.34 which indicates fair agreement. The agreement score highlights the importance of the multi-stage annotation process, aligning with similar findings observed in existing fine-grained labelled datasets (Da San Martino et al. 2019; Salman et al. 2023).

Stage 2 (annotating tweets with disagreements). A third annotator independently annotated the tweets with disagreements from *Stage 1*. We then considered the majority voting to select the final narrative class. The tweets annotated by a third annotator have an average confidence score of 4, indicating a high level of certainty. This score shows that the third annotator was generally confident, with only minor uncertainty in a few cases.

Stage 3 (consolidation). If after *Stage 2* there is still disagreement between all annotators, we asked a fourth annotator (referred to as consolidator), who is a student in political sciences, to decide a final narrative class by checking the three previously assigned labels. The average confidence score for the consolidator is 4.34, which reflects the high level of confidence on the labels for this set.

Data Overview

In this section, we provide statistics about **UKElectionNarratives** and topic analysis.

Data Statistics

The most prevalent super-narratives in **UKElectionNarratives** are *Distrust in institutions* and *Distrust in democratic system* (see Table 2). However, the most dominant class is *None* which we hypothesise to closely mirror the real-world scenario, where users comment on or question topics relevant to the elections without attempting to spread narratives. A similar pattern has been observed in datasets with stance towards rumors (Gorrell et al. 2019; Haouari and Elsayed 2024; Zheng et al. 2022). It is worth noting, that although some tweets are labeled as *None*, they may express other narratives not covered in our codebook. Additionally, similar to existing datasets with large number of classes, such as propaganda detection datasets (Da San Martino et al. 2019; Salman et al. 2023), the distribution of classes is unbalanced. Finally, whilst **UKElectionNarratives** encompasses all 32 narratives outlined in our codebook, the 2019 and 2024 elections data covers only 30 narratives and 27 narratives, respectivelly. We present example tweets from our dataset in Table 6 (in Appendix).

Topic Modeling Analysis

To analyse the key topics discussed in tweets during each of the UK elections, we conducted a topic modeling analysis using BERTopic (Grootendorst 2022). BERTopic has recently gained popularity (Hellwig et al. 2024; Zain et al. 2024) due to its ability to capture semantic relationships between words through document embeddings, which leads to the generation of more insightful topics. For tweet embeddings, we utilised the base GTE model[9] (Li et al. 2023). We then applied UMAP (McInnes et al. 2018) for dimensionality reduction, and used HDBSCAN (McInnes, Healy, and

[8]https://gatenlp.github.io/gate-teamware/development/

[9]https://huggingface.co/thenlper/gte-base

Distrust in institutions	305	(15.25%)
Failed state	64	
Criticism of national policies	241	
Distrust in democratic system	191	(9.55%)
Elections are rigged	71	
Anti-Political system	18	
Immigrants right to vote	1	
Anti-Media	101	
Political hate and polarisation	135	(6.75%)
Pro far-left	56	
Pro far-right	54	
Anti-liberal	17	
Anti-woke	8	
Anti-EU	82	(4.1%)
EU economic scepticism	19	
Crisis of EU	25	
EU political interference	33	
EU corruption	5	
Gender-related	177	(8.85%)
Language-related	7	
LGBTQ+-related	169	
Demographic narratives	1	
Religion-related	109	(5.45%)
Anti-Islam	101	
Anti-Semitic conspiracy theories	5	
Interference with states' affairs	3	
Migration-related	91	(4.55%)
Migrants societal threat	91	
Ethnicity-related	23	(1.15%)
Association to political affiliation	7	
Ethnic generalisation	4	
Ethnic offensive language	6	
Threat to population narratives	6	
Geopolitics	90	(4.5%)
Pro-Russia	5	
Foreign interference	81	
Anti-international institutions	4	
Anti-Elites	16	(0.8%)
Soros	6	
World Economic Forum / Great Reset	4	
Antisemitism	2	
Green Agenda	4	
None	781	(39.05%)
Total	2000	

Table 2: **UKElectionNarratives** statistics. **Super-narratives** are in bold.

Astels 2017) for clustering. By setting the minimum number of tweets per topic to 25, we successfully identified 9 distinct topics for each election.

As highlighted by Alammar and Grootendorst (2024), LLMs have the potential to achieve remarkable outcomes when leveraged for labeling the topics identified by BERTopic. In our study, we employed GPT-4o (Hurst et al. 2024) for generating topic labels based on the keywords and the most representative tweets extracted by BERTopic, following the official prompt guidelines recommended by the BERTopic author. In our prompt (presented in the Appendix), we set the number of representative tweets as 10.

Detected Topics Table 3 presents the topics identified for each election. Additionally, we show the most dominant super-narrative (narrative) class for each topic, illustrating their alignment with the topic labels.

For the 2019 election, we observe that the labels for topics 4, 5, 7, 8, and 9 align precisely with their most prevalent super-narrative (narrative). For topic 2, while the most prevalent super-narrative (narrative) is *Distrust in institutions (Criticism of national policies)*, at the super-narrative level alone, *Anti-EU* emerges as the most prevalent, accounting for 46.3% of the tweets associated with this topic. Differently, for topics 1, 3, and 6, the most dominant narrative class is *None*. However, examining the second most dominant class for these topics reveals alignment with the topic labels. For topic 1, the second most prevalent class is *Political hate and polarisation (Pro far-left)*. We believe this aligns with the topic label, as in the UK, discussions around left-wing populism have largely centered on Jeremy Corbyn's leadership of the Labour Party (Rios-Jara 2022). Similarly, for topic 3, the second most dominant class is *Religion-related (Anti-Islam)* which aligns with the Islamophobia topic. Lastly, for topic 6, although the second most prevalent class is *Distrust in democratic system (Anti-Media)*, a manual examination of the tweets for this topic showed that they all discuss antisemitism as shown in this tweet example "*Jews, antisemitism and Labour – a letter to the BBC about the systematic bias in their reporting of this issue and their continued pattern of accepting assertion as fact and not allowing contrary opinions a hearing opinion.*". Moreover, we found that all the tweets labeled as *Religion-related (Anti-Semitic conspiracy theories)* are within the tweets for this topic.

For the 2024 election, we observe that the labels for topics 2 and 4 align perfectly with their most dominant super-narrative (narrative). For topic 6, the most prevalent narrative is *Religion-related (Anti-Islam)*. The reason is that Sadiq Khan is London's first Muslim mayor (Zheng and de Almeida 2021) and during the 2024 election, he encountered significant Islamophobic rhetoric from certain politicians. Notably, Conservative MP Lee Anderson claimed that "Islamists" had "taken control" of Khan (Lucas 2024). Looking at the other topics where the dominant class is *None*, we observe that topics 3, 5, 7, and 8 are all relevant to *Distrust in institutions (Criticism of national policies)* narrative. Additionally, we found that for topic 3, all the tweets labelled as *Anti-Elites (Antisemitism)* and *Religion-related (Anti-Semitic conspiracy theories)* are within the tweets for this topic. Finally, for topic 9, the second most prevalent class is *Political hate and polarisation (Pro far-left)*, which is highly relevant to a topic about the Labour Party, positioned on the left of the UK political spectrum (Grant and Evans 2024).

Topic Similarity Between Both Elections To demonstrate the similarities between narratives across different

	Topic	GPT4o Topic Label (#Tweets)	Most Prevalent Super-narrative (Narrative) [%Tweets]
UK Election 2019	(1)	Jeremy Corbyn and the Labour Party (141)	None [69.5%] / Political hate and polarisation (Pro far-left) [12.1%]
	(2)	UK and EU Relations (136)	Distrust in institutions (Criticism of national policies) [23.5%]
	(3)	Islamophobia and Jihadist Attacks in Politics (119)	None [45.4%] / Religion-related (Anti-Islam) [29.4%]
	(4)	Voter ID and Election Integrity Concerns (117)	Distrust in democratic system (Elections are rigged) [41%]
	(5)	Suppression of Russian Interference Report by Boris Johnson's Government (80)	Geopolitics (Foreign interference) [75%]
	(6)	Antisemitism and Politics (61)	None [77%] / Distrust in democratic system (Anti-Media) [%9.8]
	(7)	Trans Rights and Issues (59)	Gender-related (LGBTQ+-related) [83.1%]
	(8)	NHS and Healthcare Privatization Concerns (58)	Distrust in institutions (Criticism of national policies) [51.7%]
	(9)	Immigration Policies and Impact on Public Services (55)	Migration-related (Migrants societal threat) [49.1%]
UK Election 2024	(1)	Criticism of Rishi Sunak's Leadership and Statements (130)	None [40%] / Distrust in institutions (Criticism of national policies) [27.1%]
	(2)	UK Migrant Policy and Labour's Asylum Seeker Plan (112)	Migration-related (Migrants societal threat) [46.4%]
	(3)	UK Politics, Israel Lobby, and Antisemitism Debate (96)	None [65.6%] / Distrust in institutions (Criticism of national policies) [20.8%]
	(4)	Trans Rights and Gender Ideology Debate in UK Politics (95)	Gender-related (LGBTQ+-related) [86.3%]
	(5)	Criticism of Keir Starmer and the Labour Party (95)	None [66.3%] / Distrust in institutions (Criticism of national policies) [8.4%]
	(6)	Controversy Surrounding Sadiq Khan's Mayoralty of London (90)	Religion-related (Anti-Islam) [45.6%]
	(7)	UK Economy and EU Relations Under Keir Starmer (57)	None [43.9%] / Distrust in institutions (Criticism of national policies) [24.6%]
	(8)	Diane Abbott and Her Treatment by the Labour Party (41)	None [90.2%] / Distrust in institutions (Criticism of national policies) [4.88%]
	(9)	Labour Party Call for Change in Britain (36)	None [88.88%] / Political hate and polarisation (Pro far-left) [5.56%]

Table 3: Detected topics in **UKElectionNarratives** using BERTopic and GPT4o, and the most prevalent super-narrative (narrative) for each topic based on our annotated data. We also present the second most prevalent if the first is the *None* class.

elections, we identified the overlapping topics between the two elections by leveraging their respective topic embeddings. To quantify the alignment between these topics, we calculated their cosine similarity, providing a measure of how closely related the narratives are. For visualization, we used a cosine similarity matrix, which as presented in Figure 7 (in the Appendix), maps the degree of similarity between narratives across the two elections. This approach highlights how some misleading narratives can persist across different elections such as those concerning EU relations, Immigration, Gender, Antisemitism, and Islamophobia. This also demonstrates how these issues continue to be manipulated and distorted over time, shaping political discourse.

Experimental Setup

This section presents the experimental setup to showcase the use of **UKElectionNarratives** in a narratives classification benchmark.

Narrative Detection Models

We adopted three categories of models for benchmarking:

Basic Models We considered two dummy baselines:[10] (1) *Majority Classifier*: that always predicts the most frequent class in the training data, and (2) *Random Classifier*: that randomly generates predictions with equal probability for each unique class observed in the training data.

Pre-trained Language Models (PLMs) We fine-tuned the base[11] and large[12] RoBERTa (Liu 2019) models to evaluate their performance on detecting misleading narratives. We fine-tuned both models for 5 epochs with a batch of size 16, experimenting with different learning rates [1e-5, 2e-5, 3e-5], and dropout values [0.1, 0.2, 0.3]. We then selected the best model based on Macro-F_1 on the dev set.

Large Language Models (LLMs) We adopted OpenAI GPT4o (Hurst et al. 2024), which has demonstrated high performance in detecting disinformation narratives compared to multiple LLMs (Sosnowski et al. 2024). Our experiments included both *zero-shot* and *few-shot* setups. We optimized the prompts on the development set to identify the best-performing prompts, which were then used to evaluate the test set. To ensure reliability, we executed each prompt three times and reported the average performance. We present our prompts under all setups in Appendix.

Data Splits

We applied stratified sampling (Sechidis, Tsoumakas, and Vlahavas 2011) to split the data into training (70%), development (dev) (10%), and test (20%) sets. For the narratives with 1 or 2 examples (3 narrative classes), no examples were included in the training set. The labels distribution on each set is presented in Table 5 in the Appendix.

Experimental Evaluation

GPT-4o versus RoBERTa As presented in Table 4, GPT-4o significantly outperforms both RoBERTa models across all evaluation measures. This improvement can be attributed on the one hand to GPT-4o extensive training data covering relevant topics. On the other hand, our data is unbalanced with some narratives having a small number of examples, which is challenging for PLMs. Moreover, three narrative classes are missing from the training data which poses a significant limitation as PLMs are unable to learn patterns for the missing classes.

[10]https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

[11]https://huggingface.co/FacebookAI/roberta-base

[12]https://huggingface.co/FacebookAI/roberta-large

	Development Set			Test Set		
	Macro-F_1	Macro-Rec	Macro-Prec	Macro-F_1	Macro-Rec	Macro-Prec
Majority Classifier	0.018	0.032	0.012	0.017	0.030	0.012
Random Classifier	0.019	0.018	0.032	0.021	0.073	0.039
RoBERTa-base	0.227	0.235	0.238	0.216	0.220	0.233
RoBERTa-large	0.266	0.291	0.263	0.226	0.248	0.215
GPT-4o$_{\text{(zero-shot)}}$	0.423	0.521	0.426	0.444	0.534	0.460
GPT-4o$_{\text{(zero-shot+Narrative descriptions)}}$	0.435	0.545	0.447	0.479	0.554	**0.509**
GPT-4o$_{\text{(few-shot)}}$	0.429	0.513	**0.448**	0.421	0.480	0.447
GPT-4o$_{\text{(few-shot+Narrative descriptions)}}$	**0.445**	**0.560**	0.438	**0.488**	**0.555**	0.504

Table 4: Benchmarking results on the **UKElectionNarratives** development and test sets.

GPT-4o zero-shot versus few-shot setup For the *few-shot* setup, we provided a tweet example for every narrative in the training data. As presented, in Table 4, the results show that providing tweet examples, slightly improved the performance on the dev set in terms of Macro-F_1 and precision but degraded the recall. For the test set, the *zero-shot* setup showed better performance. However, we believe more investigation exploring other prompt engineering techniques may further improve the results (Sahoo et al. 2024).

Narrative descriptions in GPT-4o prompt We included the description for each narrative, as per our codebook, in the prompt to determine whether this additional information would help the model better understand the narrative classes and distinguish between them more effectively. The results show that, for both the *zero-shot* and *few-shot* setups, including the narrative descriptions enhanced the model's performance compared to simply listing the narrative classes. Finally, combining the narrative descriptions and tweets examples resulted in the best overall performance in both dev and test sets. We argue that, when considering the trade-off between cost and effectiveness, the *zero-shot+Narrative descriptions* setup can be sufficient, as the improvement in performance is relatively modest compared to the cost of including tweet examples (30 tweet examples in our setup).

Potential Use Cases

In this section, we outline a few potential use cases for our *codebook* and *dataset*, illustrating how they can be leveraged for various tasks and challenges.

Constructing novel datasets of misleading narratives spread during elections in Europe Our multi-level codebook can aid in annotating new datasets for detecting misleading narratives that are propagated during elections in European countries other than the UK. Furthermore, the codebook can be easily adapted by extending or replacing certain narratives to better suit the context of a specific country or to address emerging misleading narratives in future elections. This flexibility ensures that the codebook remains relevant and effective in representing new narratives as they arise.

Enabling automatic annotation Despite the limited number of tweets in **UKElectionNarratives**, it has the potential to significantly enhance the process of automatic labelling for a much larger volume of tweets, as done by Ai et al. (2024). By leveraging our dataset, models that are capable of efficiently labelling additional tweets with high accuracy can be developed. This approach not only simplifies the annotation process but also expands the scope of analysis. Domain adaptation techniques can also be explored to generalise models developed with our dataset to elections in other countries and even to other domains (e.g. health-related misinformation).

Misinformation analysis Given the overwhelming volume of information that circulated about the UK General Elections, numerous misleading narratives rapidly spread and gained significant attention from the public (Vaccari, Chadwick, and Kaiser 2023; Gaber and Fisher 2022). The accelerated distribution of misinformation poses a serious challenge, as it can quickly shape public perceptions. In turn, this may result in negative outcomes, including the loss of trust in political institutions, the spread of confusion among voters, and potentially harmful effects on the democratic process itself. Our dataset can play a crucial role in supporting the study of misinformation detection and verification during this period, providing insights for countering misleading narratives in future events.

Analysing election discourse for research and media literacy Elections spark widespread discussions on social media, where individuals express opinions on policies, candidates, and political events. These conversations, often driven by strong emotions, provide valuable data for analysing public sentiment (Rita, António, and Afonso 2023), political stance (Müller, Riedl, and Drews 2022), hate speech (Agarwal et al. 2021), and other forms of offensive or harmful language (Weissenbacher and Kruschwitz 2024). The social media discussions in **UKElectionNarratives** can support research on political communication and digital discourse, enabling scholars to study narrative patterns, assess online engagement, and develop tools that promote media literacy, fostering informed public discourse.

Limitations

This study presents a codebook for the main narratives surrounding the European Elections; however, it may not cover certain country-specific narratives. Therefore, the codebook

should be expanded to accommodate the context of specific European countries or incorporate new emerging misleading narratives in the future.

One limitation of our dataset is the potential for annotator bias. Despite efforts to ensure consistency through guidelines and training, individual annotators may interpret narratives subjectively, leading to variations in classification. Additionally, due to time and cost constraints, the third stage of annotation involved a single consolidator for each tweet. A better approach would be having two consolidators to discuss and decide a final narrative class. We also acknowledge that the UK Election domain is limited and will not reflect all narratives that appear in other European countries. Furthermore, the size of the dataset presented herein is relatively small and certain narratives have very few examples.

In terms of experiments, further exploration is required to evaluate additional LLMs and diverse prompt engineering techniques. Moreover, techniques to mitigate the class imbalance should be investigated, alongside data augmentation methods to improve the performance of the models.

Conclusion and Future Work

In this paper, we introduced a multi-level codebook with a taxonomy of the main misleading narratives spread during elections in Europe. Adopting this taxonomy, we created **UKElectionNarratives**, the first dataset annotated with misleading narratives spread during the 2019 and 2024 UK General Elections. We performed a topic modelling analysis on this data, and demonstrated how topics align with annotated narratives. Moreover, we presented benchmarking results on our dataset, assessing PLMs and LLMs performance. We found that GPT-4o significantly outperforms RoBERTa-based models. However, further exploration is needed to assess additional LLMs and experiment with various prompt engineering techniques. Additionally, approaches to mitigate class imbalance should be examined, in conjunction with data augmentation methods, to enhance the models performance. As a future work, we plan to augment our dataset adopting different techniques including a human-in-the-loop approach, automatic data annotation, and generating synthetic data using our proposed codebook and novel dataset. Moreover, we aim to propose more robust models tailored for detecting misleading narratives.

Acknowledgements

This work is supported by the UK's innovation agency (InnovateUK) grant number 10039039 (approved under the Horizon Europe Programme as VIGILANT, EU grant agreement number 101073921) (https://www.vigilantproject.eu).

References

Agarwal, P.; Hawkins, O.; Amaxopoulou, M.; Dempsey, N.; Sastry, N.; and Wood, E. 2021. Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, 5–16. Association for Computing Machinery.

Ai, L.; Gupta, S.; Oak, S.; Hui, Z.; Liu, Z.; and Hirschberg, J. 2024. TweetIntent@Crisis: A Dataset Revealing Narratives of Both Sides in the Russia-Ukraine Crisis. In *ICWSM2024*.

Alammar, J.; and Grootendorst, M. 2024. *Hands-On Large Language Models: Language Understanding and Generation*. " O'Reilly Media, Inc.".

Alaphilippe, A.; Ceccarelli, C.; Charlet, L.; and Mycielski, M. 2018. Developing a disinformation detection system and sourcing it live–The case study of the 2018 Italian elections. *EU DisinfoLab*, 16.

Amanatullah, S.; Balani, S.; Fraioli, A.; McVicker, S. M.; and Gordon, M. 2023. Tell Us How You Really Feel: Analyzing Pro-Kremlin Propaganda Devices & Narratives to Identify Sentiment Implications. *The Propwatch Project. Illiberalism Studies Program Working Paper.*

Applebaum, A.; Pomerantsev, P.; Smith, M.; and Colliver, C. 2017. Make Germany great again. Kremlin, alt-right and international influences in the 2017 German elections. *London School of Economics Report*.

Coan, T. G.; Boussalis, C.; Cook, J.; and Nanko, M. O. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1): 22320.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1).

Colliver, C.; Guhl, J.; Guerin, C.; Ebner, J.; and Tuck, H. 2019. Click Here For Outrage: Disinformation in the European Parliamentary Elections 2019. *Institute for Strategic Dialogue*, 32.

Colliver, C.; Pomerantsev, P.; Applebaum, A.; and Birdwell, J. 2018. Smearing Sweden: International influence campaigns in the 2018 Swedish election. *LSE Institute of Global Affairs. Retrieved April*, 14: 2020.

Coman, R.; and Volintiru, C. 2023. Anti-liberal ideas and institutional change in Central and Eastern Europe. *European Politics and Society*, 24(1): 5–21.

Da San Martino, G.; Seunghak, Y.; Barrón-Cedeno, A.; Petrov, R.; Nakov, P.; et al. 2019. Fine-grained analysis of propaganda in news article. In *EMNLP-IJCNLP 2019*.

Duggan, A.; Milazzo, C.; and Trumm, S. 2024. Local Leaflets: Constituency Issue Messaging at the 2024 General Election. *The Political Quarterly*.

Ferrara, E. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-01-15.

Fraile-Hernández, J. M.; Peñas, A.; and Moral, P. 2024. Automatic Identification of Narratives: Evaluation framework, annotation methodology and dataset creation. *IEEE Access*.

Gaber, I.; and Fisher, C. 2022. "Strategic lying": The case of Brexit and the 2019 UK election. *The International Journal of Press/Politics*, 27(2): 460–477.

Ganti, A.; Hussein, E. A. H.; Wilson, S.; Ma, Z.; and Zhao, X. 2023. Narrative Style and the Spread of Health Misinformation on Twitter. In Bouamor, H.; Pino, J.; and Bali,

K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4266–4282.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gorrell, G.; Kochkina, E.; Liakata, M.; Aker, A.; Zubiaga, A.; Bontcheva, K.; and Derczynski, L. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In May, J.; Shutova, E.; Herbelot, A.; Zhu, X.; Apidianaki, M.; and Mohammad, S. M., eds., *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Grant, Z.; and Evans, G. 2024. A New Dilemma of Social Democracy? The British Labour Party, the White Working Class and Ethnic Minority Representation. *British Journal of Political Science*, 54(3): 793–815.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Haouari, F.; and Elsayed, T. 2024. Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter. *Social Network Analysis and Mining*, 14(1): 34.

Heinrich, P.; Blombach, A.; Doan Dang, B. M.; Zilio, L.; Havenstein, L.; Dykes, N.; Evert, S.; and Schäfer, F. 2024. Automatic Identification of COVID-19-Related Conspiracy Narratives in German Telegram Channels and Chats. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Hellwig, N. C.; Fehle, J.; Bink, M.; Schmidt, T.; and Wolff, C. 2024. Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election. *International Journal of Speech Technology*, 27(4): 901–921.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Introne, J.; Korsunska, A.; Krsova, L.; and Zhang, Z. 2020. Mapping the narrative ecosystem of conspiracy theories in online anti-vaccination discussions. In *International Conference on Social Media and Society*, 184–192.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

John, J. N.; Gorman, S.; Scales, D.; and Gorman, J. 2024. Online Misleading Information About Women's Reproductive Health: A Narrative Review. *Journal of General Internal Medicine*, 1–9.

Kotseva, B.; Vianini, I.; Nikolaidis, N.; Faggiani, N.; Potapova, K.; Gasparro, C.; Steiner, Y.; Scornavacche, J.; Jacquet, G.; Dragu, V.; et al. 2023. Trend analysis of COVID-19 mis/disinformation narratives–A 3-year study. *Plos one*, 18(11): e0291423.

Langguth, J.; Schroeder, D. T.; Filkuková, P.; Brenner, S.; Phillips, J.; and Pogorelov, K. 2023. COCO: an annotated Twitter dataset of COVID-19 conspiracy theories. *Journal of Computational Social Science*, 6(2): 443–484.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Lucas, S. 2024. Liz Truss was in Maryland with an explanation for her historically-short, 50-day tenure as UK Prime Minister. *Centre for Brexit Studies Blog*.

Maarouf, A.; Bär, D.; Geissler, D.; and Feuerriegel, S. 2023. HQP: a human-annotated dataset for detecting online propaganda. *arXiv preprint arXiv:2304.14931*.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11): 205.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Moral, P. 2024. A tale of heroes and villains: Russia's strategic narratives on Twitter during the COVID-19 pandemic. *Journal of Information Technology & Politics*.

Moral, P.; and Marco, G. 2023. Assembling stories tweet by tweet: strategic narratives from chinese authorities on Twitter during the covid-19 pandemic. *Communication Research and Practice*, 9(2): 159–183.

Mudde, C. 2024. The 2024 EU Elections: The Far Right at the Polls. *Journal of Democracy*, 35(4): 121–134.

Müller, A.; Riedl, J.; and Drews, W. 2022. Real-time stance detection and issue analysis of the 2021 german federal election campaign on twitter. In *International Conference on Electronic Government*, 125–146. Springer.

Panizio, E. 2024. Disinformation narratives during the 2023 elections in Europe. *EDMO*, 34.

Pirro, A. L. 2023. Far right: The significance of an umbrella concept. *Nations and Nationalism*, 29(1): 101–112.

Piskorski, J.; Nikolaidis, N.; Stefanovitch, N.; Kotseva, B.; Vianini, I.; Kharazi, S.; Linge, J. P.; et al. 2022. Exploring Data Augmentation for Classification of Climate Change Denial: Preliminary Study. In *Text2Story@ ECIR*, 97–109.

Pogorelov, K.; Schroeder, D. T.; Filkuková, P.; Brenner, S.; and Langguth, J. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *OASIS 2021*.

Prosser, C. 2021. The end of the EU affair: the UK general election of 2019. *West European Politics*, 44(2): 450–461.

Rios-Jara, H. 2022. Between Movements and the Party: Corbynism and the Limits of Left-Wing Populism in the UK. *Populism, Protest, New Forms of Political Organisation*, 130–49.

Rita, P.; António, N.; and Afonso, A. P. 2023. Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period. *Social Network Analysis and Mining*, 13(1): 46.

Rooduijn, M.; Pirro, A. L.; Halikiopoulou, D.; Froio, C.; Van Kessel, S.; De Lange, S. L.; Mudde, C.; and Taggart, P. 2024. The PopuList: A database of populist, far-left, and far-right parties using expert-informed qualitative comparative classification (EiQCC). *British Journal of Political Science*, 54(3): 969–978.

Rowlands, H.; Morio, G.; Tanner, D.; and Manning, C. 2024. Predicting Narratives of Climate Obstruction in Social Media Advertising. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *ACL 2024*.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Salman, M.; Hanif, A.; Shehata, S.; and Nakov, P. 2023. Detecting Propaganda Techniques in Code-Switched Social Media Text. In *EMNLP 2023*.

Sawiris, M.; Dušková, L.; Syrovátka, J.; Győri, L.; and Wierzejski, A. 2019. EUROPEAN ELECTIONS IN THE V4 From disinformation campaigns to narrative amplification. *GLOBSEC*, 30.

Sechidis, K.; Tsoumakas, G.; and Vlahavas, I. 2011. On the stratification of multi-label data. In *ECML PKDD 2011*.

Sibley, A. 2024. Behind the British New Far-Right's veil: Do individuals adopt strategic liberalism to appear more moderate or are they semi-liberal? *The British Journal of Politics and International Relations*.

Smith, D. S.; Boag, L.; Keegan, C.; and Butler-Warke, A. 2023. Land of woke and glory? The conceptualisation and framing of "wokeness" in UK media and public discourses. *Javnost-The Public*, 30(4): 513–533.

Sosnowski, W.; Modzelewski, A.; Skorupska, K.; Otterbacher, J.; and Wierzbicki, A. 2024. EU DisinfoTest: a Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives. In *EMNLP 2024*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaccari, C.; Chadwick, A.; and Kaiser, J. 2023. The campaign disinformation divide: believing and sharing news in the 2019 UK General election. *Political Communication*, 40(1): 4–23.

Weissenbacher, M.; and Kruschwitz, U. 2024. Analyzing Offensive Language and Hate Speech in Political Discourse: A Case Study of German Politicians. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*.

Wilby, D.; Karmakharm, T.; Roberts, I.; Song, X.; and Bontcheva, K. 2023. GATE Teamware 2: An open-source tool for collaborative document classification annotation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 145–151.

Williams, C.; and Ishiyama, J. 2018. Responding to the left: the effect of far-left parties on mainstream party Euroskepticism. *Journal of Elections, Public Opinion and Parties*.

Wu, B.; Li, Y.; Mu, Y.; Scarton, C.; Bontcheva, K.; and Song, X. 2023. Don't waste a single annotation: improving single-label classifiers through soft labels. In Bouamor, H.; Pino, J.; and Bali, K., eds., *EMNLP 2023*.

Zain, R. M.; Anggai, S.; Musyafa, A.; Waskita, A.; et al. 2024. Revealing a Country's Government Discourse Through BERT-based Topic Modeling in the US Presidential Speeches. In *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*.

Zheng, J.; Baheti, A.; Naous, T.; Xu, W.; and Ritter, A. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation. In *EMNLP 2022*.

Zheng, Y.; and de Almeida, M. 2021. Islamophobia in Western Media: A Case Study. *Academia Letters*, 3147.

Zhou, H.; Hobson, D.; Ruths, D.; and Piper, A. 2024. Large Scale Narrative Messaging around Climate Change: A Cross-Cultural Comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, 143–155.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our dataset and codebook provide unique resources for the study of misinformation narratives in (European) elections. By creating a codebook grounded on the literature, we believe to have minimised biases inherently present in data-driven approaches. The datasest complies with the ethics standards of the University of Sheffield Research and Ethics (UREC) guidelines[13] (in particular, Ethics note number 14),[14] making available only anonymised data and focusing on general misinformation narratives, instead of specific profiles. By collecting data focusing only on generally used hashtags and public figures, we also aimed to avoid including biased and/or personal data about individuals.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract and introduction contain an accurate description of the contributions and scope of our paper.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. The

[13] https://www.sheffield.ac.uk/rpi/ethics-integrity

[14] https://www.sheffield.ac.uk/media/29459/download?attachment

codebook is created by focusing on a extensive literature review, basing the narrative definitions in such previous work. The dataset is created using a well-established framework for social media data collection, focusing on keywords (hashtags) and public figures monitoring. Data annotation is conducted by using a robust three-stage methodology, aiming to remove biases and solve disagreements. The methodology for benchmarking the dataset also follows the state-of-the-art for the field.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, because the dataset does not contain population-specific features (apart from generally being only UK centred). See our Ethics Statement for more details.

(e) Did you describe the limitations of your work? Yes, a limitations section is explicitly added in the paper.

(f) Did you discuss any potential negative societal impacts of your work? Yes. Our Ethics Statement presents a discussion about our analysis being only on aggregated data, completely avoiding drawing conclusions about individuals.

(g) Did you discuss any potential misuse of your work? Yes. The Limitations section discuss the current limitations of our work highlighting cases of potential misuse.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. Our data sharing process is restricted to only sharing tweet IDs and the dataset license only allows research work and requests that modifications to this dataset to be released under the same license. Our code is also released to ensure reproducibility.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes. We have read the guidelines and ensured our paper conforms to them.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. We share both code and prompts used in our experiments. Data is shared as described in the Ethics Statement (under a share-alike CC license for research purposes only).

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. Model details are presented in the paper and available in the GitHub repository.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes. We describe the resources used in the Appendix. However, it is worth noting that for GPT-4o we used the OpenAI API, which does not explicitly describe the hardwared used.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. Our benchmarking follows established evaluation metrics from previous work.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, because we do not perform an error analysis which is beyond the scope of this paper. However, we provide details in the Limitations section that can be used to mitigate misclassification in future work.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? Yes. The license of our dataset is CC-BY-NC-SA.

(c) Did you include any new assets in the supplemental material or as a URL? Yes. A sample of our dataset is added as supplementary material and the codebook is added as part of the Appendix.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. Our annotators consented to the use of their annotations in our research. Although it is not possible to get consent from social media users, we follow the

GDPR and UREC guidelines for the use of social media posts.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. In our ethics statement we mention potential harms to annotators regarding the nature of social media data. However, as previous work have identified, toxic content and hate speech are not as frequent as other content, which reduces the chances of exposure to such data.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes. We made our dataset available through Zenodo and provided the details of its license and permitted usage in the Ethics Statement.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No. However, we provide all details regarding data collection, annotation and release in the paper (therefore, the information that should appear in the datasheet are already detailed in the paper). We also clearly define the motivation behind data creation and overall research topic.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes. We added the instructions given to annotators in the Appendix.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes. We follow the UREC guidelines and clearly inform participants of potential risks in participating in an annotation task involving social media (e.g. the exposure to toxic content). Participants were made aware of these risks and were given time to read the details of consent form.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. This is described in the paper under the Human Annotation subsection.

(d) Did you discuss how data is stored, shared, and deidentified? Yes. All data is securely stored in our encrypted servers. The data shared with annotators is done via our Teamware tool, that also stores all data in our encrypted servers. Annotators data (e.g. e-mail) are stored separately from the annotations and researchers only have access to annotators' IDs and annotators (i.e. no access to personal information is available). The dataset is made available for research without any identification to the annotators. The data from social media is also processed and annotated following the UREC guidelines as previously stated.

Ethics Statement

Data Collection Our data collection complies with social platform T&Cs and the UK Data Protection Act.[15]

Data Annotation The annotation task has been approved by the University Research and Ethics Committee (UREC) with reference number 059166.[16] Annotators were recruited through an internal process facilitated by the University. They were informed of potential harms related to their participation and given training for data annotation. They were informed that they could leave the task at any time, without any negative consequences to them and they also signed a consent form. They were informed that the data could contain toxic language and hate speech and they were allowed to skip tweets if they do not feel comfortable. Our approach involving three stages also mitigate annotators' biases regarding subjective aspects (e.g. political and cultural views).

Data Release Adhering to X T&Cs,[17] we only release tweet IDs together with their respective annotations under a CC-BY-NC-SA 4.0 license.[18] The dataset is available in the following link: https://doi.org/10.5281/zenodo.15228283

Data Analysis Although we use anonymised data in our experiments, we are aware of the potential risks of using social media data and profiling. Therefore, our analysis are always about aggregated results and we do not have or provide access to demographic information.

Appendix

Narrative Descriptions in Our Codebook

1. **Anti-EU (EU economic scepticism):** Narratives criticising EU's economic policies and how they affect local economies.
2. **Anti-EU (Crisis of EU):** Narratives expressing scepticism or concern about the EU, focusing on alleged broader issues affecting the union's legitimacy and effectiveness.
3. **Anti-EU (EU political interference):** Narratives alleging that the EU is interfering or manipulating local politics and specific policy areas, including food, health, environmental, and migration policies, sparking controversy and debate over EU influence on national sovereignty.
4. **Anti-EU (EU Corruption):** Narratives claiming corruption within the European Union and its institutions.
5. **Political hate and polarisation (Pro far-left):** Refers to the manipulative narratives that fundamentally reject the existing socio-economic structure of contemporary capitalism. These narratives advocate for an alternative economic and power structure based on principles of economic justice and social equality. They view economic inequality as a core issue within the current political and

[15] https://www.gov.uk/data-protection

[16] https://www.sheffield.ac.uk/rpi/ethics-integrity

[17] https://developer.x.com/en/more/developer-terms/agreement-and-policy

[18] https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en

social arrangements and propose a major redistribution of resources from existing political elites. Additionally, they support a state role in controlling the economy. These narratives target both "radical left" who accepts democracy but favors "workers' democracy" and the direct participation of labor in the management of the economy, and the "extreme left" emphasizes both parliamentary, as well as the extra-parliamentary struggle against globalized capitalism, and sees no place for free-market enterprise (Williams and Ishiyama 2018; Rooduijn et al. 2024).

6. **Political hate and polarisation (Pro far-right):** Refers to the manipulative narratives that exploit concerns about immigration, national security, and the loss of control over domestic affairs to prioritize the interests of native citizens. These narratives advocate for strict societal authority and order, often framing these issues in a way that appeals to fear and insecurity, thereby promoting divisive and exclusionary policies (Rooduijn et al. 2024; Mudde 2024; Sibley 2024; Pirro 2023).
7. **Political hate and polarisation (Anti-liberal):** Refers to the manipulative narratives that reject key principles of liberalism—political, economic, and cultural—and seek to reshape politics and societal norms. These narratives blame political and economic crises on liberal policies, particularly viewing the process of EU integration as a source of domestic issues. They advocate for a political community that promotes a narrow, shared cultural and moral identity, often framing these issues in a way that appeals to fear and exclusionary instincts (Coman and Volintiru 2023).
8. **Political hate and polarisation (Anti-woke):** Narratives that frame efforts to address social inequalities—such as racism, sexism, and anti-LGBTQ+ discrimination—as exaggerated, illegitimate, or a threat to societal norms. These narratives often shift focus away from marginalized groups and instead position privileged groups, such as wealthy individuals, as victims of "wokeness" or cultural shifts. They emphasize cultural divisions, dismiss systemic inequalities as non-issues, and portray initiatives promoting minority rights or inclusion as oppressive, unnecessary, or harmful (Smith et al. 2023).
9. **Religion-related (Anti-Islam):** Criticising Islam and portraying it as a threat to national security and cultural identity. Also, claiming that Muslim communities pose a risk to public safety and that the spread of Islam in Europe could lead to the destruction of traditional Western values and cultural heritage.
10. **Religion-related (Anti-Semitic conspiracy theories):** Narratives promoting conspiracy theories that allege Jewish people or organisations are responsible for major global events or issues.
11. **Religion-related (Interference with states' affairs):** Narratives of foreign religious authorities interfering in a country's internal political processes.
12. **Gender-related (Language-related):** Narratives that claim certain gender terminologies or vocabularies are being imposed on individuals or communities. Controversy over language use and inclusivity.
13. **Gender-related (LGBTQ+-related):** Narratives related to the LGBTQ+ community. Those include discussions around issues of acceptance, rights, discrimination, etc.
14. **Gender-related (Demographic narratives):** Narratives that argue against masturbation or the 'gender propaganda' and claiming that they promote infertility and population decline.
15. **Ethnicity-related (Association to political affiliation):** Narratives that associate politicians with specific ethnic groups and label them in negative ways. Elections are viewed through the lens of a competition between different ethnic groups.
16. **Ethnicity-related (Ethnic generalisation):** Narratives that generalise entire ethnic groups, treating them as homogeneous 'others'.
17. **Ethnicity-related (Ethnic offensive language):** Narratives that use offensive language to describe or address individuals based on their membership in certain ethnic groups.
18. **Ethnicity-related (Threat to population narratives):** Narratives claiming that the native population is being replaced by immigrants of different ethnicities.
19. **Migration-related (Migrants societal threat):** Narratives that portray migrants as a societal threat, emphasising issues such as violence, terrorism, inequality, and crime.
20. **Distrust in institutions (Failed state):** The country is in a state of demise/collapse. Focusing on issues such as immigration, police, etc.
21. **Distrust in institutions (Criticism of national policies):** Criticising national policy positions on contentious issues like pension, military conscription, and wealth disparity.
22. **Distrust in democratic system (Elections are rigged):** Narratives alleging election manipulation through various means, including political interference, exclusion, anti-vote measures, lack of transparency, and distrust in electoral institutions or polling agencies.
23. **Distrust in democratic system (Anti-Political system):** Narratives critical of the political system, asserting that political parties are neither beneficial nor necessary for a well-functioning country.
24. **Distrust in democratic system (Anti-Media):** Narratives critical of the media, including attacks on journalists, allegations of corruption or complicity in undermining the country, lack of balance and impartiality, and claims of favoritism towards a particular political faction.
25. **Distrust in democratic system (Immigrants right to vote):** Narratives surrounding immigrants' right to vote that claim their participation biases election outcomes or that they are deliberately excluded from the process, with potential impacts on election fairness and representation.
26. **Geopolitics (Pro-Russia):** Narratives that show Russia as strong country that is the victim of the west.

27. **Geopolitics (Foreign interference):** Narratives claiming that foreign actors are meddling in the affairs of EU countries.
28. **Geopolitics (Anti-international institutions):** Narratives that oppose or criticise international organisations such as NATO.
29. **Anti-Elites (Soros):** Narratives targeting George Soros, accusing him of manipulating political processes.
30. **Anti-Elites (World Economic Forum / Great Reset):** Narratives claiming that the World Economic forum has/seeks control over society.
31. **Anti-Elites (Antisemitism):** Narratives accusing Jews or Jewish organisations of having disproportionate power and influence over society and politics.
32. **Anti-Elites (Green Agenda):** Narratives that criticise the "Green Agenda" and green activism.

Annotation Methodology

Screenshots from the Annotation Framework

We present in this section some screenshots from our annotation framework in Figure 2, 3, 4, 5, and Figure 6. It is worth noting, that the annotators attended a training session besides being given the annotation guidelines. These session aimed to familiarise them with the task and included the presentation of example tweets for better understanding.

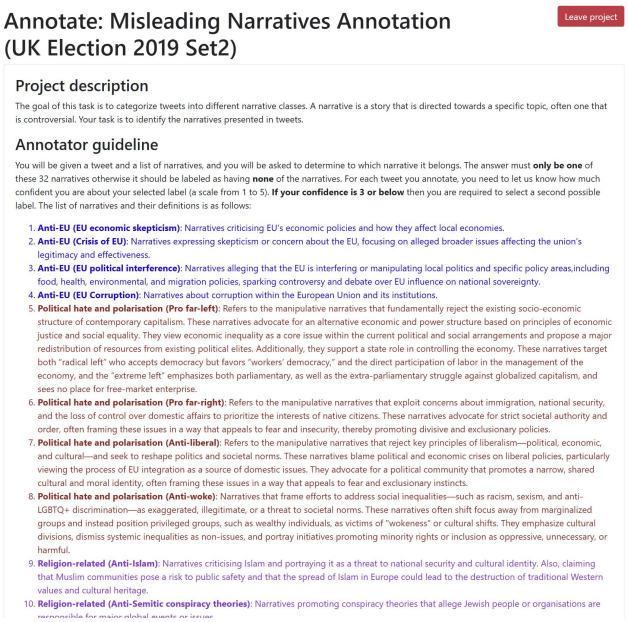

Figure 2: Annotation guidelines.

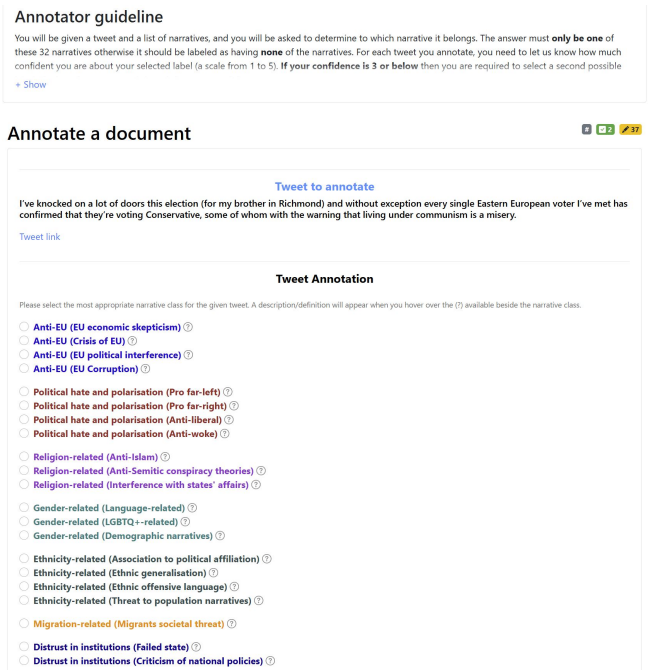

Figure 3: Selecting the first choice during annotation.

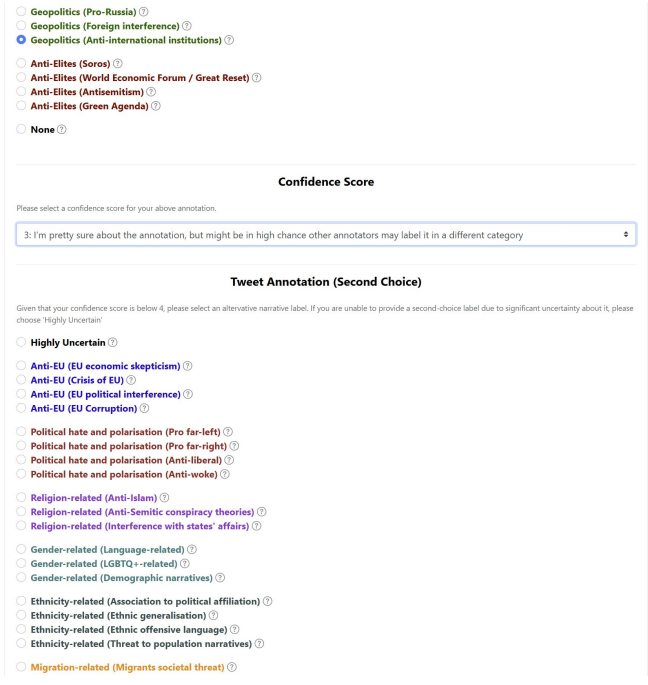

Figure 4: Selecting the second choice during annotation.

Annotator Confidence Scores

The following confidence scores descriptions were presented to the annotators:

- **Score 1:** I'm really unsure about the annotation. It may belong to another category as well, you may wish to discard this instance from the training.
- **Score 2:** I'm not sure about the annotation, it seems it also belongs to other categories.
- **Score 3:** I'm pretty sure about the annotation, but might be in high chance other annotators may label it in a different category
- **Score 4:** I'm confident about the annotation, but might

be in small chance other annotators may label it in a different category.

- **Score 5:** I'm certain about the annotation without a doubt.

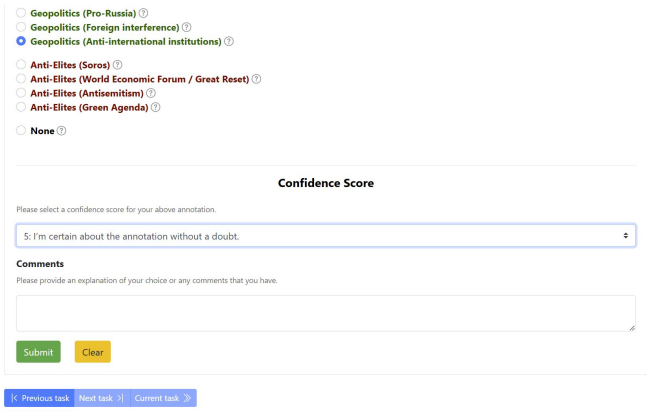

Figure 5: A second choice was not required if the confidence score is above 3.

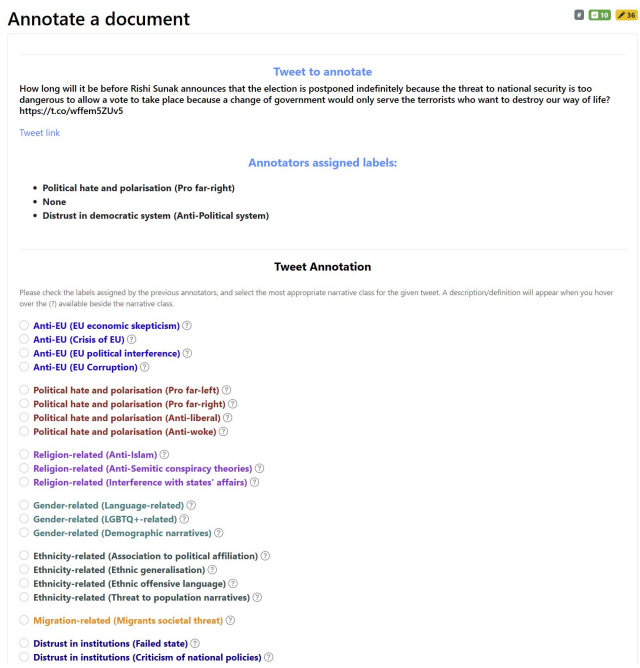

Figure 6: Stage 3 of the annotation process (consolidation).

Prompt Template for the pre-annotation annotation task using LLMs

```
You are a social media moderator that
will determine the narratives in tweets
posted by users.
You will be given a single tweet and
a list of narratives, and you will be
asked to determine to which narrative
it belongs.
The answer must only be one of these
narratives [Anti-EU (EU economic
scepticism), Anti-EU (Crisis of EU),
Anti-EU (EU political interference),
Anti-EU (EU Corruption), Political
hate and polarisation (Pro far-left),
Political hate and polarisation
(Pro far-right), Political hate and
polarisation (Anti-liberal), Political
hate and polarisation (Anti-woke),
Religion-related (Anti-Islam),
Religion-related (Anti-Semitic
conspiracy theories), Religion-related
(Interference with states' affairs),
Gender-related (Language-related),
Gender-related (LGBTQ+-related),
Gender-related (Demographic
narratives), Ethnicity-related
(Association to political
affiliation), Ethnicity-related (Ethnic
generalisation), Ethnicity-related
(Ethnic offensive language),
Ethnicity-related (Threat to population
narratives), Migration-related
(Migrants societal threat), Distrust
in institutions (Failed state),
Distrust in institutions (Criticism
of national policies), Distrust in
democratic system (Elections are
rigged)', Distrust in democratic
system (Anti-Political system),
Distrust in democratic system
(Anti-Media), Distrust in democratic
system (Immigrants right to vote),
Geopolitics (Pro-Russia), Geopolitics
(Foreign interference), Geopolitics
(Anti-international institutions),
Anti-Elites (Soros), Anti-Elites
(World Economic Forum / Great Reset),
Anti-Elites (Antisemitism), Anti-Elites
(Green Agenda), None]
You should not provide any explanation
or justification.
Question: Can you detect the narrative
in this tweet given the list of
narratives presented to you?
Tweet: ''Tweet''
```

Prompt Template for Topic Labelling (Topic Modelling)

```
I have a topic that contains the
following documents: [DOCUMENTS]
The topic is described by the following
keywords: [KEYWORDS]
Based on the information above, extract
a short topic label in the following
```

format: topic: ``short topic label''

Prompt Template for Narrative Detection

Prompt Template for Narrative Detection (zero shot)

You are a content moderator who will monitor if there are any misleading narratives in the tweets posted by users.
You will be given a single tweet and a list of narratives, and your job is to read the tweet and determine if it is expressing any one of the narratives listed to you.
The answer must only be one of these narratives [Anti-EU (EU economic scepticism), Anti-EU (Crisis of EU), Anti-EU (EU political interference), Anti-EU (EU Corruption), Political hate and polarisation (Pro far-left), Political hate and polarisation (Pro far-right), Political hate and polarisation (Anti-liberal), Political hate and polarisation (Anti-woke), Religion-related (Anti-Islam), Religion-related (Anti-Semitic conspiracy theories), Religion-related (Interference with states' affairs), Gender-related (Language-related), Gender-related (LGBTQ+-related), Gender-related (Demographic narratives), Ethnicity-related (Association to political affiliation), Ethnicity-related (Ethnic generalisation), Ethnicity-related (Ethnic offensive language), Ethnicity-related (Threat to population narratives), Migration-related (Migrants societal threat), Distrust in institutions (Failed state), Distrust in institutions (Criticism of national policies), Distrust in democratic system (Elections are rigged)', Distrust in democratic system (Anti-Political system), Distrust in democratic system (Anti-Media), Distrust in democratic system (Immigrants right to vote), Geopolitics (Pro-Russia), Geopolitics (Foreign interference), Geopolitics (Anti-international institutions), Anti-Elites (Soros), Anti-Elites (World Economic Forum / Great Reset), Anti-Elites (Antisemitism), Anti-Elites (Green Agenda), None]
You should not provide any explanation or justification.
Question: Can you detect any misleading narrative in this tweet given the list of narratives presented to you?
Tweet:``Tweet''

Prompt Template for Narrative Detection (few shot)

You are a content moderator who will monitor if there are any misleading narratives in the tweets posted by users.
You will be given a single tweet and a list of narratives, and your job is to read the tweet and determine if it is expressing any one of the narratives listed to you.
The answer must only be one of these narratives [Anti-EU (EU economic scepticism), Anti-EU (Crisis of EU), Anti-EU (EU political interference), Anti-EU (EU Corruption), Political hate and polarisation (Pro far-left), Political hate and polarisation (Pro far-right), Political hate and polarisation (Anti-liberal), Political hate and polarisation (Anti-woke), Religion-related (Anti-Islam), Religion-related (Anti-Semitic conspiracy theories), Religion-related (Interference with states' affairs), Gender-related (Language-related), Gender-related (LGBTQ+-related), Gender-related (Demographic narratives), Ethnicity-related (Association to political affiliation), Ethnicity-related (Ethnic generalisation), Ethnicity-related (Ethnic offensive language), Ethnicity-related (Threat to population narratives), Migration-related (Migrants societal threat), Distrust in institutions (Failed state), Distrust in institutions (Criticism of national policies), Distrust in democratic system (Elections are rigged)', Distrust in democratic system (Anti-Political system), Distrust in democratic system (Anti-Media), Distrust in democratic system (Immigrants right to vote), Geopolitics (Pro-Russia), Geopolitics (Foreign interference), Geopolitics (Anti-international institutions), Anti-Elites (Soros), Anti-Elites (World Economic Forum / Great Reset), Anti-Elites (Antisemitism), Anti-Elites (Green Agenda), None]
Here are some example tweets and their associated narrative label to help you decide:

...
...
Tweet: Brussels chaos: Spain follows Poland in shock threat to quit EU 'No more humiliation!' The EU are finished and I'm glad to say we started the destruction of it. Europe will be a better place without it
Narrative: Anti-EU (Crisis of EU)
...
Tweet: Scary scenes on London streets. Every day, UK and London in particular, look like a city from the Middle East, with angry wild Islamists bullying and threatening people! Sick and scary! Shame on @MayorofLondon'
Narrative: Religion-related (Anti-Islam)
...
You should not provide any explanation or justification.
Question: Can you detect any misleading narrative in this tweet given the list of narratives presented to you?
Tweet:``Tweet''

Prompt Template for Narrative Detection (zero shot) with Narrative Descriptions

You are a content moderator who will monitor if there are any misleading narratives in the tweets posted by users.
You will be given a single tweet and a list of narratives and their descriptions, and your job is to read the tweet and determine if it is expressing any one of the narratives listed to you.
The answer must only be one of these narratives [Anti-EU (EU economic scepticism), Anti-EU (Crisis of EU), Anti-EU (EU political interference), Anti-EU (EU Corruption), Political hate and polarisation (Pro far-left), Political hate and polarisation (Pro far-right), Political hate and polarisation (Anti-liberal), Political hate and polarisation (Anti-woke), Religion-related (Anti-Islam), Religion-related (Anti-Semitic conspiracy theories), Religion-related (Interference with states' affairs), Gender-related (Language-related), Gender-related (LGBTQ+-related), Gender-related (Demographic narratives), Ethnicity-related (Association to political affiliation), Ethnicity-related (Ethnic generalisation), Ethnicity-related (Ethnic offensive language), Ethnicity-related (Threat to population narratives), Migration-related (Migrants societal threat), Distrust in institutions (Failed state), Distrust in institutions (Criticism of national policies), Distrust in democratic system (Elections are rigged)', Distrust in democratic system (Anti-Political system), Distrust in democratic system (Anti-Media), Distrust in democratic system (Immigrants right to vote), Geopolitics (Pro-Russia), Geopolitics (Foreign interference), Geopolitics (Anti-international institutions), Anti-Elites (Soros), Anti-Elites (World Economic Forum / Great Reset), Anti-Elites (Antisemitism), Anti-Elites (Green Agenda), None]
Please check the following narrative descriptions before making a decision:
Anti-EU (EU economic scepticism): Narratives criticising EU's economic policies and how they affect local economies.
Anti-EU (Crisis of EU): Narratives expressing scepticism or concern about the EU, focusing on alleged broader issues affecting the union's legitimacy and effectiveness.
Anti-EU (EU political interference): Narratives alleging that the EU is interfering or manipulating local politics and specific policy areas,including food, health, environmental, and migration policies, sparking controversy and debate over EU influence on national sovereignty.
...
None: None of the above narratives.
You should not provide any explanation or justification.
Question: Can you detect any misleading narrative in this tweet given the list of narratives presented to you?
Tweet:``Tweet''

Prompt Template for Narrative Detection (few shot) with Narrative Descriptions

You are a content moderator who will monitor if there are any misleading narratives in the tweets posted by users.
You will be given a single tweet and a list of narratives and their descriptions, and your job is to read the tweet and determine if it is expressing any one of the narratives

```
listed to you.
The answer must only be one of these
narratives [Anti-EU (EU economic
scepticism), Anti-EU (Crisis of EU),
Anti-EU (EU political interference),
Anti-EU (EU Corruption), Political
hate and polarisation (Pro far-left),
Political hate and polarisation
(Pro far-right), Political hate and
polarisation (Anti-liberal), Political
hate and polarisation (Anti-woke),
Religion-related (Anti-Islam),
Religion-related (Anti-Semitic
conspiracy theories), Religion-related
(Interference with states' affairs),
Gender-related (Language-related),
Gender-related (LGBTQ+-related),
Gender-related (Demographic
narratives), Ethnicity-related
(Association to political
affiliation), Ethnicity-related (Ethnic
generalisation), Ethnicity-related
(Ethnic offensive language),
Ethnicity-related (Threat to population
narratives), Migration-related
(Migrants societal threat), Distrust
in institutions (Failed state),
Distrust in institutions (Criticism
of national policies), Distrust in
democratic system (Elections are
rigged)', Distrust in democratic
system (Anti-Political system),
Distrust in democratic system
(Anti-Media), Distrust in democratic
system (Immigrants right to vote),
Geopolitics (Pro-Russia), Geopolitics
(Foreign interference), Geopolitics
(Anti-international institutions),
Anti-Elites (Soros), Anti-Elites
(World Economic Forum / Great Reset),
Anti-Elites (Antisemitism), Anti-Elites
(Green Agenda), None]
Please check the following narrative
descriptions before making a decision:
Anti-EU (EU economic scepticism):
Narratives criticising EU's economic
policies and how they affect local
economies.
...
None: None of the above narratives.
Here are some example tweets and their
associated narrative label to help you
decide:
...
Tweet: Scary scenes on London streets.
Every day, UK and London in particular,
look like a city from the Middle East,
with angry wild Islamists bullying and
threatening people! Sick and scary!
Shame on @MayorofLondon'
Narrative: Religion-related
(Anti-Islam)
...
You should not provide any explanation
or justification.
Question: Can you detect any misleading
narrative in this tweet given the list
of narratives presented to you?
Tweet:``Tweet''
```

Topic Modeling

In Figure 7, we present the topic similarity across both 2019 and 2024 elections.

Data Statistics and Examples: Further Details

This section presents further details about the UKElection-Narratives dataset. Table 5 shows the label distribution in the dataset's train, development, and test subsets which were used in our benchmark experiments. Table 6 presents example tweets from the dataset.

Computational Resources

To perform our experiments, we used our internal servers using two different GPUs specifically *NVIDIA A100-PCIE-40GB* and *NVIDIA RTX A4500*.

	Train	Dev	Test
None	548	76	157
Distrust in institutions (Criticism of national policies)	171	23	47
Gender-related (LGBTQ+-related)	119	17	33
Religion-related (Anti-Islam)	72	10	19
Distrust in democratic system (Anti-Media)	72	10	19
Migration-related (Migrants societal threat)	65	9	17
Geopolitics (Foreign interference)	57	8	16
Distrust in democratic system (Elections are rigged)	50	7	14
Distrust in institutions (Failed state)	45	6	13
Political hate and polarisation (Pro far-left)	39	5	12
Political hate and polarisation (Pro far-right)	38	5	11
Anti-EU (EU political interference)	23	3	7
Anti-EU (Crisis of EU)	17	2	6
Anti-EU (EU economic scepticism)	13	2	4
Distrust in democratic system (Anti-Political system)	13	1	4
Political hate and polarisation (Anti-liberal)	12	1	4
Political hate and polarisation (Anti-woke)	6	1	1
Ethnicity-related (Association to political affiliation)	5	1	1
Gender-related (Language-related)	5	1	1
Ethnicity-related (Ethnic offensive language)	4	1	1
Anti-Elites (Soros)	4	1	1
Ethnicity-related (Threat to population narratives)	4	1	1
Anti-EU (EU Corruption)	3	1	1
Geopolitics (Pro-Russia)	3	1	1
Religion-related (Anti-Semitic conspiracy theories)	3	1	1
Ethnicity-related (Ethnic generalisation)	2	1	1
Anti-Elites (World Economic Forum / Great Reset)	2	1	1
Anti-Elites (Green Agenda)	2	1	1
Geopolitics (Anti-international institutions)	2	1	1
Religion-related (Interference with states' affairs)	1	1	1
Anti-Elites (Antisemitism)	0	1	1
Gender-related (Demographic narratives)	0	0	1
Distrust in democratic system (Immigrants right to vote)	0	0	1
Total	1400	200	400
Election 2019 tweets	700	100	200
Election 2024 tweets	700	100	200

Table 5: Data Splits and Label Distributions.

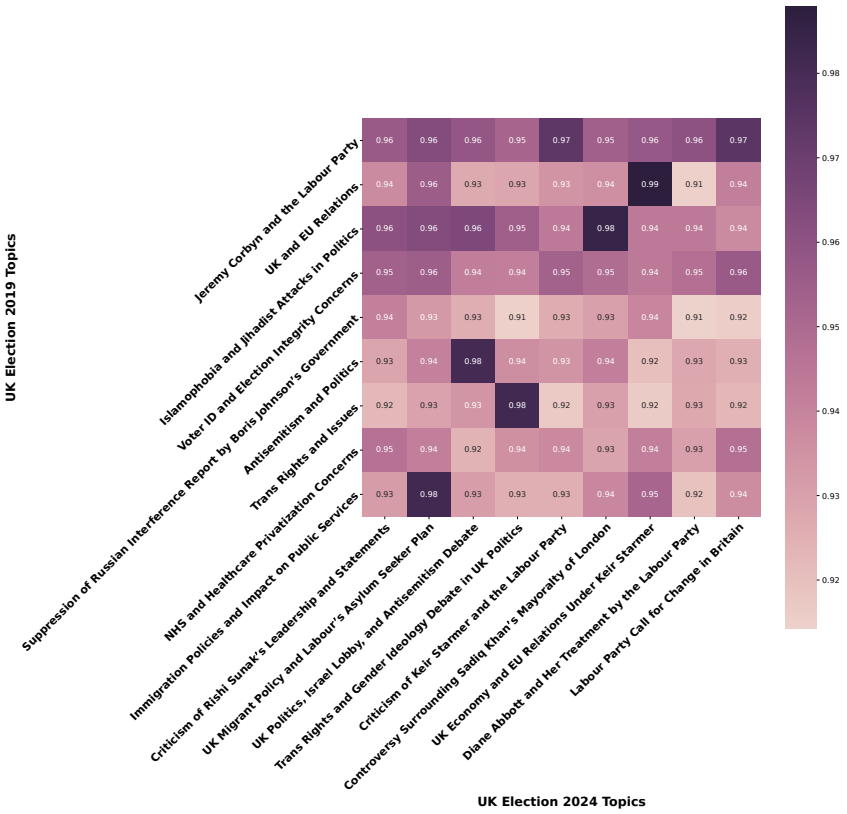

Figure 7: Cosine Similarity Matrix for Election 2019 and 2024 Topics Similarity.

Tweet: The process for casting a vote in UK elections is nothing short of archaic. No ID. No CCTV of who is coming & going. Who watches the boxes when they are driven to the central place for counting? Pencils!? The whole procedure is ripe for fraud & must change. #GE2109 **Super-narrative:** Distrust in democratic system **Narrative:** Elections are rigged
Tweet: As Rishi Sunak brags about his plans for prisons #PMQs Ministers today announced emergency measures to keep defendants in police cells as PRISONS HAVE RUN OUT OF SPACE What living political TORY hell is this? **Super-narrative:** Distrust in institutions **Narrative:** Failed state
Tweet: To all young people grappling with rising house costs... Mass immigration has caused this. Demand has outstripped supply. We will NEVER be able to build enough houses to support an extra 400,000 people EACH YEAR. In reality, freedom of movement has been a disaster for you. **Super-narrative:** Migration-related **Narrative:** Migrants societal threat
Tweet: The Greens should be reminded the UK accounts for 1% of global CO2 emissions, whilst their policies would inflict major economic damage. **Super-narrative:** Anti-elites **Narrative:** Green Agenda

Table 6: Sample tweets and corresponding super-narrative and narrative from **UKElectionNarratives**.