Hostility Detection in UK Politics: A Dataset on Online Abuse Targeting MPs

Mugdha Pandya and **Mali Jin** and **Kalina Bontcheva** and **Diana Maynard**
School of Computer Science, The University of Sheffield, United Kingdom
{mugdha.pandya|m.jin|k.bontcheva|d.maynard}@sheffield.ac.uk

Abstract

Social media platforms, particularly X, enable direct interaction between politicians and constituents but also expose politicians to hostile responses targetting both their governmental role and personal identity. This online hostility can undermine public trust and potentially incite offline violence. While general hostility detection models exist, they lack the specificity needed for political contexts and country-specific issues. We address this gap by creating a dataset of 3,320 English tweets directed at UK Members of Parliament (MPs) over two years, annotated for hostility and targeted identity characteristics (race, gender, religion). Through linguistic and topical analyses, we examine the unique features of UK political discourse and evaluate pre-trained language models and large language models on binary hostility detection and multi-class targeted identity type classification tasks. Our work provides essential data and insights for studying politics-related hostility in the UK.

CONTENT WARNING: This paper contains some examples of abusive and hateful content that some readers may find offensive or distressing.

1 Introduction

With the rise of social media use among politicians, especially on X, there has been an increase in direct interaction with the public (Agarwal et al., 2019). This interaction, while beneficial for communication and feedback, also exposes politicians to a significant number of hostile replies due to the anonymity of online platforms (Solovev and Pröllochs, 2022). Such hostility is considered a major concern as it erodes public trust in political processes and institutions, which disrupts constructive communication (Gross et al., 2023). Furthermore, it affects the personal lives and mental health of politicians, with online abuse sometimes leading to real-world threats and violence (Enock et al., 2023). In extreme cases, sustained hostility has driven politicians to step down from their roles and retreat from public life altogether (Scott, 2019).

Hostility targeting politicians is a global phenomenon characterised by widespread misogyny, sexism, and racism. Political and social science research indicates that all politicians receive hostility, but those from minority groups (e.g., Black, female, LGBTQ+) often face increased hostility based on their identity characteristics (Carson et al., 2024).

In NLP, sentiment analysis tools have been used to identify negative posts and facilitate studies on abuse trends (Hua et al., 2020; Ward and McLoughlin, 2020). Although general hostility detection is prevalent, identifying political hostility requires specialised approaches as political discussions often reflect a country's unique linguistic and cultural characteristics, incorporating regional colloquialisms, profanity and prejudices. For example, hostility towards people of colour is more prevalent in the US (Lavalley and Johnson, 2022), while the phenomenon of Islamophobia is more severe in India (Amarasingam et al., 2022). Furthermore, hostile posts are frequently tied to trending issues.

As the body of work on hate speech, abuse and hostility detection in NLP grows (Jahan and Oussalah, 2023), there has been a move towards developing resources specifically for political hate speech detection across different countries (Grimminger and Klinger, 2021; Jafri et al., 2023). In the UK, Members of Parliament (MPs) represent a wide range of backgrounds, and this diversity is mirrored in the nature of the abusive comments they receive (Gorrell et al., 2020). Studies have compiled datasets to analyse abuse trends specific to UK politics, though these datasets are not publicly available (Southern and Harmer, 2021; Bakir et al., 2024). While existing political datasets are available, only two contain hostility-related labels potentially usable for automated detection: Ward and McLoughlin (2020), with manual annotations,

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 254–266

and Agarwal et al. (2021), which relies entirely on automated labels without manual verification, limiting its reliability. A third dataset focuses solely on Islamophobia in UK politics (Vidgen and Yasseri, 2020), a specific form of identity-based hostility, which restricts its broader applicability. None comprehensively capture identity-based hostility.

We aim to bridge this gap by constructing a high-quality hostility dataset spanning a two-year period to cover diverse political topics in the UK. Our main contributions are:

- A publicly available dataset for political hostility towards UK MPs, containing 3,320 tweets with expert annotations for hostility and targeted identity characteristics (race, gender, religion, none), including individual annotations with confidence scores and gold labels;[1].
- In-depth linguistic and topical analyses identifying patterns and trending topics in the data;
- Demonstrating the utility of the dataset for political hostility detection by evaluating pre-trained language models (PLMs) and large language models (LLMs) on binary hostility classification and multi-class identity type classification.

Our work is distinctive in creating a dataset specifically for training models to detect identity-based political hostility towards UK MPs. Through topic analysis, we show how political hostility correlates with current events, which has crucial implications for model training (Jin et al., 2023). Our analysis reveals that the governing party faces proportionally more hostility, with race-based attacks being most prevalent. The dataset's two-year span offers greater topic diversity and generalisability than existing datasets, while uniquely capturing intersectional hostility through identity characteristic labels—a particularly harmful form of online hostility (Kuperberg, 2018, 2021).

2 Related Work

2.1 Online Hostility

The rise in social media usage has led to growing hostility (Walther, 2022; MacAvaney et al., 2019), spurring NLP research into online hostility tasks (Mansur et al., 2023; Jahan and Oussalah, 2023) like detecting hate speech, abuse, toxicity, and cyberbullying (Pavlopoulos et al., 2020; Mathew et al., 2021). While existing datasets include labels for targeted groups and various forms of harassment (Rosa et al., 2019; Hartvigsen et al., 2022), overlapping definitions complicate annotation and dataset comparison (Fortuna et al., 2020; Waseem et al., 2017). We address this by using "hostile" as an umbrella term. Though general hostility detection has been studied across social media platforms like Gab, Reddit, X, etc, (Mollas et al., 2022; Rieger et al., 2021), political hostility requires specialised research due to the distinct characteristics of the data (e.g. language, topic, country).

2.2 Online Hostility towards Politicians

Existing work on such data typically focuses on qualitative insights or analysis of summary statistics, revealing overarching themes of sexism, racism and religious hostility. Studies document gender-based hostility in Japan (Fuchs and Schäfer, 2021), disproportionate hate towards Democratic politicians of colour and women in the US (Solovev and Pröllochs, 2022; Grimminger and Klinger, 2021; Hua et al., 2020), and racial and gender-based abuse of UK MPs (Bakir et al., 2024; Kuperberg, 2018). While country-specific political hate speech detection models exist (Arabic in Algeria (Guellil et al., 2020), Chinese in Taiwan (Wang et al., 2022), Hindi in India (Jafri et al., 2023)), they typically overlook identity characteristics despite their prominence in political hate.

2.3 UK-Specific Hostility towards MPs

In the UK, studies of political hostility have examined both topics and identity characteristics. Bakir et al. (2024) and Farrell et al. (2021) found abuse towards MPs peaked during the first year of COVID-19, with women MPs, particularly those from non-white backgrounds, receiving higher levels of abuse. Gorrell et al. (2019) examined racial and religious abuse trends towards MPs relating to Brexit, along with abuse patterns before the 2015, 2017 (Gorrell et al., 2018) and 2019 (Gorrell et al., 2020) General Elections. Their research revealed correlations between abuse and MPs' prominence, Parliamentary events, and identity characteristics. Research on gender-based hostility shows female MPs face othering, belittling, discrediting, and stereotyping. Gender-based harassment correlates with lower success rates for female electoral candidates (Collignon and Rüdig, 2021), while

[1]Dataset is available at https://doi.org/10.5281/zenodo.10809694

Dataset	Time	Tweets	Labels
Agarwal et al. (2021)	1 Oct 2017 - 29 Nov 2017	2.5 M	hate; not hate
Vidgen et al. (2020)	Jan 2017 - June 2018	4000	none; weak islamophobia; strong islamophobia
Ward et al. (2020)	14 Nov 2016 - 28 Jan 2017	3000	non-abusive; not-directed; abusive; hate-speech
Our dataset	**Nov 2020 - Dec 2022**	**3320**	**not hostile; hostile - religion, gender, race, none**

Table 1: Datasets for automatic UK political hostility detection.

YouTube reinforces gender stereotypes and misogyny through hateful videos and comments (Esposito and Zollo, 2021). Female MPs encounter more incivility, including stereotyping and credibility challenges, than their male counterparts (Southern and Harmer, 2021). Gender intersects with other identity characteristics—age, class, race, and religious beliefs—in shaping hostility towards MPs (Kuperberg, 2021; Esposito and Breeze, 2022).

2.4 Existing Datasets for UK Political Hostility

Despite widespread awareness of UK political hostility, few NLP datasets and models exist. To the best of our knowledge, only 3 suitable datasets are currently available, detailed in Table 1.

Agarwal et al. (2021) compiled 2.5 million tweets spanning 2 months, containing binary hate labels and an analysis of topics and MP characteristics. However, these labels were generated entirely through automated means using 18 hate speech classifiers not trained on political data, without manual verification, which limits their reliability for training or evaluation purposes. Vidgen and Yasseri (2020) developed a dataset and classifier for detecting Islamophobia in political contexts, comprising 4000 expert-annotated tweets collected over 1.5 years with reported inter-annotator agreement metrics, but focus only on this single form of identity-based hostility. Ward and McLoughlin (2020) examined abuse trends by collecting 3000 negative tweets over 2.5 months through sentiment analysis, manually annotating hate and abuse, and showing that abuse related to both identity characteristics and reactions to political issues. However, their dataset appears to have been labelled by a single annotator, with no reported inter-annotator agreement, making the annotation quality difficult to assess.

Our work differs in that it specifically targets the automatic detection of UK political hostility across multiple identity characteristics, with multi-annotator manual labelling and reported inter-annotator agreement scores to ensure label reliability. Unlike existing datasets, our two-year collection period covers diverse topics over an extended timeframe, enabling more effective classifier generalisation (Jin et al., 2023). Additionally, we utilise the dataset to present preliminary findings about the nature of this hostility, as well as best methods for identifying it.

3 Data

3.1 Data Collection

Following Bakir et al. (2024), we used the X Streaming API to follow all 568 MPs with active X accounts. We collected 4 types of tweets related to each MP between November 2020 and December 2022: original tweets and retweets by the MPs, and replies to and retweets of these by others, resulting in over 30 million tweets, denoted as C.

3.2 Data Sampling

Manual annotation is not feasible for the entire dataset, so we sample a subset S, covering diverse time periods and topics, using the following steps:

- We choose a **subset of 18 MPs** covering diverse representation of identities and political affiliations. The pool includes both minority and majority identity groups (race: White, non-White; gender: male, female; religion: Christian, non-Christian).[2] 9 of the selected MPs are from the Conservative Party, 8 from the Labour Party, and 1 from the Scottish National Party. Table 7 in Appendix C presents the distribution of identities and parties.
- A **long temporal span** was ensured by sampling tweets from the 5 highest posting activity days for each MP, which occur in C.
- We exclude duplicate tweets and use an abusive language classifier (Gorrell et al., 2020) to identify **hostility** of all 2.54M individual tweets. For each of the 5 days, we sample 17 hostile and 20 non-hostile tweets, resulting in potentially 85 hostile and 100 non-hostile tweets per MP for manual annotation.

[2]The MPs' identity characteristics are based on self-declared public information.

In total, S contains 3,330 tweets in English.

3.3 Data Annotation

This process involves defining the guidelines, performing the annotation task, and undertaking quality control.

3.3.1 Annotation Guidelines

To address the challenge of differentiating between the closely related concepts of hate, abuse and toxicity, we combined their definitions from NLP literature into an umbrella term, hostile.

We consider political hostility detection as a hierarchical classification task. Given a tweet t, the aim is to classify t based on hostility (binary classification) and the target identity characteristics (multiclass classification). We formulate the task in a hierarchical manner similar to existing datasets like OffensEval (Zampieri et al., 2019) and HatEval (Basile et al., 2019). First, t is classified into two hostility labels: hostile and not hostile. If t is classified as hostile, then it will be further classified into at least 1 of the 4 target identity characteristic labels: religion, gender, race and none. Table 2 shows the definitions of each category and example tweets. Note that hostility can be intersectional (i.e., target multiple identity characteristics simultaneously), so a tweet can have more than 1 identity label. To provide a measure of reliability of each annotation, we include a confidence score of 1 to 5, from very low confidence to extreme confidence, for both hostility and identity characteristic labels.

3.3.2 Annotation Method

The annotation task was conducted in three steps: training, testing, and annotation. Steps 1 and 2 ensured high-quality annotations. Details of further measures taken to ensure high-quality annotations are in Appendix B. The entire annotation process was conducted using the collaborative web-based annotation tool Teamware 2 (Wilby et al., 2023).

1. **Training sessions**: These were conducted via in-person presentations explaining label definitions and detailed examples. Annotators were guided on setting up their accounts and familiarising themselves with the platform.

2. **Testing sessions**: Each annotator then underwent a test to ensure a proper understanding of the task and guidelines, consisting of 20 tweets covering all the labels. Annotators were required to label at least 70% correctly. Finally, annotators were provided with both the correct answers and explanations.

3. **Annotation**: On passing the test, annotators were assigned the actual annotation task. Figure 3 in Appendix B shows the platform user interface.

3.4 Dataset

The fully annotated dataset consists of 3,320 tweets, after removing posts containing URLs or user mentions only. We use 3 sets of gold labels:

- **Set 1:** The gold labels were assigned based on majority vote, i.e. the label chosen by at least 2 out of 3 annotators. For cases where multiple identity labels were chosen (intersectional), an expert assigned a single label.

- **Set 2:** Annotations with confidence <3 were removed to derive gold labels. For cases with 1 remaining annotation, that label was used. When there were 2 annotations, the higher confidence one was selected; if tied, an expert manually assigned the dominant label. For 3 remaining annotations, majority vote was applied as in Set 1.

- **Set 3:** To investigate intersectionality in the data and model performance, we used the same method as Set 2 for the binary hostility labels. For the identity labels, if there was an intersectional label with confidence >2, we chose that as the gold label. [3]

Table 3 shows the statistics of each set.[4] The top 6 rows present the frequency of each label for each set. Non-hostile tweets are predominant, followed by no identity and race-based hostile tweets. Set 3 includes the 43 intersectional labels, of which 5 target religion and gender, 21 religion and race, and 17 gender and race. The bottom 2 rows present the Fleiss' κ annotator agreement score (Fleiss, 1971) for hostility and target identity annotation. Set 2 exhibits the highest κ-value for both hostility (0.79) and identity (0.65) annotation, indicating substantial agreement (Artstein and Poesio, 2008). This suggests selecting annotations based on confidence scores helps to improve the quality of the dataset. The differences in the amount and type of hostility

[3] We had no cases of different intersectional labels with confidence >2.

[4] For Set 3, the value in parentheses shows the count of identity-based hostility that comes from intersectional labels.

Label	Definition	Example
Hostile	**Hostility towards a target group or individual. Intended to be derogatory, abusive, threatening, humiliating, inciting violence or hatred towards an individual/members of the group.**	*<USER >and <USER >Put back on your leash were you? There's a good boy*
Race	Hostility directed at a person/group based on racial background/ethnicity. Including discrimination based on somatic traits (e.g. skin colour), origin, cultural traits, language, nationality, etc..	*<USER >You're in England speak bloody ENGLISH!*
Gender	Hostility directed at a person/group based on their gender. Including negative stereotyping, objectification, using gendered slurs to insult, and threats of a sexual nature.	*<USER >If you can't stand the heat get the hell out of the kitchen next time elect a man to be PM, Liz Truss just proved there are things women can't do.*
Religion	Hostility directed at a person/group based on their religious beliefs. including misrepresenting the truth and criticism of a religious group without a well-founded argument.	*<USER >sick of you tweeting about muslims or any other religion. Your silence speaks the same bullshit, but its ok as Ramadan is over?!?!*
None	Do not refer to gender, race/ethnicity or religion.	*<USER >sucks! I wish someone would shoot her*
Not hostile	**Posts that are not hostile. Posts with profanity are not hostile unless their context makes it so.**	*<USER >will make a bad PM. Don't make this a race war. Please notice that he is a lousy politician*

Table 2: Hostility taxonomy with targeted identity type definitions and examples.

Hostility	**Identity**	**Set 1**	**Set 2**	**Set 3**
Hostile	Religion	36	41	52 (26)
	Gender	108	119	119 (22)
	Race	188	182	205 (38)
	None	1135	1112	1121 (0)
	Total	1467	1454	1454 (43)
Not Hostile	Total	1853	1866	1866
Fleiss' κ	Hostility	0.68	0.79	0.79
	Identity	0.51	0.65	0.47

Table 3: Label counts for each set.

MPs receive based on their political party and identity characteristics can be found in Appendix C.

4 Data Characterisation

4.1 Linguistic Analysis

We conduct a comparative linguistic analysis to investigate differences between content and language of hostile and non-hostile tweets. We use the Bag of Words (BOW) model and Linguistic Inquiry and Word Count (LIWC) Dictionary (Boyd et al., 2022) to identify linguistic patterns. We then use a univariate Pearson's correlation test to identify which linguistic patterns significantly correlate with hostile and non-hostile tweets. Tweets are pre-processed to replace URLs and @mentions with <URL >and <USER >, respectively) and stop words are removed using NLTK (Bird et al., 2009).

4.1.1 BOW

We represent each post as a TF-IDF weighted distribution of the 3,000 most frequent unigrams and bigrams using the BOW model. Figure 1 shows the differences in BOW features associated with hostile and non-hostile tweets as word clouds.

Unsurprisingly, we observe that hostile tweets are characterised by negative and abusive phrases (e.g."scum", "vile", "nothing good", "absolute disgrace"). They express anger or dissatisfaction at politicians, from questioning their abilities and distrusting their policies to insulting their personal traits. We also see emojis, e.g. "face_with_symbols_on_mouth" and "face_vomiting", representing the use of profanity and disgust. Below is an example from our dataset:

Tweet 1: *"Some in Cabinet are incompetent, others corrupt or evil. You are all 3. I have only contempt and disgust for you!"*

Secondly, phrases such as "go away", "shame resign" and "run country" in hostile tweets suggest that much of the hostility is directed at the Conservative (ruling) Party. Below is an example requesting the MP to resign:

Tweet 2: *"Too late with <USER>in charge & his cabinet of mendacious halfwits. Demand his resignation."*

Phrases such as "liar", 'corrupt", "never trust" and "know nothing" indicate general distrust in the MPs. Also, we notice some trending topics in hostile tweets (e.g. "vaccine passports", "illegal immigrants"), which reveal specific issues that cause dissatisfaction. The example tweet expresses the anger at policies relating to illegal immigration:

Tweet 3: *"What would you do about the illegal immigration welcome them with open arms wish we could send you to Rwanda and your filthy son"*

For non-hostile tweets, the correlation r is lower (as can be seen from the text size in Figure 1). However, they are correlated with words and phrases

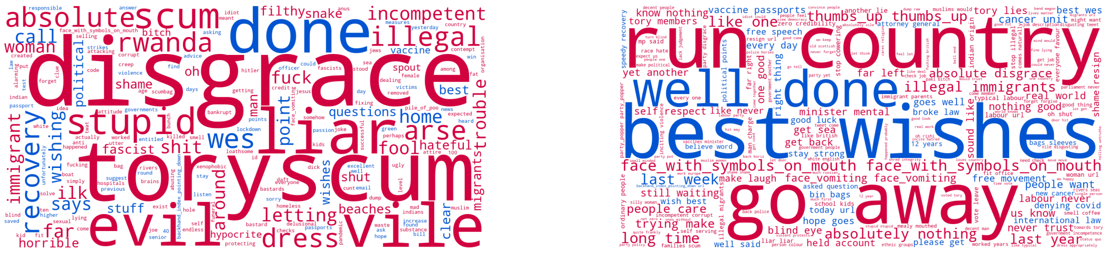

Figure 1: Top 100 BOW unigrams (left) and bigrams (right) associated with hostile and non-hostile tweets. The larger the text size, the higher the Pearson correlation coefficient r, and vice versa.

such as “excellent”, “best wishes” and “well done”. These suggest that non-hostile tweets often contain appreciative and positive emotions towards MPs. Some phrases (e.g. “asked questions”, “free movement”) indicate users’ attempts to voice their political concerns. The following tweet is an example conveying appreciation to the MP:

Tweet 4: *“<USER >was on fire! Another spectacular debate. Well done sir!”*

4.1.2 LIWC

Hostile	*r*	Not hostile	*r*
socrefs	0.186	Tone	0.192
you	0.181	OtherP	0.189
swearwords	0.162	AllPunc	0.183
clout	0.160	focuspast	0.133
tone_neg	0.160	comm	0.104
moral	1.51	prosoc	0.084
affect	0.142	polite	0.064
ppron	0.131	i	0.063
ethnicity	0.111	work	0.062
sex	0.109	tone_pos	0.061

Table 4: Top 10 LIWC categories for hostile and non-hostile tweets sorted by Pearson correlation (r) between the normalised frequency and the labels. All correlations are significant at p <.001, two-tailed t-test.

Each tweet is characterised using psycholinguistic categories from LIWC (Boyd et al., 2022). Table 4 presents the top 10 categories most strongly correlated with hostile and non-hostile tweets.

Similar to the BOW findings, we see that hostile tweets tend to have a negative tone (*tone_neg*) and convey negative emotions like anger and sadness (*affect*). They contain more assertive and judgmental language (*clout* and *moralisation*). Unsurprisingly, they also contain more swear words (*swearwords*) and sexual terms (*sex*). Interestingly, race-related (*ethnicity*) terms are frequent, suggesting that hostility is often related to race. The following tweet is an example from the dataset:

Tweet 5: *“What about black violence! Ur just a race divider. U Marxists have ruined this country & divided it further!”*

Topic	Representative Words
Brexit	brexit, uk, ireland, eu, europe, leave, deal, citizens, free, border
Illegal immigration	refugees, illegal, boats, rwanda, immigrants, asylum, raped, terrorists, seekers, migrants
Conservative party	tory, conservative, resign, vote, rishi, scum, tories, johnsonout, torysewageparty, cabinet
Labour party	labour, starmer, voters, corbyn, party, win, mps, election, abbott, protest
COVID-19	covid, virus, vaccine, lockdown, died, pandemic, mask, vulnerable, jab, nhs
Cost of living crisis	economy, bills, winter, job, tax, inflation, energy, nhs, heating, gas

Table 5: Topic groups and representative words

We notice that non-hostile tweets correlate highly with the tone marker (*tone*), particularly a positive tone (tone_pos). They are polite, more communicative (*comm*), and adhere to social norms (*prosoc* and *polite*), often consisting of explanations, feedback and questions. They also express concerns about work, jobs, schooling, etc. (*work*). Below is an example tweet expressing concerns about the new scheme:

Tweet 6: *“Please consider the scheme’s effect before acting on it. We suffered a lot during covid. The economy will not recover. Think carefully!”*

4.2 Topic Analysis

We perform topic analysis using BERTopic (Grootendorst, 2022) after removing stop words with NLTK (Bird et al., 2009). Due to the frequency of MP names and profanity, the topics are rather unclear. Once we remove these, clear topics and their representative words emerge. Table 5) presents the six most frequent topics, selected based on the frequency of their representative terms in the corpus. The topics relate to major events and issues in the UK, like Brexit (e.g., “europe”, “border”), illegal immigration (e.g., “refugees”, “terrorists”), and the cost of living crisis (e.g., “bills”, “tax”, “inflation”).

The following example is a hostile tweet expressing anger due to increased costs of bills:

Tweet 7: *"What planet do you live on? You haven't saved the day. Fuel is +40%. Energy bills are +50%. We're still f**ked. Make it make sense"*

Other popular topics are the two main political parties in our dataset (Conservative and Labour). However, the ruling Conservative party is likely to receive more hostility based on the larger proportion of negative terms we find, such as "scum", "johnsonout". Here is an example of hostile tweets mentioning the Conservative Party:

Tweet 8: *"<USER>is this you? Scum! You ludicrous pork Hay-bale. You bin bag full of custard. #ToryCriminalsUnfitToGovern"*

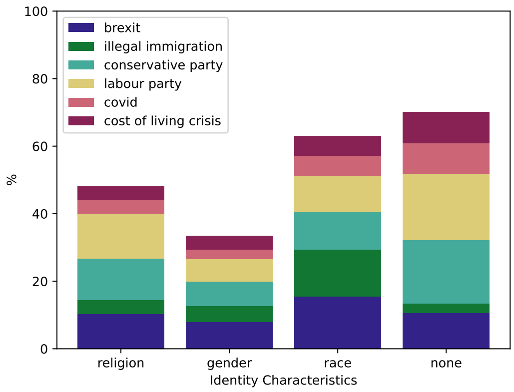

Figure 2: Proportion of topic-related tweets belonging to each identity characteristic label

Most topics appear in the same proportion in both hostile tweets and non-hostile tweets. The exception is "illegal immigration" which appears twice as much in hostile tweets. Figure 2 shows the proportions of topic-related tweets belonging to identity-based hostility. Looking at the distribution of "illegal immigration" and "Brexit", they appear most frequently in race-based hostile tweets. While the "Conservative party" and "Labour party" topics contribute to race-based hostile tweets, they appear more frequently in non-race, gender or religion-based hostility.

While all the tweets relate to MPs, they still naturally fall into topics related to current issues at the time. Due to its 2-year span, the dataset thus covers a diverse range of topics, since issues discussed on social media can change rapidly. This topic characterisation means that the dataset could eventually be used for analysis and comparison of hostility in relation to different issues over time.

5 Online Hostility Detection

We finetune PLMs for political hostility detection to test how they perform on our dataset. We also evaluate the ability of LLMs to identify political hostility on our dataset, demonstrating its value.

Given a text snippet, we define online hostility detection as two classification tasks: (1) binary hostility classification (if a tweet contains hostility or not) and (2) multi-class classification to see if it contains one of the four identity-based hostility types (religion, gender, race, none) or no hostility at all. For multi-class classification, we use a two-level hierarchical classification method.[5] The first classifiers classify tweets as hostile or not, while the second classifiers then classify the identity types of those identified as hostile.

5.1 Predictive Models

We use three PLMs for binary hostility classification and multi-class classification. We finetune **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), and a domain adaptation model, **RoBERTa-Hate** (Antypas and Camacho-Collados, 2023) (trained on 13 different hate speech datasets in English including political content), by adding a classification layer with softmax activation function on top of the [CLS].

We also evaluate two widely used LLMs on identifying hostile tweets and their targeted identity types. We use the **instruction-tuned LLaMA 3 8B model**[6] through the Hugging Face platform and the **GPT-3.5 model**[7] via the API, providing the model with a sequence of texts and a prompt with a task description to guide its output.

5.2 Experimental Set-up

Tweets are pre-processed, replacing URLs and user @mentions with special tokens <URL >and <USER >. We use BERT-base-uncased and RoBERTa-base models with a maximum sequence length of 256 tokens and batch size of 32. Training uses 5-fold cross-validation (4-fold training, split 9:1 for validation, 1-fold testing) with Cross Entropy Loss and AdamW optimizer at 5e-5 learning rate. Models are selected based on minimum validation loss over 15 epochs and trained on an

[5] We also tried a flat classification method, but we exclude the results as it performs slightly worse.

[6] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[7] https://platform.openai.com/docs/models/gpt-3-5-turbo

Model	Accuracy	Precision	Recall	F1
Binary Hostility Classification				
BERT	66.96±1.35	66.55±1.45	65.75±1.32	65.84±1.35
RoBERTa	68.13±0.83	68.04±0.62	67.55±0.51	67.44±0.48
RoBERTa-Hate	67.38±1.51	67.47±1.15	67.10±0.66	66.84±1.09
BERT	72.47±3.56	72.27±3.82	71.62±3.22	71.77±3.37
RoBERTa	71.77±3.37	72.26±2.05	69.15±2.99	68.86±3.65
RoBERTa-Hate	72.27±3.82	**73.44±1.00**	**73.16±1.44**	**73.03±1.27**
LLaMA	71.30±0.96	71.17±0.86	71.44±0.86	71.11±0.91
LLaMA-Def	**73.55±1.39**	73.21±1.42	72.76±1.43	72.91±1.43
GPT	60.57±1.93	69.97±1.21	64.20±1.72	58.67±2.41
GPT-Def	70.69±1.27	71.90±1.29	71.85±1.28	70.69±1.27
Multi-class Hostility Classification				
BERT	60.78±1.00	27.44±5.76	25.60±2.01	24.79±2.14
RoBERTa	61.99±1.32	37.66±9.18	27.53±2.08	28.87±2.44
RoBERTa-Hate	62.47±2.29	38.77±5.79	28.42±1.58	31.21±2.39
BERT	66.30±4.52	32.42±2.08	28.41±2.95	29.09±3.08
RoBERTa	66.14±1.70	40.77±8.44	30.47±6.38	32.85±7.03
RoBERTa-Hate	68.10±1.57	39.93±4.37	32.18±4.57	**33.81±4.63**
LLaMA	64.79±1.97	54.62±3.75	51.77±3.83	52.15±3.65
LLaMA-Def	64.70±2.37	53.11±11.04	53.98 ±3.67	54.16±4.43
GPT	54.19±2.77	**55.61±5.11**	54.29±5.79	50.53±5.08
GPT-Def	64.43±1.52	54.15±3.42	**60.02±3.11**	**55.98±3.08**
BERT	66.30±4.32	21.53±2.29	19.14±1.51	19.49±1.60
RoBERTa	65.84±2.24	30.52±8.89	23.01±6.86	23.60±6.55
RoBERTa-Hate	**67.80±2.07**	26.00±2.28	25.09±3.29	24.22±2.96

Table 6: Performance metrics (± std. dev.) for binary and multi-class hostility classification for Set 1, Set 2 and Set 3 (only multi-class).

NVIDIA A100 GPU. All LLM experiments use 0.1 temperature. For evaluation, we report average Accuracy, Precision, Recall and macro F1 over 5 folds with standard deviations.

For LLMs, we input the prompt to specify the task for binary hostility classification: *Classify the tweet as hostile or not hostile* with (**LLaMA-Def**, **GPT-Def**) or without definitions (**LLaMA**, **GPT**). For 2-level hierarchical classification, we input the prompt based on the outputs from the binary hostility classification: *Classify the tweet as hostility based on race, gender, religion or other*. For a fair comparison, we also report the average performance over 5 folds with the same data in each fold.

5.3 Results

5.3.1 Binary Hostility Classification

Table 6 presents the predictive results of all models on binary hostility classification using Set 1 and Set 2 (top 10 rows). We exclude Set 3 because the intersectional labels in identity type annotation do not affect the binary labels. Overall, RoBERTa-Hate on Set 2 achieves the best performance among all models, reaching a macro F1 score up to 73.03 (in bold). We observe that models trained on Set 2 achieve better performance than those trained on Set 1 (e.g., 68.86 vs. 67.44 F1 for RoBERTa on Set 2 and Set 1), highlighting the importance of selecting annotations based on confidence scores. Also, the domain adaptation model (i.e., RoBERTa-Hate) outperforms the vanilla models on Set 2 (e.g., 68.86 F1 for RoBERTa vs. 73.03 F1 RoBERTa-Hate) and has comparable performance with the vanilla models on Set 1 (e.g., 67.44 F1 for RoBERTa vs. 68.84 F1 for RoBERTa-Hate).[8]

We test LLMs on Set 2, where better results are achieved. Among four LLM settings, LLaMA-Def achieves the best performance with a macro F1 score of 72.91, followed by GPT-Def (70.69 F1). We notice that adding label definitions in the prompt improves performance (+1.80 F1 for LLaMA and +12.02 F1 for GPT). We argue that advanced LLMs do not show significant advantages on binary hostility classification as it is a simple and straightforward 2-class classification task.

5.3.2 Multi-class Hostility Classification

Table 6 presents the results of all models on multi-class hostility type classification using three sets of data in 2-level hierarchical method (bottom 13 rows). Among all PLMs, the best performing model is RoBERTa-Hate on Set 2 with an F1 score of 33.81 (in bold). Similar to the binary hostility classification, models in Set 2 achieve the best predictive results compared with the same models trained on other sets (e.g., 32.85 F1 for RoBERTa), followed by Set 1 (e.g., 31.21 F1 for RoBERTa-Hate). The domain adaptation model, RoBERTa-Hate, outperforms the vanilla RoBERTa model with a larger difference compared to binary hostility classification (e.g., +4.17 F1 vs. +0.96 F1 on Set 2 in binary hostility classification and in multi-class hostility classification). Additionally, RoBERTa outperforms BERT across three sets of data (e.g., 32.85 vs. 29.09 F1 on Set 2).

Similar to the hostility classification task, we only apply LLMs on Set 2. First of all, GPT-Def outperforms all PLMs and LLMs, reaching a macro F1 score up to 55.98, which is 12.67 higher than the best-performing PLM, RoBERTa-Hate. Secondly, in general, adding definitions of each hostility type boosts the performance. Moreover, prompts with definitions result in a larger improvement on the multi-class classification than the binary one (e.g., +5.45 F1 for GPT in hierarchical classification).

6 Conclusion

This work focuses on the creation of data for investigating online hostility towards UK politicians. We

[8]We also evaluate Set 1 and Set 2 on the same test set with the same labels (we exclude Set 3 as adding intersectional labels leads to different test sets). RoBERTa and RoBERTa-Hate using Set 2 achieve better results than using Set 1 (72.46 vs. 71.11 F1 and 74.10 vs. 73.26 F1 accordingly).

developed an English dataset of 3,320 tweets, manually annotated with hostility as well as targeted identity characteristics: religion, gender, and race. We also conducted extensive linguistic and topical analyses to provide deeper insights into the specific content of these hostile interactions. By constructing and analysing such a dataset, we identify key patterns, such as the prevalence of race-based hostility, especially regarding immigration issues in the UK. Our findings also suggest that there is a general lack of trust in MPs in the UK. Finally, we evaluated the performance of various PLMs and LLMs on binary hostility classification and multi-class targeted identity type classification using our dataset. This study not only offers valuable data but also lays the groundwork for future research aimed at understanding and mitigating the impact of online hostility in UK political contexts.

7 Limitations

We included only 18 MPs out of 568 possible MPs with active Twitter accounts in our final dataset. We also focus on only 3 political parties in the UK. This limited sample size was necessitated by both the demands of manual annotation and the varying levels of social media engagement across MPs. Our work does not address sexuality-based hostility, due to practical constraints: unlike gender, race, and religion, which were based on self-declared public information, sexuality is not consistently publicly declared by MPs. We limited our identity characteristics to only those that could be reliably determined from public self-declarations. Our analytical approach employed binary categorisations that may oversimplify the UK's diverse ethnic and religious landscape. We adopted these simplifications to make the annotation task and subsequent analysis more tractable. Future work would benefit from more nuanced approaches to categorising identity characteristics. While we aimed to select a diverse representation, this sample may not fully capture the breadth of experiences across all UK parliamentarians.

8 Acknowledgments

This study was conducted as part of the "Responsible AI for Inclusive, Democratic Societies: A cross-disciplinary approach to detecting and countering abusive language online" project [grant number R/163157-11-1].

References

Pushkal Agarwal, Oliver Hawkins, Margarita Amaxopoulou, Noel Dempsey, Nishanth Sastry, and Edward Wood. 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*, pages 5–16.

Pushkal Agarwal, Nishanth Sastry, and Edward Wood. 2019. Tweeting MPs: Digital engagement between citizens and members of parliament in the UK. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 26–37.

Amarnath Amarasingam, Sanober Umar, and Shweta Desai. 2022. "Fight, die, and if required kill": Hindu nationalism, misinformation, and Islamophobia in India. *Religions*, 13(5):380.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Mehmet Emin Bakir, Tracie Farrell, and Kalina Bontcheva. 2024. Abuse in the time of COVID-19: the effects of Brexit, gender and partisanship. *Online Information Review*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Andrea Carson, Gosia Mikolajczak, Leah Ruppanner, and Emily Foley. 2024. From online trolls to 'slut shaming': Understanding the role of incivility and gender abuse in local government. *Local Government Studies*, 50(2):427–450.

Sofia Collignon and Wolfgang Rüdig. 2021. Increasing the cost of female representation? the gendered effects of harassment, abuse and intimidation towards parliamentary candidates in the UK. *Journal of elections, public opinion and parties*, 31(4):429–449.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Florence Enock, Pica Johansson, Jonathan Bright, and Helen Zerlina Margetts. 2023. Tracking experiences of online harms and attitudes towards online safety interventions: Findings from a large-scale, nationally representative survey of the British public. *Nationally Representative Survey of the British Public (March 21, 2023)*.

Eleonora Esposito and Ruth Breeze. 2022. Gender and politics in a digitalised world: Investigating online hostility against UK female MPs. *Discourse & Society*, 33(3):303–323.

Eleonora Esposito and Sole Alba Zollo. 2021. "how dare you call her a pig, I know several pigs who would be upset if they knew" a multimodal critical discursive approach to online misogyny against UK MPs on youtube. *Journal of language aggression and conflict*, 9(1):47–75.

Tracie Farrell, Mehmet Bakir, and Kalina Bontcheva. 2021. MP twitter engagement and abuse post-first COVID-19 lockdown in the UK: White paper. *arXiv preprint arXiv:2103.02917*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

FORCE11. 2020. The fair data principles. `https://force11.org/info/the-fair-data-principles/`.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794.

Tamara Fuchs and Fabian Schäfer. 2021. Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter. In *Japan forum*, volume 33, pages 553–579. Taylor & Francis.

Genevieve Gorrell, Mehmet E Bakir, Mark A Greenwood, Ian Roberts, and Kalina Bontcheva. 2019. Race and religion in online abuse towards UK Politicians: Working paper. *arXiv preprint ArXiv:1910.00920 [Cs]*.

Genevieve Gorrell, Mehmet E Bakir, Ian Roberts, Mark A Greenwood, and Kalina Bontcheva. 2020. Which politicians receive abuse? four factors illuminated in the UK general election 2019. *EPJ Data Science*, 9(1):18.

Genevieve Gorrell, Mark Greenwood, Ian Roberts, Diana Maynard, and Kalina Bontcheva. 2018. Twits, twats and twaddle: Trends in online abuse towards UK politicians. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

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

Joelle Gross, Samuel Baltz, Mara Suttmann-Lea, Lia Merivaki, and Charles Stewart III. 2023. Online hostility towards local election officials surged in 2020. *Available at SSRN 4351996*.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Sara Chennoufi, Hanene Maafi, and Thinhinane Hamitouche. 2020. Detecting hate speech against politicians in Arabic community on social media. *International Journal of Web Information Systems*, 16(3):295–313.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing Twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during Indian election campaign: A new low-resource dataset and baselines.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Mali Jin, Yida Mu, Diana Maynard, and Kalina Bontcheva. 2023. Examining temporal bias in abusive language detection. *arXiv preprint arXiv:2309.14146*.

Rebecca Kuperberg. 2018. Intersectional violence against women in politics. *Politics & Gender*, 14(4):685–690.

Rebecca Kuperberg. 2021. Incongruous and illegitimate: Antisemitic and Islamophobic semiotic violence against women in politics in the United Kingdom. *Journal of Language Aggression and Conflict*, 9(1):100–126.

Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, Aisling Third, and Miriam Fernandez. 2022. Misogynoir: challenges in detecting intersectional hate. *Social Network Analysis and Mining*, 12(1):166.

Ryan Lavalley and Khalilah Robinson Johnson. 2022. Occupation, injustice, and anti-Black racism in the United States of America. *Journal of Occupational Science*, 29(4):487–499.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Zainab Mansur, Nazlia Omar, and Sabrina Tiun. 2023. Twitter hate speech detection: a systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, 11:16226–16249.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and reddit. *Social Media+ Society*, 7(4):20563051211052906.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Jennifer Scott. 2019. Women MPs say abuse forcing them from politics.

Kirill Solovev and Nicolas Pröllochs. 2022. Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *Proceedings of the ACM Web Conference 2022*, pages 3656–3661.

Rosalynd Southern and Emily Harmer. 2021. Twitter, incivility and "everyday" gendered othering: An analysis of tweets sent to UK members of parliament. *Social science computer review*, 39(2):259–275.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Joseph B Walther. 2022. Social media and online hate. *Current Opinion in Psychology*, 45:101298.

Chih-Chien Wang, Min-Yuh Day, and Chun-Lian Wu. 2022. Political hate speech detection and lexicon building: A study in Taiwan. *IEEE Access*, 10:44337–44346.

Stephen Ward and Liam McLoughlin. 2020. Turds, traitors and tossers: the abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, 26(1):47–73.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

David Wilby, Twin Karmakharm, Ian Roberts, Xingyi Song, and Kalina Bontcheva. 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*, pages 145–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

A Dataset Availability

Our dataset is publicly available in accordance with the FAIR principles (FORCE11, 2020):

- **Findable:** Our dataset is published in the Zenodo dataset-sharing service with a unique DOI. It can be found at `https://doi.org/10.5281/zenodo.10809694`.
- **Accessible:** Original tweets can be retrieved using their tweet IDs via the standard X API.[9]

[9]https://developer.twitter.com/en/docs/twitter-api/tweets/lookup/api-reference/get-tweets-id

- **Interoperable:** File structure and column descriptions are detailed in a readme file and the CSV format ensures broad compatibility across data processing tools.

- **Re-usable:** Our dataset can be re-used by anyone who has an X developer account.

B Annotation Task

Annotation Platform

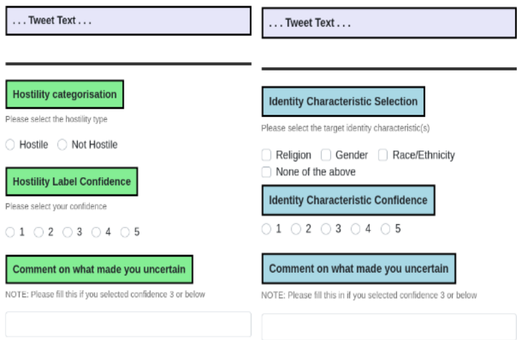

Figure 3: Annotation platform user interface.

Annotation Task Quality

A number of steps were taken to ensure high-quality manual annotations. Annotators were recruited from postgraduate courses in Politics and Computer Science. The only prerequisite was that they had to be familiar with UK politics and colloquialisms. We placed no restriction on age, gender, ethnicity, etc. so as to not bias the labels. We contacted potential annotators by emailing the respective course groups. Each annotator was paid 30 GBP for the annotation of 200 tweets. We recruited a total of 48 annotators. Each tweet in S is labelled by 3 annotators.

During the task, annotators were instructed to look up unfamiliar terms and slang. Each annotator was allowed to annotate only 200 tweets in total, and the task did not need to be completed in one sitting. This allowed annotators to take breaks and prevented them from getting overly desensitised to the hostile content.

A manual analysis of the annotations revealed that some annotators had incorrectly confused the race and religion labels in a few cases where Muslims and Jews were being targeted. Therefore, expert annotators made corrections to these labels.

C Dataset Information

MP Identity and Political Party Statistics

Party	Conservative	Labour	SNP	Total
Female	6	4	1	11
Male	3	4	0	7
Non-white	7	4	1	12
White	2	4	0	6
Not Christian	5	2	1	8
Christian	4	6	0	10

Table 7: Statistics of MP identity characteristics and political parties.

Quantity and Quality of Hostility

Figures 4 and 5 show the number and types of hostile tweets MPs receive based on their political party and identity group. The horizontal pink (Figure 4) and black (Figure 5) lines mark the mean value for each group. On average, Conservative MPs receive more race-based hostility. For gender and religion-based hostility, on average, MPs from both parties receive similar levels. However, there are some Labour MPs who receive more identity-based hostility than others (e.g. Diane Abbott, David Lammy).Due to only one SNP MP in our study, we do not include SNP in this comparison.

In Figure 5, we see that while male (M) MPs receive more hostile tweets, female (F) MPs face disproportionately more gender-based hostility, as expected. Similarly, non-white (NW) and non-Christian (NC) MPs face significantly higher levels of general, race- and religion-based hostility. Interestingly, we see that MPs from racial and religious minority groups consistently receive more general hostility and identity-based hostility (consistently higher mean values for all types of hostile tweets) than their white (W) or Christian (C) counterparts. This highlights the issues of intersectional hostility (Kwarteng et al., 2022), where individuals belonging to multiple minority groups experience compounded forms of discrimination and harassment.

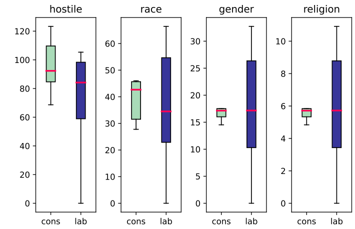

Figure 4: Comparing political party-based differences in the amount and type of hostility received

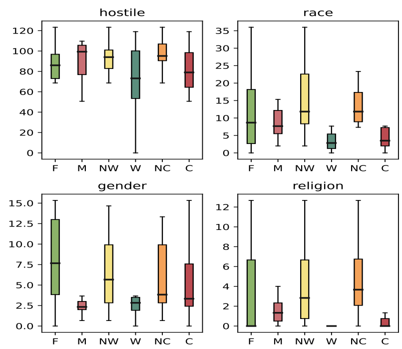

Figure 5: Comparing identity-based differences in the amount and type of hostility received