Preprint:

Pandya, M., Jin, M., Bontcheva, K. orcid.org/0000-0001-6152-9600 et al. (1 more author) (Submitted: 2024) Hostility detection in UK politics: A dataset on online abuse targeting MPs. [Preprint] (Submitted)

https://doi.org/10.48550/arXiv.2412.04046

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

Mugdha Pandya, Mali Jin, Kalina Bontcheva, Diana Maynard

Department of Computer Science, University of Sheffield
{mugdha.pandya, m.jin, k.bontcheva, d.maynard}@sheffield.ac.uk

Abstract

Numerous politicians use social media platforms, particularly X, to engage with their constituents. This interaction allows constituents to pose questions and offer feedback but also exposes politicians to a barrage of hostile responses, especially given the anonymity afforded by social media. They are typically targeted in relation to their governmental role, but the comments also tend to attack their personal identity. This can discredit politicians and reduce public trust in the government. It can also incite anger and disrespect, leading to offline harm and violence. While numerous models exist for detecting hostility in general, they lack the specificity required for political contexts. Furthermore, addressing hostility towards politicians demands tailored approaches due to the distinct language and issues inherent to each country (e.g., Brexit for the UK). To bridge this gap, we construct a dataset of 3,320 English tweets spanning a two-year period manually annotated for hostility towards UK MPs. Our dataset also captures the targeted identity characteristics (race, gender, religion, none) in hostile tweets. We perform linguistic and topical analyses to delve into the unique content of the UK political data. Finally, we evaluate the performance of pre-trained language models and large language models on binary hostility detection and multi-class targeted identity type classification tasks. Our study offers valuable data and insights for future research on the prevalence and nature of politics-related hostility specific to the UK.

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, Sastry, and Wood 2019). This interaction, while beneficial for communication and feedback, also exposes politicians to a significant amount 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 some politicians to step down from their roles and retreat from public life altogether (Scott 2019).

Hostility towards politicians is a global phenomenon characterised by widespread misogyny, sexism, and racism. Political and social scientists investigate it through surveys, interviews, and extensive quantitative and qualitative analyses of social media data (Håkansson 2024; Collignon and Rüdig 2021; Scott 2019). Their findings indicate that all politicians receive hostility to some degree, but those from minority groups (e.g., Black, female, LGBTQ+) often face increased hostility based on their identity characteristics (Hua, Naaman, and Ristenpart 2020; Carson et al. 2024).

In Natural Language Processing (NLP), sentiment analysis tools have been used to identify negative tweets and facilitate studies on abuse trends (Hua, Naaman, and Ristenpart 2020; Ward and McLoughlin 2020). Although general hostility detection is prevalent, identifying political hostility requires specialised approaches because political discussions often reflect a country's unique linguistic and cultural characteristics, incorporating regional colloquialism, 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, Umar, and Desai 2022). Furthermore, hostile posts are frequently tied to popular issues at that time. For instance, recent hostility towards politicians has largely centred around illegal immigration in the UK (Goodman and Locke 2024).

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; Guellil et al. 2020; 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 (Southern and Harmer 2021; Bakir, Farrell, and Bontcheva 2024; Gorrell et al. 2018). As discussed in the related work section, most of these datasets lack hostility-related labels. Among them, only two datasets include labels suitable for automatic political hostility detection, but they do not take into account identity characteristics (Agarwal et al. 2021; Ward and McLoughlin 2020). A third dataset was created in the context of UK political hostility and used to train a classifier, but with a focus on detecting Islamophobia alone (Vidgen and Yasseri 2020).

In this paper, 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 specific to political hostility towards UK MPs consisting of 3,320 tweets with expert annotations for hostility and the targeted identity characteristics (race, gender, religion, none), including individual annotations with confidence scores and gold labels;[1]
- In-depth linguistic and topical analyses to identify linguistic patterns and trending topics in our hostility dataset towards UK MPs;
- Evaluation of pretrained language models (PLMs) and large language models (LLMs) on the task of binary hostility identification and multi-class targeted identity type classification in flat and 2-level hierarchical classification settings.

Our work distinguishes itself from others by creating a dataset specifically designed for training models to automatically detect political hostility towards UK politicians based on targeted identity characteristics. Through topic analysis, we demonstrate that political hostility is closely tied to contemporaneous events. While this is unsurprising, it nevertheless has important ramifications for training models (Jin et al. 2023). Linguistic analysis and dataset statistics reveal that the governing political party receives the most hostility proportionally, with race-based hostility being the most prominent among the identity characteristics studied. The extended two-year data collection period of our dataset thus provides a broader range of topics than existing datasets, improving both topic diversity and the generalisability of classifiers trained on this data (Jin et al. 2023). Additionally, our dataset includes labels for identity characteristics and their combinations, as intersectional abuse is a significant and particularly damaging feature of online hostility (Kuperberg 2018, 2021).

Related Work

Online Hostility

The rise in social media usage has resulted in a growing amount of hostility (Walther 2022; MacAvaney et al. 2019). Consequently, this has prompted NLP research into different hostility detection tasks such as (Mansur, Omar, and Tiun 2023; Jahan and Oussalah 2023) hate speech, abuse, toxicity, offence, trolling, cyberbullying, etc (Pavlopoulos et al. 2020; Mathew et al. 2021). Hate speech and abuse datasets have labels for hate, abuse, and offence with additional labels for the targeted groups like gender and race (Zampieri et al. 2019; Basile et al. 2019). Toxicity and cyberbullying datasets have label such as harassment, aggression, toxic, etc. (Rosa et al. 2019; Hartvigsen et al. 2022) However, the definitions of these tasks are very similar (Fortuna, Soler, and Wanner 2020; Basile et al. 2019), making it challenging for annotation and for comparison of datasets (Zampieri et al. 2019; Waseem et al. 2017) To avoid this problem, we combine the definitions of these terms in NLP literature into one umbrella term - hostile. While lots of work has been done to combat general hostility detection issues on different social media like Gab, Reddit, X (formerly Twitter), etc, (Jahan and Oussalah 2023; Mollas et al. 2022; Rieger et al. 2021) the problem of the specificity of political hostility data (i.e., language, topic, country) needs more specialised research.

Online Hostility towards Politicians

Existing work in political hostility typically focuses on qualitative insights or analysis of summary statistics. This work is widespread, and there seems to be an overarching theme of sexism, racism and religious hostility. For instance, female politicians face negative sentiments and attitudes in Japan (Fuchs and Schäfer 2021); in the US, people of colour from the Democratic party and female politicians receive disproportionate hate (Solovev and Pröllochs 2022; Grimminger and Klinger 2021; Hua, Naaman, and Ristenpart 2020); MPs in the UK face substantial racial and gender-based abuse (Bakir, Farrell, and Bontcheva 2024; Kuperberg 2018).

Datasets and machine learning models to detect political hate speech have been created for different countries and their corresponding languages (Arabic in Algeria (Guellil et al. 2020), Chinese in Taiwan (Wang, Day, and Wu 2022), Hindi in India (Jafri et al. 2023)). While these datasets can be used to detect country-specific political hate speech, they do not take identity characteristics into account despite their prominence in political hate speech.

UK-Specific Hostility towards MPs

In the UK, political hostility has been studied based on topics and identity characteristics. Bakir, Farrell, and Bontcheva (2024) and Farrell, Bakir, and Bontcheva (2021) found that abuse towards MPs was at an all-time high during the first year of the COVID-19 pandemic and that women MPs, especially from non-white backgrounds, received higher abuse. Gorrell et al. (2019) investigated trends in racial and religious abuse towards MPs around Brexit and also abuse trends leading up to the 2015, 2017 (Gorrell et al. 2018) and 2019 (Gorrell et al. 2020) General Elections. They found that prominence, Parliamentary events, and MP identity characteristics correlated with receiving abuse.

A large body of work has focused on gender-based hostility showing that the hostility female MPs face is often in the form of othering, belittling, discrediting, and stereotyping. For example, female electoral candidates' lower success rates were correlated with gender-based harassment (Collignon and Rüdig 2021); gender stereotypes and misogyny are reinforced on YouTube through hateful videos and comments (Esposito and Zollo 2021); female MPs face more incivility, including stereotyping and questioning credibility than male MPs (Southern and Harmer 2021). However, gender alone is not the only dimension where MPs face hostility. Rather, it is intertwined with other identity characteristics such as age, class, race, and religious beliefs. (Kuperberg 2021; Esposito and Breeze 2022).

[1]Dataset is available at https://zenodo.org/records/10809695

Existing Datasets for UK Political Hostility

Despite so much awareness about political hostility in the UK, only a small amount of work has been carried out in developing NLP datasets and models specifically for its automatic detection. To the best of our knowledge, there are only 3 datasets that are suitable for this task. Details of these are in Table 1. Agarwal et al. (2021) created a dataset of 2.5 million tweets collected over 2 months. They used 18 existing social media hate speech classifiers to generate binary hate labels and then studied the topic and MP characteristic trends in political hate speech. However, these hate speech classifiers were not trained specifically on political hate speech data. Vidgen and Yasseri (2020) created a dataset and classifier to detect Islamophobia in a political hate speech context. The dataset consists of 4000 tweets collected over 1.5 years with expert manual annotations. While this study uses political social media data, its focus is on Islamophobia alone. Therefore, the labels pertain specifically to Islamophobia rather than to hostility. Ward and McLoughlin (2020) used sentiment analysis to collect negative tweets from which they created a dataset of 3000 tweets collected over 2.5 months manually annotated for hate and abuse. They used this data to study abuse trends and found that other than identity characteristics, one of the main causes of abuse was reacting to political topics and issues.

The work presented in this paper differs from existing datasets as it is specifically designed to facilitate the automatic detection of political hostility in the UK, focusing on multiple identity characteristics. Unlike the limited suitable existing datasets, our data collection spans two years, covering a broad range of topics over an extended period. This timeframe is crucial for creating classifiers that generalise more effectively (Jin et al. 2023). Additionally, we use the dataset to show some preliminary findings about the nature of this hostility, as well as methods to best identify it.

Data

We develop our dataset in 3 steps: data collection, data sampling and annotation.

Data Collection

Following the method used by Bakir, Farrell, and Bontcheva (2024), the Twitter (now X) Streaming API is used to follow the accounts of all MPs (568) with active X accounts. We collect four types of tweets relating to each MP between November 2020 and December 2022: the original tweets posted by the MPs, replies to them, retweets of tweets posted by them and retweets created by them. This collection contains over 30 million tweets, which we denote as C.

Data Sampling

The sheer volume of tweets in C makes manual annotation infeasible. Therefore, we sample a subset of C for the annotation task covering diverse time periods and topics, we denote as S. We employ the following sampling steps:

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

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

Data Annotation

The data annotation process consists of defining the guidelines, performing the annotation task and quality control.

Annotation Guidelines To address the challenge of differentiating between the closely related concepts such as hate, abuse and toxicity, we combined the definitions of these terms from NLP literature into an umbrella term, hostile (see Table 3). We revised the definitions multiple times to strike a balance between the prescriptive and descriptive paradigm (Röttger et al. 2022). The former enforces definitions and rules that annotators must abide to, while the latter provides guidelines that allow annotators to apply their own understanding. To this end, small focus groups were conducted to discuss the rigidity of the label definitions.

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 one of the four target identity characteristic labels: religion, gender, race and none. Table 3 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 one 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. Table 9 in the appendix presents the confidence scores with explanations.

Annotation Method The annotation task was conducted in three steps: training, testing, and annotation. Steps 1 and 2

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

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**	**hostile-none, hostile-religion, hostile-gender, hostile-race, non-hostile**

Table 1: Datasets for automatic UK political hostility detection.

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 2: Statistics of MP identity characteristics and belong parties.

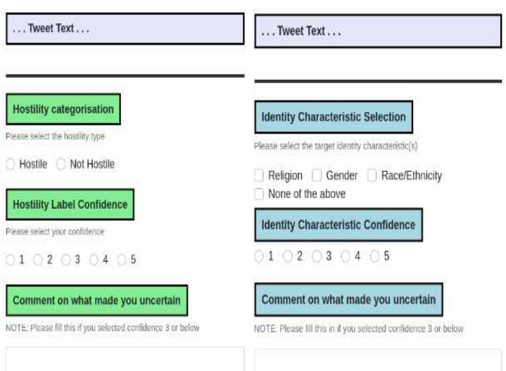

Figure 1: Annotation platform user interface.

ensured high-quality annotations. The entire annotation process was conducted using the collaborative web-based annotation tool Teamware 2 (Wilby et al. 2023).

1. **Training sessions**: Training sessions were conducted in which annotators received in-person presentations explaining label definitions with detailed examples. Annotators were also guided on setting up their annotator account and familiarising themselves with the platform.
2. **Testing sessions**: Each annotator then underwent a short 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 14 of the 20 annotations correctly. Finally, annotators were provided with the correct answers as well as explanations.
3. **Annotation**: Once annotators passed the test, they were added to the actual annotation task and were shown tweets sequentially. Figure 1 shows the platform user interface.

Annotation Task Quality A number of steps were taken to ensure high-quality manual annotations. Annotators were recruited from postgraduate study 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 annotation results established that some annotators had incorrectly confused the race and religion labels in cases where Muslims and Jews were being targeted. Therefore, expert annotators corrected this small number of annotations.

Dataset

The fully annotated dataset consists of 3,320 tweets in total after removing posts that contain URLs or user mentions only. We use 3 sets of gold labels for our modelling experiments, as follows:

- **Set 1:** The gold labels were assigned based on the majority vote, i.e. the label which has at least 2 annotations out of 3. For the cases where multiple identity labels were chosen (intersectional), an expert manually assigned the dominant label in their opinion.
- **Set 2:** Annotations with confidence <3 were removed, and gold labels are derived from the remaining annotations. In the cases where only one annotation remained for a tweet, that was selected as the gold label. When there were 2 annotations, the annotation with the higher confidence was selected. If both had equal confidence, an expert manually assigned the dominant label in their opinion. If all 3 annotations remained, we used the majority vote method as used in Set 1.
- **Set 3:** To investigate intersectionality in the data and model performance,we used the same method as Set 2 for high-level hostility labels. For the lower-level identity labels, if there was an intersectional label with confidence >2, we chose that as the gold label. We had no cases of different intersectional labels with confidence >2.

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 >Your 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 prime minister, Liz Truss just showed us 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 >is the worst human being. I wish someone would shoot her*
Not hostile	**Posts that are not hostile. a tweet containing profanity is not hostile unless its context makes it so.**	*<USER >will make a bad PM. Please don't turn this into a race war. Please notice that he is a terrible politician*

Table 3: 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 4: Label counts for each set. For Set 3, the value in parentheses shows the count of identity-based hostility that comes from intersectional labels.

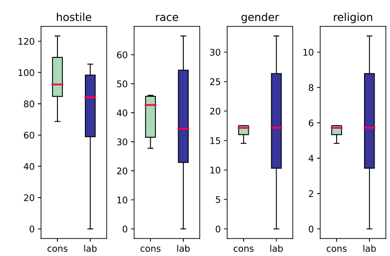

Figure 2: Comparing political party-based differences in the amount and type of hostility received

Table 4 shows the statistics of each set. The top 6 rows present the frequency of each label for each set. In general, non-hostile tweets account for the largest proportion, followed by no identity and race-based hostile tweets. Set 3 includes intersectional labels, but there are only 43 of these, of which 5 target both religion and gender, 21 target both religion and race, and 17 target both 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 tasks, which belong to substantial agreement (Artstein and Poesio 2008). This suggests selecting annotations based on confidence scores helps to improve the quality of the dataset.

Figure 2 and Figure 3 show the amount and type of hostile tweets MPs receive based on their political party and identity group. The horizontal pink (Figure 2) and black (Figure 3) lines mark the mean value for each group. On average, MPs belonging to the Conservative Party receive more race-based hostility.For gender and religion-based hostility, on average, MPs from both parties receive a similar amount of hostility. However, there are some MPs from the Labour Party who receive more identity-based hostility than others (e.g. Diane Abbott, David Lammy, etc.). Due to only one MP in our study belonging to SNP, we do not include SNP in this comparison.

In Figure 3, as expected, we see that while male (M) MPs receive more hostile tweets, female (F) MPs are disproportionately subjected to gender-based hostility. Similarly, non-white (NW) and non-Christian (NC) MPs face significantly higher levels of general and both 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) wherein different minority groups intersect with and reinforce each

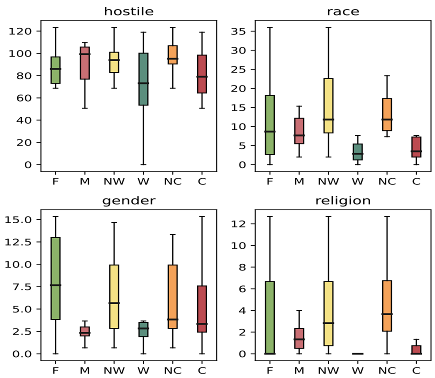

Figure 3: Comparing identity-based differences in the amount and type of hostility received

Figure 4: Top 100 BOW unigrams associated with hostile and non-hostile tweets. The larger the text size, the higher the Pearson correlation coefficient r, and vice versa.

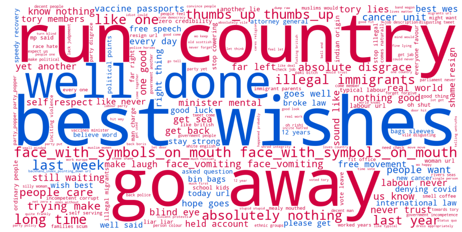

Figure 5: Top 100 BOW bigrams associated with hostile and non-hostile tweets. The larger the text size, the higher the Pearson correlation coefficient r, and vice versa.

other.

Data Characterisation

Linguistic Analysis

To investigate the difference between both the use of language and the content of hostile and not hostile tweets, we conduct a comparative linguistic analysis. We use the Bag of Words (BOW) model and Linguistic Inquiry and Word Count (LIWC) Dictionary (Boyd et al. 2022) to identify linguistic patterns. Then a univariate Pearson's correlation test is used to identify which of these linguistic patterns significantly correlate with hostile and not hostile tweets respectively. During pre-processing, URLs and user @mentions in the tweets are replaced with special tokens (<URL >and <USER >, respectively), and stop words are removed using NLTK (Bird, Klein, and Loper 2009).

BOW We begin by employing the Bag-of-Words (BOW) model to represent each post as a TF-IDF weighted distribution over a vocabulary of the 3,000 most frequent unigrams and bigrams. To visually highlight the differences in BOW features associated with hostile and not hostile tweets, we create word clouds (see Figure 4 and Figure 5 for unigrams and bigrams, respectively).

First of all, we observe that hostile tweets are characterised by negative and abusive words and phrases such as "scum", "vile", "incompetent", "evil", "stupid", "nothing good" and "absolute disgrace". These tweets are mostly posted to vent anger or dissatisfaction at politicians, ranging from questioning their abilities and distrusting their policies to insulting their personal traits. There are also emojis like "face_with_symbols_on_mouth," and "face_vomiting," which represent the use of profanity and disgust. Here is an example from our dataset:

> Tweet 1: *<USER>Some in the Cabinet are incompetent. Some are corrupt. Some are evil. You are all three. You stand for nothing that is good. I have nothing but contempt and disgust for you.*

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

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

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

> Tweet 3: *<USER >Just what would you do about the illegal immigration welcome them with open arms just 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 4). However, they are correlated with words and phrases such as "excellent", "recovery", "best wishes" and "well done". These suggest that not hostile tweets often contain appreciative and positive emotions towards MPs. Some phrases such as "asked questions" and "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!*

LIWC Each tweet is also characterised using psycholinguistic categories from the LIWC 2022 dictionary (Boyd et al. 2022). Table 5 presents the top 10 LIWC categories most strongly correlated with hostile and not hostile tweets.

Similar to BOW, We find that (unsurprisingly) hostile tweets tend to have a negative tone (*tone_neg*) and convey

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 5: Top 10 LIWC categories associated with hostile and not hostile tweets sorted by Pearson's correlation (r) between the normalised frequency and the labels. All correlations are significant at p <.001, two-tailed t-test.

negative emotions like anger and sadness (*affect*). They contain more assertive and judgemental 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 example shows race-based hostility towards the MP from the dataset:

Tweet 5: *<USER >What about black violence! Your just another race divider. Marxists like you have ruined this country and divided it further!*

We notice that non-hostile tweets are highly correlated with the tone marker (*tone*) in general and, specifically, a positive tone (tone_pos). The tweets are polite, adhere to social norms (*prosoc* and *polite*), and are more communicative (*comm*), 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: *<USER >please consider the effect of this new scheme before putting it into action. We have lost jobs and suffered a lot during covid. The UK economy will not recover. We need to think of our next steps very carefully*

Topic Analysis

We perform topic analysis using BERTopic (Grootendorst 2022) after removing stop words with NLTK (Bird, Klein, and Loper 2009). Because of the dominance of MP names and profanity, the topics are rather unclear. Table 6 shows the top 6 topic groups and their representative words after removing these terms. The topics relate to major events and issues in the UK, such as 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: *<USER >What planet do you live on in your head ? You haven't saved the day. Fuel is still +40% on years average. Energy bills are increasing 50%. We're all still fucked. Make it make sense !!!*

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 6: Topic groups and representative words derived from BERTopic

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 negative terms from representative words such as "scum", "johnsonout". Here is an example of hostile tweets mentioning the Conservative Party:

Tweet 7: *<USER ><USER >this you? You ludicrous pork Hay-bale. You bin bag full of custard. #ToryScum #ToryCriminalsUnfitToGovern it's true. <USER >shouldn't apologise you scum*

Most topics appear in the same proportion in hostile tweets as they do in non-hostile tweets. The exception is "illegal immigration" which appears twice as much in hostile tweets than in non-hostile tweets. Figure 6 shows the proportion of topic-related tweets belonging to identity-based hostility. Looking at the distribution of "illegal immigration" and "Brexit", they appear mainly in race-based hostile tweets. While the "Conservative party" and "Labour party" topics contribute to race-based hostile tweets, they appear much 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 wide range of topics. This topic characterisation means that the dataset could eventually be used for analysis and comparison of hostility in relation to different issues over time.

Online Hostility Detection

Given a text snippet, we define online hostility detection as two classification tasks: (1) binary hostility identification (if a tweet contains hostility or not) and (2) multi-class classification for classifying if a tweet contains one of the four identity-based hostility types (i.e., religion, gender, race, none) or no hostility at all. For multi-class classification, we compare two classification methods, namely flat and two-level hierarchical classification.

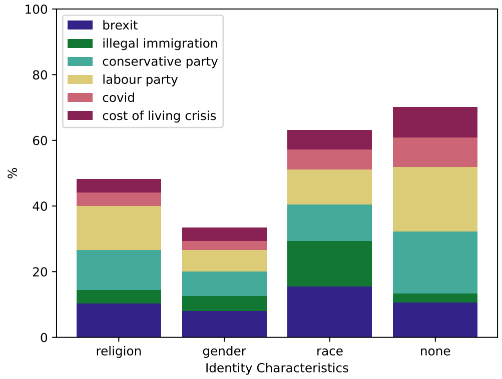

Figure 6: Proportion of topic-related tweets belonging to each identity characteristic label

Flat Classification Given a tweet, models classify it as religion-based hostility, gender-based hostility, race-based hostility, hostility with no protected identity characteristics (none-hostility) or not hostility.

Two-level Hierarchical Classification Given a tweet, the first classifiers identify if it is hostile or not. Then, the second classifiers classify the identity types of the hostile tweets (religion, gender, race or none) based on predictive results from the first classifiers.

Predictive Models

We use three PLMs for binary hostility identification and multi-class classification. Also, we evaluate two widely used LLMs on identifying hostile tweets as well as their targeted identity types.

BERT We fine-tune Bidirectional Encoder Representations from Transformers (BERT) (Kenton and Toutanova 2019) by adding a classification layer with softmax activation function on top of the [CLS].

RoBERTa We fine-tune RoBERTa (Liu et al. 2019) model in a similar way to BERT.

RoBERTa-Hate Similarly, we fine-tune a domain adaptation model, RoBERTa-Hate[3], which is trained on 13 different hate speech datasets in the English language including political content.

LLaMA-3-8B (LLaMA) We use the instruction tuned LLaMA 3 8B model through the Hugging Face platform[4]. We provide the model with a sequence of texts and a prompt with a task description to guide its output.

GPT-3.5 (GPT) Similarly, we use the GPT-3.5 model[5] with an API key providing texts and corresponding prompts.

[3]https://huggingface.co/cardiffnlp/twitter-roberta-base-hate-latest

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

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

Model	Accuracy	Precision	Recall	F1
		Set 1		
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
		Set 2		
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 w/ 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 w/ Def.	70.69±1.27	71.90±1.29	71.85±1.28	70.69±1.27

Table 7: Accuracy, Precision, Recall and macro F1-Score (F1) for binary hostility identification (± std. dev.). Best results are in bold.

Experimental Set-up

Tweets are pre-processed where URLs and user @mentions are replaced with special tokens (<URL >and <USER >, respectively). For evaluation, we report average Accuracy, Precision, Recall and macro F1 over 5 folds with standard deviations. Details of model parameters can be found in the appendix.

For LLMs, we input the prompt to specify the task for binary hostility identification and multi-class flat classification: (1) *Classify the tweet as hostile or not hostile* and (2) *Classify the tweet into religion-based hostile, gender-based hostile, race-based hostile, other-hostile or not hostile* with (**LLaMA w/ Def.**, **GPT w/ Def.**) or without definitions (**LLaMA**, **GPT**) of each category (see Table 3 for definitions). For 2-level hierarchical classification, we input the prompt based on the outputs from the binary hostility identification: *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.

Results

Binary Hostility Identification Table 7 presents the predictive results of all models on binary hostility identification using Set 1 and Set 2. We exclude Set 3 because the intersectional labels in identity type annotation do not affect binary labels (i.e., hostile or not hostile). Overall, RoBERTa-Hate on Set 2 achieves the best performance among all models, reaching a macro F1 score up to 73.03 (in bold). Then, we observe that models trained on Set 2 achieve better performance than those trained on Set 1 (e.g., 68.86 vs. 67.44 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

Model	Accuracy	Precision	Recall	F1
Flat Classification				
Set 1				
BERT	61.93±2.03	31.42±2.14	28.31±2.36	28.18±2.91
RoBERTa	63.13±0.35	24.40±0.32	26.03±0.20	20.79±0.21
RoBERTa-Hate	64.01±1.41	45.81±6.34	34.74±3.90	36.67±4.54
Set 2				
BERT	65.93±1.68	37.12±1.62	34.77±2.79	35.07±2.72
RoBERTa	67.26±1.05	33.01±5.96	32.25±3.95	32.04±4.68
RoBERTa-Hate	**69.64±1.24**	50.82±7.61	41.66±4.90	43.31±5.65
LLaMA	55.48±2.29	42.01±5.38	46.60±8.18	38.09±5.69
LLaMA w/ Def.	61.33±1.41	48.61±3.58	52.39±3.68	47.02±3.25
GPT	65.57±1.77	52.64±4.21	55.02±5.21	52.65±4.45
GPT w/ Def.	65.03±1.71	53.70±2.71	56.89±2.69	54.74±2.22
Set 3				
BERT	65.03±2.31	19.45±3.80	18.43±2.03	18.23±2.66
RoBERTa	66.66±1.72	20.30±3.68	20.43±3.44	20.08±3.64
RoBERTa-Hate	68.70±2.64	28.55±8.12	23.51±3.70	23.98±4.49
Hierarchical Classification				
Set 1				
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
Set 2				
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 w/ 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 w/ Def.	64.43±1.52	54.15±3.42	**60.02±3.11**	**55.98±3.08**
Set 3				
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 8: Accuracy, Precision, Recall and macro F1-Score (F1) for hostility type classification in flat (top) and 2-level hierarchical ways (bottom) (± std. dev.). Best results are in bold.

for RoBERTa vs. 68.84 F1 for RoBERTa-Hate).[6]

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

[6]We also evaluate Set 1 and Set 2 on the same test set with the same labels (we exclude Set 3 since 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 and 74.10 vs. 73.26 accordingly).

Multi-class Hostility Classification Table 8 presents the results of all models on multi-class hostility type classification using three sets of data in flat (top) and 2-level hierarchical (bottom) ways. Among all PLMs, the best performing model is RoBERTa-Hate on Set 2 in the flat classification method with an F1 score of 43.31 (in bold). Similar to the binary hostility identification, models in Set 2 achieve the best predictive results compared with the same models trained on other sets (e.g., 32.04 F1 for RoBERTa in flat classification, 33.81 F1 for RoBERTa-Hate in hierarchical classification), followed by Set 1 (e.g., 20.79 F1 for RoBERTa in flat classification, 31.21 F1 for RoBERTa-Hate in hierarchical classification). The domain adaptation model, RoBERTa-Hate, outperforms the vanilla RoBERTa model with a larger difference compared to binary hostility identification (e.g., +4.17 F1 vs. +11.27 F1 on Set 2 in binary hostility identification and in multi-class hostility classification using the flat method). Additionally, BERT and RoBERTa exhibit comparable performance, with BERT sometimes outperforming RoBERTa and vice versa depending on different settings.

Furthermore, the flat classification method outperforms the 2-level hierarchical classification method (e.g., 36.67 F1 in flat classification vs. 27.65 F1 in hierarchical for RoBERTa-Hate on Set 1). This may be explained by the fact that after training the first classifier, predictive errors introduce more noise in the hierarchical method. Also, the second classifier is trained on a smaller data set (non-hostile tweets are excluded) compared with that in the flat classification.

Similar to the hostility identification task, we only apply LLMs on Set 2. First of all, GPT w/ Def. in hierarchical classification 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., +5.85 F1 for LLaMA in flat classification, +5.45 F1 for GPT in hierarchical classification). Furthermore, LLMs in flat and hierarchical settings achieve comparable results. This might be explained by the fact that we use zero-shot classification, where they do not rely on our training data. Future research may extend it to few-shot learning.

Conclusion

This work focuses on investigating online hostility towards UK politicians. We develop an English dataset of 3,320 tweets, which are manually annotated with hostility as well as their targeted identity characteristics: religion, gender, and race. Also, we conduct 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. Also, our findings suggest that there is a general lack of trust in MPs in the UK. Furthermore, we evaluate various PLMs and LLMs on binary hostility identification and multi-class targeted identity type classification in flat and hierarchical ways. 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 political contexts specific to the UK.

Acknowledgments

The 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

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*, 5–16.

Agarwal, P.; Sastry, N.; and Wood, E. 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, 26–37.

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

Artstein, R.; and Poesio, M. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational linguistics*, 34(4): 555–596.

Bakir, M. E.; Farrell, T.; and Bontcheva, K. 2024. Abuse in the time of COVID-19: the effects of Brexit, gender and partisanship. *Online Information Review*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F. M. R.; Rosso, P.; and Sanguinetti, M. 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*, 54–63.

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

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47.

Carson, A.; Mikolajczak, G.; Ruppanner, L.; and Foley, E. 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.

Collignon, S.; and Rüdig, W. 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.

Enock, F.; Johansson, P.; Bright, J.; and Margetts, H. Z. 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)*.

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

Esposito, E.; and Zollo, S. A. 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.

Farrell, T.; Bakir, M.; and Bontcheva, K. 2021. Mp twitter engagement and abuse post-first covid-19 lockdown in the uk: White paper. *arXiv preprint arXiv:2103.02917*.

Fleiss, J. L. 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/.

Fortuna, P.; Soler, J.; and Wanner, L. 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*, 6786–6794.

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

Goodman, S.; and Locke, A. 2024. Supporting and challenging hate in an online discussion of a controversial refugee policy. *Discourse Studies*, 14614456231225448.

Gorrell, G.; Bakir, M. E.; Greenwood, M. A.; Roberts, I.; and Bontcheva, K. 2019. Race and Religion in Online Abuse towards UK Politicians: Working Paper. *arXiv preprint ArXiv:1910.00920 [Cs]*.

Gorrell, G.; Bakir, M. E.; Roberts, I.; Greenwood, M. A.; and Bontcheva, K. 2020. Which politicians receive abuse? Four factors illuminated in the UK general election 2019. *EPJ Data Science*, 9(1): 18.

Gorrell, G.; Greenwood, M.; Roberts, I.; Maynard, D.; and Bontcheva, K. 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.

Grimminger, L.; and Klinger, R. 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*, 171–180. Online: Association for Computational Linguistics.

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

Gross, J.; Baltz, S.; Suttmann-Lea, M.; Merivaki, L.; and Stewart III, C. 2023. Online Hostility Towards Local Election Officials Surged in 2020. *Available at SSRN 4351996*.

Guellil, I.; Adeel, A.; Azouaou, F.; Chennoufi, S.; Maafi, H.; and Hamitouche, T. 2020. Detecting hate speech against politicians in Arabic community on social media. *International Journal of Web Information Systems*, 16(3): 295–313.

Håkansson, S. 2024. Explaining citizen hostility against women political leaders: A survey experiment in the United States and Sweden. *Politics & Gender*, 20(1): 1–28.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 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)*, 3309–3326.

Hua, Y.; Naaman, M.; and Ristenpart, T. 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*, 1–13.

Jafri, F. A.; Siddiqui, M. A.; Thapa, S.; Rauniyar, K.; Naseem, U.; and Razzak, I. 2023. Uncovering Political Hate Speech During Indian Election Campaign: A New Low-Resource Dataset and Baselines. *arXiv e-prints*.

Jahan, M. S.; and Oussalah, M. 2023. A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing*, 126232.

Jin, M.; Mu, Y.; Maynard, D.; and Bontcheva, K. 2023. Examining temporal bias in abusive language detection. *arXiv preprint arXiv:2309.14146*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2.

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

Kuperberg, R. 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.

Kwarteng, J.; Perfumi, S. C.; Farrell, T.; Third, A.; and Fernandez, M. 2022. Misogynoir: challenges in detecting intersectional hate. *Social Network Analysis and Mining*, 12(1): 166.

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

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8): e0221152.

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

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Mollas, I.; Chrysopoulou, Z.; Karlos, S.; and Tsoumakas, G. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6): 4663–4678.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4296–4305.

Rieger, D.; Kümpel, A. S.; Wich, M.; Kiening, T.; and Groh, G. 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.

Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P. C.; Carvalho, J. P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A. V.; and Trancoso, I. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93: 333–345.

Röttger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–190.

Scott, J. 2019. Women MPs say abuse forcing them from politics. *BBC News*.

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

Southern, R.; and Harmer, E. 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.

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

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

Wang, C.-C.; Day, M.-Y.; and Wu, C.-L. 2022. Political hate speech detection and lexicon building: A study in taiwan. *IEEE Access*, 10: 44337–44346.

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

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84.

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.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 75–86.

Dataset Availability

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

- **Findable:** Our dataset will be published in the Zenodo dataset-sharing service with a unique digital object identifier. For now it can be found at https://anonymous.4open.science/r/ohtukmp-21D8.
- **Accessible:** Original tweets can be retrieved using their tweet IDs via the standard Twitter API[7]
- **Interoperable:** The readme file in the repository explains the dataset structure and the description of each column in the CSV data file. CSV datasets are easily imported and processed by most widely used data processing tools.
- **Re-usable:** Our dataset can be re-used by anyone who has Twitter developer account.

Model Parameters

For BERT, we use base uncased model, and for RoBERTa, we use base model. The maximum sequence length is set to 256 tokens, and the batch size is set to 32. We run all models using a 5-fold cross-validation method where 4-fold data is used for training and 1-fold data is used for testing. We split 4-fold data into training and validation sets with a ratio of 9:1. We use Cross Entropy Loss as the training loss function with the AdamW optimizer. For flat classification, we train using a learning rate of lr = 3e-6, while hierarchical classification is trained with lr = 5e-5. During training, we choose the model with the smallest validation loss over 15 epochs. All models are trained on an NVIDIA A100 GPU. For all experiments of LLMs, the temperature is set to 0.1

Confidence Scores

Table 9 shows the meaning and explanation of different levels of confidence scores (1-5) used in the annotation task.

Score	Meaning	Explanation
5	Extremely confident	I'm certain without a doubt.
4	Fairly confident	I'm confident, but there might be a small chance other annotators may label it as a different category.
3	Pretty confident	I'm pretty sure, but there might be a high chance other annotators may label it as a different category.
2	Not confident	I'm not sure; it could belong to this or another category.
1	Very low confidence	I'm really unsure; it might belong to another category instead.

Table 9: Confidence Scores used in the annotation task.

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