Offensive Language in Reactions to Public Figures in Polarised Discourse Online

Maciej Kulik[1], Katarzyna Budzynska[1], He Zhang[1], Marie-Amélie Paquin[2], Barbara Konat[3]

[1] Laboratory of The New Ethos, Warsaw University of Technology, Warsaw, Poland

[2] Panthéon-Sorbonne University, Paris, France

[3] Adam Mickiewicz University, Poznan, Poland

Abstract. Offensive language affects contemporary societies by hindering communication and increasing polarisation. In this study, we apply computational linguistics to investigate offensive reactions to public figures in climate debate on Twitter across their roles and popularity. We also use sentiment analysis to inspect the accuracy of lexical criteria in detecting negative attitudes and examine the types of social media users based on the frequency of offensive content in their posts. With an in-depth, large-scale corpus analysis comprising one million words, we demonstrate that frequent offensiveness in responses to politicians relatively rarely expresses personal attacks, and the popularity of public figures does not always come together with the highest density of offensive reactions. We also show that the majority of users publish predominantly non-abusive posts. The study sets foundations for strategies to be employed to reduce polarisation that constitutes a threat to deliberative democracy.

Keywords. offensive reactions, public figures, climate change, social media, large-scale corpus analysis, AI-based technology of data analytics

1 Introduction

The aim of this paper is to describe the use of offensive language in online reactions to public figures who took part in the polarised climate debate. We follow the account of Hobolt et al. (2021), who take the differentiation of the in-group from the out-group that leads to in-group favorability and out-group denigration to be a distinctive feature of polarisation. The impact of offensive language on polarisation thus understood has already been proven (Simchon et al. 2022; Vasist et al. 2023), and the relevant studies provide evidence for the claim that the climate debate is polarised (Falkenberg et al. 2022; Schäfer and North 2019). Thus, the analysis of offensive language in online discussions on climate change is likely to unveil activities of the public figures that trigger offensive reactions. At the same time, we establish foundations for the

development of tools such as online moderators as well as for the enforcement of policies aimed at reducing polarisation. Such strategies are needed, as polarisation is detrimental to productive communication, which lies at the core of deliberative democracies (Bächtinger and Parkinson 2019; Parkinson and Mansbridge 2012), and negatively affects the quality of individual lives in modern societies (Stahel and Baier 2023).

For the analysis, we selected Twitter discussions during the 2021 United Nations Climate Change Conference (COP) due to its high media profile and influence on public discourse. The analysis of the use of offensive language required us to examine this issue not as a feature of public figures' own rhetoric but as a social phenomenon. Thus, we decided to analyse users' reactions rather than tracking offensiveness in posts published by public figures themselves which is rare in their case anyway.

In this study, we use the core method of computational linguistics: *quantitative* large-scale corpus (55,927 posts of over 1 million words) analysis combined with *qualitative* analysis that explores specific utterances. The adopted computational approach defines offensive language using lexical criteria, i.e., based on the presence of specific words. Thus, to identify offensive reactions, we draw from Gorrell et al. (2020) lexicon, mainly extracted from the Hatebase repository (https://hatebase.org/). While applying this method, we cannot consider the context-dependent semantic content of utterances, as it is beyond the linguistic surface captured in the lexicon. To inspect the accuracy of the lexical criteria in detecting personal attacks, we apply sentiment types, defined according to Camacho-Collados et al. (2022) and Loureiro et al. (2022) as positive, negative, or neutral. This helps us to distinguish between reactions classified as offensive based on specific words and those deemed offensive in a context-dependent, semantic sense, i.e., arising from the sentence's insulting meaning.

The adopted double quantitative-qualitative approach allows us to provide a detailed description of the use of offensiveness as it appears in the climate debate. First, we inspect the distribution of offensive reactions across the groups of public figures to establish whether their ordering in terms of both range and intensity of offensive language aligns with our informed supposition: (from more to less: politicians, contrarians, businessmen, and activists). Second, we investigate whether the popularity of public figures and the frequency of offensive responses to their posts are strictly correlated. Third, we examine the correlation between sentiments and offensiveness to verify whether the frequency of positive sentiment in abusive responses is indeed, as it seems obvious, minimal. Finally, we test the accuracy of the conventional wisdom about social media and inspect whether the majority of the users publish offensive reactions.

This paper is structured as follows. First, we present related work and summarise our methodology. Next, we present and examine the results of our analysis. Finally, we summarise our observations and discuss future work.

2 Related Work

In this paper, we combine quantitative and qualitative methods to inspect a large corpus of data regarding the use of offensive language, understood according to the lexical criteria, in polarised discourse on climate change. To the best of our knowledge, other studies either apply only one of these methods, focus on medium- or small-sized corpora, or address offensiveness defined in distinct terms.

The purely qualitative studies, such as Burke et al. (2020), Chojnicka (2013), Leezenberg (2015), Määttä (2023), Retta (2023), Terkourafi et al. (2018), and Wodak et al. (2020) provide an analysis of the character of speakers' rhetoric and their possible covert goals. Yet, such an approach suffers from an inability to support its claims with large-scale, unbiased evidence. For instance, Leezenberg (2015) shows that the far-right populist discourse sometimes presents a legitimate criticism of religion instead of being openly racist and xenophobic. Also, Määttä (2023) and Retta (2023) explore the role of implicitness in the transmission of hateful content. Wodak et al. (2020) show that the open use of racist and sexist remarks, which they label as 'shameless normalisation', is one of the common strategies used by populists. Further, Burke et al. (2020) demonstrate that such open manifestations of far-right extremism as Holocaust denial and encouragement to 're-open' the concentration camps can still be found in the online communities of far-right supporters. Concerning other domains, Chojnicka (2013) analyses qualitatively the conflict in terms of offensive, aggressive, and prejudicial language in debates between Russian and non-Russian MPs in the Latvian parliament. Terkourafi et al. (2018) describe offensive language in the context of an ethnic conflict.

On the other hand, purely quantitative accounts, such as Bentivegna and Rega (2022), Hong et al. (2023), Park et al. (2021), Rösner and Krämer (2016), Vergani et al. (2022) provide the possibility to identify the large-scale tendencies. Yet, not all the details can be described without the analysis of specific instances. The account of Park et al. (2021) is close to ours in terms of the content, as the study analyses offensive language targeted against the climate activist Greta Thunberg. The comparison of the frequency with which sexist, ableist, and ageist comments are displayed is yet limited to these types of offensiveness. Similarly, Hong et al. (2023) and Vergani et al. (2022) analyse only anti-Chinese and anti-Asian speech. Despite that Bentivegna and Rega (2022) and Rösner and Krämer (2016) inspect all types of offensive language, their experimental method does not apply to real-life examples.

Some of the studies that combine quantitative and qualitative approaches explore large corpora (Carvalho et al. 2023; Eschmann et al. 2020; Rega et al. 2023;Riedl et al. 2022; Schroeter 2018). Yet, contrary to our study, these accounts are devoted to one of the sub-types of offensive language. Schroeter (2018) is outstanding in terms of the corpus size. Its purpose is to inspect German Nazi vocabulary. Rega et al. (2023) provide a typology of uncivil language into the intolerant, that is targeted against groups, and the impolite, that encompasses the other instances of offensiveness. Yet, their conclusions, which point out the role of the far-right actors in spreading the phenomenon, concern the specific use of intolerant language rather than

offensive speech in general. The scope of Riedl et al. (2022) is restricted to the Anti-Semitic expressions. The study of Carvalho et al. (2023) is devoted to the racist, xenophobic, and homophobic language, and the manifestations of offensiveness that are not associated with specific communities are left aside. The focus of Eschmann et al. (2020) is not on the use of abusive words but on the word 'microaggression'. Thus, the study analyses people's opinions but not the phenomenon itself. Rasulo (2021), who explores a medium corpus of data, analyses a broad spectrum of verbal aggression. However, the study examines traditional media (newspaper headlines) that are targeted only against one public figure.

Among the phenomena related to our focus are impoliteness and hate speech. The conceptual foundations of the study of impoliteness can be traced back to Brown and Levinson (1987) and Leech (1983), and were further developed in Culpeper's (2011) framework. Within this line of research, impoliteness is defined as an act that threatens the speaker's 'face', i.e., their self-image within the context of communication. Recently, both quantitative and qualitative approaches have been applied to this issue in studies such as Alvanoudi (2024), Graham and Hardaker (2017), and Teneketzi (2021). Similarly, hate speech has been the subject of corpus-based research integrating both quantitative and qualitative methods, as evidenced in works such as Davidson et al. (2017), De Gibert et al. (2018), Jaszczyk-Grzyb et al. (2023), Karapetjana et al. (2023), Lepoutre et al. (2023), and Parvaresh (2023). However, the focus of these accounts differs from ours, as impoliteness and hate speech, unlike offensive language as we define it, do not necessarily involve the use of words we later identify as abusive.In contrast to the discussed works, this study employs quantitative analysis to identify key patterns in the distribution of offensive language, defined here as the use of abusive words in responses to selected groups of public figures categorised by their social roles and levels of popularity. The qualitative method allows us to describe discrepancies between various types of offensive language in terms of the words used (slur or rude) or sentiment type (positive vs negative). Thus, the benefits of the quantitative and qualitative approach are combined in answering our research questions and providing their empirical, large-scale grounding. The applied definition is construed in lexical terms, i.e., we identify offensive language based on specific words drawn from the lexicon, as opposed to the context-dependent semantic criteria that involve insulting meaning of the sentence.

3 Methodology

The method used in this paper is large-scale corpus analysis. We work with a data set that contains over one million words. The application of our method required three steps: data collection, data annotation and data analysis (see Subsections below). Firstly, we collected tweets published in reaction to public figures posting on climate change. Secondly, we annotated the corpus according to offensive language use and sentiment distribution. Thirdly, we visualised and interpreted data (see Figure 1).

This approach scales up a method proposed in (Pereira-Fariña et al. 2022) that uses manual annotation of ethos in order to analyse discursive strategies in debates on cultural heritage statistically and further extends a method proposed in (Landowska et al. 2024) that combines manual annotation of arguments and automatic annotation of moral values in order to build data analytics to investigate moral arguments in discourse. The method proposed in this paper is fully automated both in terms of detection of offensive language (see action 1 in Figure 1) as well as in terms of detection of sentiment (see action 2 in the figure). This means that it is generalisable to any amount of data—the selection of one million word dataset was constrained only by the research questions we investigate in this paper (see Section 4). Thus, in principle, it can be applied to other studies of offensive language on *any* amount of data, *any* discourse topic, *any* public figures, *any* social media platform, and so on.

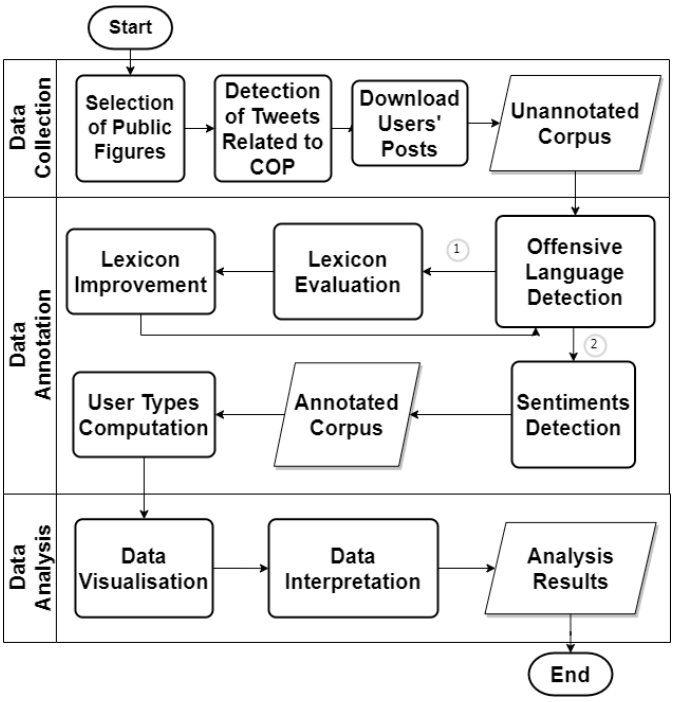

Figure 1: Methods of large-scale discourse analysis.

3.1 Data Collection

The goal of our research was to analyse offensive language defined by lexical criteria, which had to be established for a specific language. Thus, we limited the scope of our corpus to the content in English due to the popularity of online communication. To specify the scope of our research further, we selected Twitter discussions during the 2021 United Nations Climate Change Conference, which was also known as the 26th Conference of Parties (COP26) and took place in Glasgow from 31 October to 13 November 2021. After the event had been selected, we chose the public figures who (1) participated in the COP event, (2) published posts on climate change during COP, and (3) allowed users to post reactions. These requirements significantly

constrained the number of public figures selected for examination. For instance, despite that both Bill Gates and Jane Fonda were active on Twitter (now X) during COP26, we did not include them in the analysis, as the former disallowed users to comment and the latter did not take part in the event. Even though we have not taken into account any cultural or geographical criteria of selection (except the language of the online activity), most of the figures in our sample originate from North America. This is due to the fact that the COP26 participants, who were active on Twitter while the event took place, typically come from this region. The notable exceptions include Greta Thunberg and Elisabeth Wathuti. Including the former might be considered questionable, as she took part only in the initial part of the event. However, we decided to include her tweets, as she highly influenced the climate debate. Similarly, even though Donald Trump did not take part in the COP26 conference, we included him in the corpus due to his influence on the public discourse. As Trump was inactive during COP26, in his case, we take into account the 2019 United Nations Climate Change Conference (COP25), which took place from 2 to 13 December 2019 in Madrid.

After the selection, we grouped the public figures with regard to their *social roles* into: politicians, climate change contrarians, businessmen, and climate activists, as well as according to their *popularity* measured as the number of followers into: high, medium, and low number of followers, where high means public figures with over 10 million of followers, medium means public figures with between 1 and 10 millions of followers, and low means public figures with less than 1 million of followers (see Table 1).

Table 1: Public figures to whom we consider offensive reactions, distinguished by: (1) their social roles; and (2) their popularity.

Reactions to:	**Politicians**	**Contrarians**	**Businessmen**	**Activists**
High	Biden, Obama, Trump			
Medium	Johnson		Bezos, Bloomberg	DiCaprio, Thunberg
Low		Lomborg, Milloy		Wathuti

Next, we identified tweets published during COP26 (or COP25 for Trump) that were related to climate change by retrieving all tweets with the keywords ‘climate’ or ‘COP’. We then retrieved user posts published in reaction to these tweets. We extended the collection to the tweets published after the COPs to include delayed responses. For COP26, the reactions were collected from October 31 to November 26, 2021, and for COP25, from December 2 to December 27, 2019. The data thus compiled constitute a corpus of 55,927 posts and 1,030,714 words (see Table 3).

Table 3: Summary of the collected data.

	Tweets	Responses	Words in Responses
Politicians	47	41,794	757,882
Businessmen	32	2,421	45,074
Contrarians	373	1,639	256,645
Activists	50	10,073	7,813,831
Total	502	55,927	1,030,714

3.2 Data Annotation

In the related literature, offensive language is analysed under various labels, such as uncivil, impolite or intolerant language, hate speech, microaggression and so on. Here, instead of examining distinctions between various forms of offensiveness, we define the phenomenon through lexical criteria, i.e., through the presence of specific words. The term *offensive language* refers to language use in general, whereas *abusive words* and *abusive posts* refer to specific lexical items and online comments, respectively. To define abusive words, we use the lexicon developed by Gorrell et al. (2020), which is mostly extracted from the Hatebase repository that contains over 3,893 offensive words (https://hatebase.org/). From Gorrell et al. (2020), we draw 1,081 'slurs' and 131 'offensive words'. For the sake of terminological clarity, we changed the names respectively to: 'slur words' and 'rude words', which we jointly classify as 'abusive words'. Furthermore, we also adopt Gorrell et al. (2020)'s definitions of *slurs*: "insults, racist and homophobic slurs, as well as terms that denigrate a person's appearance or intelligence" and *rude words* ('words' in in their terminology): "words that don't in and of themselves constitute abuse, but worsen abuse when found in conjunction with a slur and become abusive when used with an identity term such as *black*, *Muslim* or *lesbian*" (see Table 2). Some words are classified as slurs when accompanied by the personal pronoun and as rude when they appear alone because the use of a pronoun indicates a personal attack that is typical to slurs (e.g., 'fuck you' vs 'fuck').

We identify slur and rude words based on the explicit list. This allows for the words recorded in the lexicon to be classified as abusive, even if they are not used in personal attacks. Thus, our method is lexical rather than semantic, as we identify the specific words as abusive, even if they do not convey insulting meanings.

This is necessary to be able to carry out a sentiment (as positive, negative, or neutral) analysis as defined by Camacho-Collados et al. (2022) and Loureiro et al. (2022) . This way, we can identify the sentiments expressed in utterances classified beforehand as offensive on the lexical criteria. This allows us to determine whether the utterances in question, apart from containing specific words that we classify as abusive, also function to express negative attitudes, i.e., whether they are offensive in the context-dependent semantic sense as well.

Table 2: Examples of slur and rude words.

slur words	**rude words**
idiot, moron, fuck you, buffoon, loser, coward	fuck, shit, crap, damn

Next, posts are classified as abusive if they contain slurs or rude words. The category is informative, as we observed that both slur and rude words constitute offensive reactions. As mentioned, the classification follows Gorrell et al. (2020), though we replace 'abusive text' with a slightly more descriptive term, 'abusive post'. Example (1) contains the slur word 'idiot', and Example (2) contains the rude word 'shit' (slur and rude words are underlined). Thus, in this approach, both posts are labelled as abusive. Example (3) shows that posts that combine several slur or rude words are also labelled as abusive. A post in Example (4) is non-abusive, as it contains neither a slur nor a rude word.

(1) @CondorLives: *You're an idiot and this graph shows why*
[Labelling in Gorrell et al. (2020): Slur; Abusive Text]
[Labelling in this paper: Slur (Abusive Word); Abusive Post]

(2) @JunkScience: *I wonder who thinks he even gives a shit about the Leftists 'canceling' him?!*
[Labelling in Gorrell et al. (2020): Offensive Word; Abusive Text]
[Labelling in this paper: Rude (Abusive Word); Abusive Post]

(3) @persecutorXXD: *The Chinese will mold you like an old piece of clay they will fuck you every which but loose and you won't even know it (just like 'rocket man' made you his personal bitch*)
[Labelling in Gorrell et al. (2020): Slur; Abusive Text]
[Labelling in this paper: Slur (Abusive Words): Abusive Post]

(4) @HuntingPelican: *This one could be different.*
[Non-Abusive Text/Post]

The lexical tagging allowed us to annotate user responses with these labels automatically. Then, the automated annotation was evaluated on a random sample of 600 responses by two human annotators. We observed a notably high level of inter-annotator agreement with a Cohen's Kappa of 0.7639 (Cohen 2018).

For sentiment analysis, we applied a roBERTa-base model (Camacho-Collados et al. 2022; Loureiro et al. 2022) that identifies positive, negative, and neutral sentiments. The model was initially trained on approximately 58 million tweets. It was further fine-tuned specifically for sentiment analysis using the TweetEval benchmark. This fine-tuning process enabled the model to achieve a notable performance metric of 71.3 M-Rec, indicating its effectiveness in sentiment analysis tasks within the context of Twitter data.

3.3 Procedure of analysis

An AI-based technology, Offensive Language Analytics (OLAn), was developed to analyse and visualise patterns in our data and provide insight into the use of offensive language in reactions to public figures in polarised discourse online. OLAn is a data analytics tool (Kelleher and Tierney 2018; Walker 2015) and is built upon technologies of argument analytics (Lawrence et al. 2016, 2017) and rhetoric analytics (Budzynska et al. 2024) online devices based on the Streamlit interface, developed to facilitate the analysis of data. Its purpose is to determine and display the distribution of offensive reactions understood as abusive words or abusive posts towards groups of public figures or individual public figures in the form of bar charts (see Figure 2). This allows us to analyse our corpus by comparing the proposed hypotheses with the visualised data and to draw conclusions about the use of offensive language. OLAn interface is Streamlit open-source Python Library (https://streamlit.io/). The tool is available for open access (https://newethos.org/technologies/). For more information about this technology, please refer to its User Manual available at this link. Please include link

For data analysis, we apply a mixed methods approach. The former is applied to statistics computed by OLAn for the offensive reactions, the distribution of sentiment, and the coverage of the types of users. With the latter, we describe the distinct character of offensive language and the sentiment expression in various contexts. This hybrid methodology is applied to each research question examined in the next section. Initially, we propose a hypothesis and

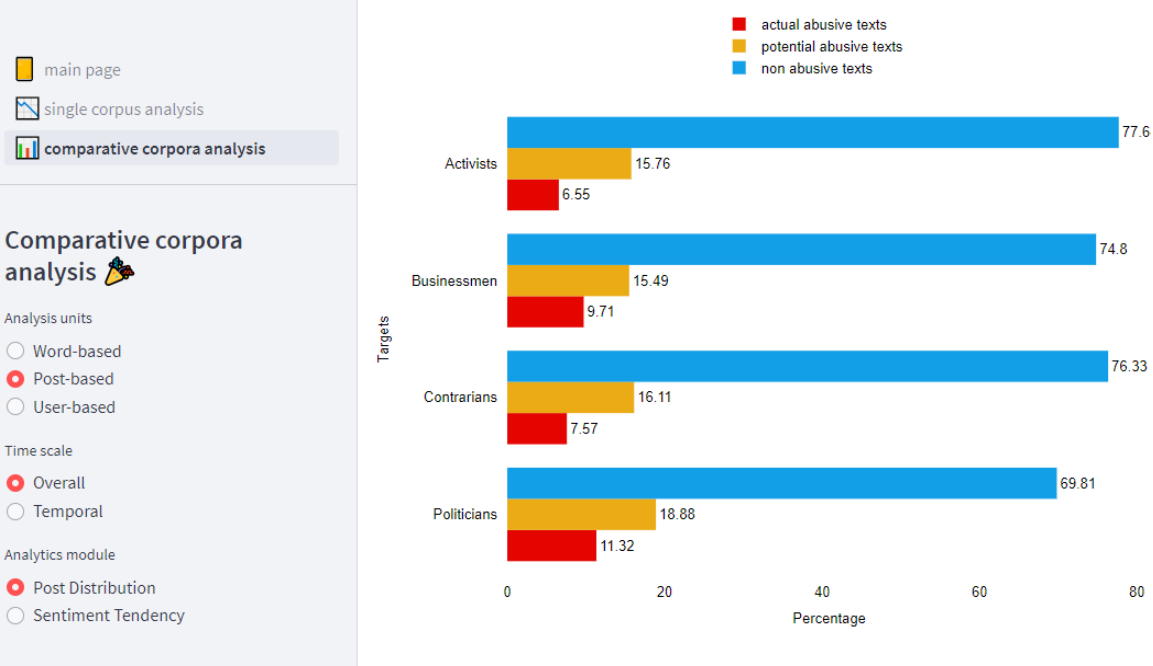

Figure 2: OLAn visualisation of the distribution of the abusive posts.

then test it with data visualised in OLAn. For instance, in Hypothesis 1, we propose the following ordering of social roles from the highest possibility of attracting offensive reactions to

the lowest: politicians, contrarians, businessmen, activists. While the data revealed slightly different ordering calculated as the frequency of abusive posts in reaction to what the public figures posted: politicians, businessmen, contrarians, activists, we investigate possible reasons behind this discrepancy. This leads us to change how we measure offensiveness of reactions, i.e., we turn to investigating it with the use of abusive words rather than abusive posts. As this reveals the higher intensity of offensiveness in responses to contrarians than to politicians, we used qualitative analysis in order to explore the distinct character of offensive language as posted in responses to these groups.

4 Results and Discussion

This section comprises the presentation and discussion of the results of our analysis.

4.1 Range and Intensity of Offensiveness Do Not Come Together

The supporters of political parties tend to be more loyal to their leaders under conditions of high polarisation (Druckman et al. 2013; Wilson et al. 2020). Thus, efforts to increase societal antagonisms could be incorporated deliberately as part of a political strategy. The use of offensive language is one of the means that political leaders could employ to achieve such a purpose, as it has been proven to have a polarising effect (Gervais 2020). Nevertheless, as Frimer and Skitka (2018) show, such a strategy is double-edged since the use of incivilities reduces the politician's public approval even within their political base. Thus, public figures can incite social media users to publish abusive responses without using profanity themselves. On the other hand, as argued by James et al. (2016), the character of politicians' public activity, which involves opposition to the interests and aspirations of certain groups within society, makes them particularly vulnerable to aggressive and intrusive behaviours. Meutia and Gulyanto (2022) compare the frequency of various forms of verbal harassment in responses to politicians, celebrities, and practitioners of other professions. They show that the most serious forms of mistreatment, such as character attacks and insults, are most frequent in responses to politicians. This usefulness of offensive language in approaching political goals and the higher general vulnerability of political leaders to aggressive behaviours inclined us to hypothesise that this group of public figures would also reach an excessive amount of verbal aggression in the context of the climate debate.

Furthermore, the controversial content posted by climate change contrarians facilitated the claim that they also face a relatively high amount of offensiveness, yet the grounds here are weaker than in the case of politicians. Next, Meutia and Gulyanto (2022) show that practitioners (a group similar to businessmen in terms of having primary occupation within the domains other than public discourse) attract rather moderate offensive reactions. Finally, even though the polarisation effects of what climate activists are claiming have not been investigated, the

sympathy that public opinion has for this group (Gaupp and Eker 2024; Leiserowitz et al. 2019; Thomas-Walters et al. 2025) formed the basis for a supposition that they would not encounter abusive responses as frequently as the other groups. This led us to formulate the following hypothesis:

Hypothesis 1: Politicians attract more offensive reactions than contrarians, contrarians more than businessmen, and businessmen more than activists.

Using OLAn, we then test this hypothesis against the collected and annotated dataset. The data revealed that the frequency of abusive posts is the following (see Table 4): politicians (12%), businessmen (10%), contrarians (8%), activists (7%), which is coherent with the assumptions that the highest frequency of offensive responses would belong to politicians and the lowest to activists. On the other hand, contrary to our expectations, businessmen triggered offensive reactions more often than contrarians. The ordering is thus partly consistent (politicians, activists) and partly reversed (businessmen, contrarians) compared to Hypothesis 1.

Table 4: Abusive posts and abusive words across the roles of the public figures (%). Abusive words mean the sum of slur and rude words. The baseline is calculated as an arithmetic mean of a given category for all public figures. The asterisk means the largest deviations from the baseline in each column.

	posts %		**words %**		
	abusive	**non-abusive**	**abusive**	slur	rude
Politicians	12*	88*	31	10	21*
Contrarians	8	92	28	12	16*
Businessmen	10	90	35*	17*	18
Activists	7	93*	25*	6*	19
Baseline	9	91	30	11	19

In order to explain this discrepancy between the empirical evidence and our hypothesis, we inspected the type of offensiveness in terms of abusive words, which reveals a ranking different from that of abusive posts: businessmen (35%), politicians (31%), contrarians (28%), activists (25%). We observe that only the last result is consistent with the hypothesis, while the results for businessmen, politicians, and contrarians are contrary to expectations. To deepen our analysis further, we considered the most offensive type of abusive words, that is, slurs, against the different groups which turned out to be ranked as follows: businessmen (17%), contrarians (12%), politicians (10%), activists (6%). Here, the results for the contrarians and the activists are consistent with the hypothesis, while the results for the businessmen and the politicians are contrary.

Several conclusions can be drawn at this point. First, the last position of the activists is

common to all ranking. Thus, regardless of the measure employed, this group is the category that attracts the lowest level of offensive reactions. In this respect, the hypothesis 1 was confirmed by the data. This observation suggests that the sympathy of the general public towards activists protects them from getting abusive responses. Second, the businessmen ranking differs from expectations: the frequency of abusive responses to their posts is higher than in reactions to the contrarians. In case of both abusive words and slur words, businessmen got offensive responses most frequently, surpassing not only the contrarians, but also the politicians. Thus, businessmen turn out to attract more offensive reactions than expected, regardless of the measure.

Third, the ranking of the political leaders in terms of abusive words and slur words is also contrary to Hypothesis 1. Regarding abusive words, they were ranked after businessmen. For slur words, they were third, falling behind not only businessmen but also contrarians. This result is surprising, especially given the top ranking position of this group in abusive posts. The apparent inconsistency can be explained by differences in the intensity of the offensive language. Rude words are more than twice as frequent as slur words in responses to the political leaders (21% vs 10%). In the case of the businessmen, the results differ only by 1% (18% vs 17%) and in the case of the contrarians by 4% (16% vs 12%). Thus, the proportion of slur to rude words in responses to politicians is much lower than in reactions to the businessmen and the contrarians. This means that even though the political leaders receive abusive posts most frequently, the intensity of offensive language in these posts is relatively low.

Finally, the contrarians received unexpectedly low results in abusive posts and abusive words, occupying third place in both cases. Nevertheless, in terms of slur words, they are placed second with a result of 4% lower than the businessmen and 2% higher than the political leaders. The proportion of slur to rude words is pretty high for contrarians (12% vs 16%) and indicates pretty high intensity of offensiveness.

The general patterns described above are representative for almost all of the individual results (see Table 5). Among activists, DiCaprio (3%) and Wathuti (1%) are the figures with the lowest frequency of abusive posts in our sample, and the result of Thunberg (8%) is still under the baseline (10%). This also holds for abusive and slur words. Thunberg's results are considerably, though not outstandingly, higher, with her result equaling the baseline (21%) in the last measure. In case of businessmen, offensive reactions appear to be twice as frequent in responses to Bloomberg than to Bezos. Among political leaders, Trump (14%) and Johnson (13%) appear to trigger more abusive posts than Biden (8%) and Obama (8%). The same ordering is observed in terms of abusive, slur and rude words. Possible factor that contributes to this effect is the right-wing populist rhetoric typically identified as the style of public activity of Trump and Johnson. The results of climate change contrarians do not differ from each other significantly.

Table 5: Abusive posts and abusive words across public figures (%). Letters in brackets indicate social role (P stands for politicians, B for businessmen, C for climate change contrarians, and A for climate activists). The asterisk means the largest deviations from the baseline in each column.

	posts %		words %		
	abusive	**non-abusive**	**abusive**	slur	rude
Trump (P)	14*	86	31	10	21
Johnson (P)	13*	87	37*	12	25*
Biden (P)	8	92	27	8	19
Obama (P)	8	92	21	7	14
Bloomberg (B)	14*	86	34	18*	16
Bezos (B)	7	93	36*	15	21
Lomborg (C)	7	93	31	12	19
Miloy (C)	8	92	27	12	15
Thunberg (A)	8	92	28	7	21
DiCaprio (A)	3	97	17	5	12
Wathuti (A)	1*	99	6*	2*	4*
Baseline	10	90	31	10	21

In order to explore further the differences between the specific character of offensive language in responses to politicians and responses to contrarians, we apply the qualitative method. We observed that many rude words, used in responses to politicians, convey a message that is not offensive at all but simply represent a manner of speaking, as in the following example:

(5) @StrayDog67: *Haha. It's the same fucking deal!!!*
[Reaction to: Lomborg; Abusive Word: Rude]

This kind of use contrasts with abusive posts that contain slur words i.e., they are not only vulgar but always used to attack a person:

(6) @CondorLives: *Fuck You! Fuck your administration! Russia and Putin own you! You have defiled the office of the President of the United States.*
[Reaction to: Trump; Abusive Words: Slur]

The lower frequency of slur words compared to rude words in responses to politicians indicates a lower intensity of offensive language. This means that in the case of politicians, the low intensity of offensive language was compensated by the wide range, while in the case of contrarians, its narrow range was compensated by the high intensity.

In summary, the distribution of online offensiveness across social roles of public figures confirmed that offensive reactions are most frequent in responses to politicians and least frequent in responses to climate activists. However, the intensity of offensive language in responses to political leaders, despite a wide range of abusive posts, turned out to be relatively low. On the other hand, offensive reactions in responses to contrarians are less frequent than expected, but

unexpectedly more intense in comparison to politicians. While the low intensity of offensiveness in responses to political leaders was compensated by the wide range, the narrow range in reactions to contrarians was compensated by the high intensity.

4.2 It Is Not Only About Popularity

According to the meta-analysis of Castano-Pulgarín et al. (2021), cyberhate is amplified by the use of social media. This effect is likely related, among others, to the anonymous character of online activities since anonymity, as argued by Mondal et al. (2017), has been proven to foster aggressive behaviours. Theocharis et al. (2020) argue that thefrequency of uncivil reactions is higher during periods of heightened public activity, such as electoral campaigns, and in responses to 'hard' topics such as economic and social issues. Chatzakou et al. (2017) examine the popularity of users who post offensive content, as measured by their number of followers, but find no observable correlation between the two factors. If such correlations could be found, they would shed the light on the relationship between increased publicity and offensiveness, the topic we also inspect in the present study. This holds even though we focus on the correlation between the popularity of public figures and the offensive content in responses to their posts, rather than on the popularity of the offensive users themselves. Mathew et al. (2019) provide more informative claims on the associated issue, as they prove hateful content to spread faster, farther, and reach a wider audience than non-abusive posts. As argued, aggressive language has been recognised to be prevalent on social media and the use of offensive content turned out to spread more than non-abusive communication. Therefore, in Hypothesis 2, we assume the strict correlation between the popularity of a public figure and the frequency of offensive language in responses to their posts. Such an assumption brings about certain ordering of the groups of public figures distinguished with respect to their popularity measured by the number of followers: having high (over ten million), medium (between ten and one million) and low (less than one million) number of followers.

Hypothesis 2: The more popular the public figure is, the more offensive reactions they will receive.

When the hypothesis is tested empirically, the data do not provide direct evidence of this ordering (see Table 6). In terms of abusive posts, public figures with a medium number of followers turned out to receive offensive reactions more frequently than public figures, followed by those public figures by a high number of users; however, the disparity is not significant (11% vs 10%). Moreover, the results are similarly ordered for abusive posts, although the disparity is more pronounced (34% vs 28%). The frequency of slur words is even more surprising: here, the highly popular group (9%) is outweighed not only by the medium-number group (11%) but even by the low-number group (10%). Admittedly, the results of each of the three are similar. Nevertheless, the ranking of slur words shows no clear relationship to popularity and appears

largely independent of it. Thus, it can be concluded that offensive reactions were the most frequent in responses to public figures followed by the medium number of users regardless of the measure employed. The hypothesis that assumed a strict correlation between public figures' popularity and offensiveness is therefore falsified.

Table 6: Abusive posts and words across the groups of public figures according to their popularity (%) in terms of the number of followers (high means over 10 million followers, medium means between 1 and 10 millions of followers, and low means less than 1 million of followers). The asterisk means the largest deviations from the baseline in each column.

	posts %		**words %**		
	abusive	**non-abusive**	**abusive**	slur	rude
High	10	90	28	9*	19
Medium	11*	89*	34*	11*	23*
Low	6*	94*	24*	10	14*
Baseline	10	90	31	10	21

The unexpected distribution could be a result of having a kind of authority (for example, the authority of a president of an influential country) that might protect the most popular public figures. Even though the offensive reactions are not necessarily targeted at the account's owner, a considerable number of them are likely negative responses to what they have posted, especially when the public figure triggers controversies and does not hold authority. In such cases, popularity could go hand in hand with authority, immunising a person against offensive reactions. In contrast, the medium-number group is popular enough to attract hate but its members do not have authority to be protected by. In other words, it can be the case that these public figures are controversial so that they are followed by many users, but as a result of these controversies they attract more offensive responses.

Another possible explanation for the unexpected result could be the uneven distribution of offensiveness within a given group. To check it, we inspected the results for individual public figures within each group (see Table 7). The results of highly popular public figures differ: while both Biden and Obama attract 8% offensive reactions, Trump attracts 14% abusive posts. Among the figures with a medium number of followers, the abusive posts in responses to Bloomberg (14%) and Johnson (13%) are similarly frequent as in the case of Trump. Thunberg (8%) and Bezos (7%), on the other hand, got moderate results and the results for DiCaprio (3%) are very low.

Given the observations above, neither all the figures followed by the medium number of users face high levels of offensiveness, nor all the highly popular figures are protected against getting offensive reactions. This suggests that yet another factor different from popularity and authority contributes to the high level of offensiveness. As Trump, Johnson, and Bloomberg are known for a specific style of public activity, this style is likely to be associated with the kind of comments they get. Wodak et al. (2020) show that deliberate use of impoliteness to attract attention and

polarise discourse (so-called shameless normalisation) is a characteristic feature of right-wing populism. Trump and Johnson are typically mentioned as figures adhering to that.

Table 7: Abusive posts and abusive words across public figures (%). Letters in brackets indicate popularity (H stands for high, M for medium, L for low.) The asterisk means the largest deviations from the baseline in each column.

	posts %		**words %**		
	abusive	**non-abusive**	**abusive**	slur	rude
Biden (H)	8	92	27	8	19
Obama (H)	8	92	21	7	14
Trump (H)	14*	86	31	10	21
Johnson (M)	13*	87	37*	12	25*
Bezos (M)	7	93	36*	15	21
Bloomberg (M)	14*	86	34	18*	16
DiCaprio (M)	3	97	17	5	12
Thunberg (M)	8	92	28	7	21
Lomborg (L)	7	93	31	12	19
Miloy (L)	8	92	27	12	15
Wathuti (L)	1*	99	6*	2*	4*
Baseline	10	90	31	10	21

political movement. Bloomberg's case is more problematic, as he is known for supporting some liberal positions. On the other hand, he was elected to be the mayor of New York City, having been nominated by the Republican Party supported by Rudy Giuliani, a figure well-known for enforcing conservative policies. Bloomberg's display of sexist comments is well documented (Pengelly 2020). Since sexism is often taken to be one of the definitional characteristics of right-wing populism (Wodak et al. 2020), Bloomberg can also be classified as conversant with this realm of politics. Thus, exploring possible interdependence between populist rhetoric and offensive reactions is a subject worthy of further inquiry.

To summarise, we have demonstrated that there is no strict correlation between the popularity of public figures and the offensiveness in responses to their posts. The overall results suggest that authority associated with popularity might immunise highly popular people against offensive reactions, while figures followed by a medium number of users are more likely to attract offensive comments. Moreover, the inspection of individual results has shown that those public figures who attract offensiveness the most may differ in popularity but are similar in displaying features typical to populist rhetoric.

4.3 Reactions Can Be Offensive and Positive at the Same Time

Sentiment analysis allows for identifying speakers' positive, negative, and neutral attitudes. These are assumed to be typically correlated with an emotional load (positive, negative, neutral, respectively). The usefulness of sentiment as an indicator of various forms of linguistic misconduct, such as the use of profanities or hate speech, has been proven by Plaza del Arco et al. (2022) and Rodríguez et al. (2019). On the other hand, it is also well recognised that what may be considered vulgar words do not necessarily express hostility and often serve as intensifiers or represent a manner of speaking (Maisto et al. 2017; Yin et al. 2009). Nevertheless, the use of profanity in order to express derogatory and hostile content is typical. In this section, we test to what extent offensive language in responses to various groups of public figures serves such purposes. To this end, we test Hypothesis 3, according to which offensive language and negative sentiments are strictly correlated and inspect the frequency of sentiment types in abusive posts.

Hypothesis 3: Negative sentiment and offensive reactions are strictly correlated.

Abusive posts are the most reliable measure of offensive language, as they contain whole offensive content rather than single words that can be used repetitively in one offending comment. If negative sentiment correlates with offensiveness, it is expected to be the most frequent in abusive posts. Indeed, negative sentiment is expressed in almost all posts (93%), whereas only a minority (2%) convey positive sentiment. Thus, the sentiment distribution in abusive posts confirms the hypothesis with a minor deviation.

The strict correlation between negative sentiment and offensive reactions should also mean that the ordering of the public figures in terms of sentiment expression is the same as the ordering of offensive reactions (see Table 4): politicians attract abusive posts most frequently, followed by businessmen, contrarians, and climate activists. We assume that the ranking of negative sentiment should be identical to this one, while the ranking of positive sentiment should be reversed (as it represents the opposite attitude). The ranking of neutral sentiment is non-informative, as it indicates neither attack nor support and thus will not be included in the analysis.

This assumption turned out to not be fully confirmed by the data (see Table 8). Similarly, as in the case of offensive reactions, the activists received the lowest result in negative sentiment. As this sentiment indicates a personal attack, the result also shows that the offensive language in responses to the activists is rarely used to insult. This confirms not only Hypothesis 3 but also Hypothesis 1, according to which activists are the group affected by online offensiveness to the lowest extent. On the other hand, the result of businessmen (91%), by 4% lower than the equal results of the politicians and the contrarians (95%), disconfirms Hypothesis 3 since the businessmen were ordered higher in offensive reactions than the contrarians.

Table 8: Sentiment distribution in abusive posts across the roles of the public figures (%). The ordering reflects the actual distribution of abusive posts, as was described in the previous section. The asterisk means the largest deviations from the baseline in each column.

	sentiment %		
	negative	**neutral**	**positive**
Politicians	95*	4	1*
Businessmen	91	8	1*
Contrarians	95*	5	-
Activists	84*	5	11*
Baseline	93	5	2

In terms of positive sentiment, the top ranking of activists is not a surprise. Yet, its high frequency (11%, while the baseline is 2%) gives additional evidence that they have the sympathy of the general public: activists are not only a group with the lowest frequency of abusive posts but also with the lowest frequency of personal attacks among abusive posts. In case of individual results (see Table 9), the two lowest frequencies of negative sentiment among abusive posts belong to activists: Thunberg (83%) and DiCaprio (82%). Those are also the figures with the highest frequency of positive sentiments in abusive posts: 12% and 6% respectively. For sake of comparison, the highest frequency for non-activists equals 2%. The high frequency of negative sentiment in responses to Wathuti (100%) is anomaly worth noting, yet likely due to the limited number of comments to her posts.

Table 9: Sentiment distribution in abusive posts across public figures (%). Letters in brackets indicate social role (P stands for politicians, B for businessmen, C for climate change contrarians, and A for climate activists).

	sentiment %		
	negative	**neutral**	**positive**
Trump (P)	95	3*	2
Biden (P)	95	4	1
Obama (P)	95	3*	2
Johnson (P)	94	5	1
Miloy (C)	96*	4	-*
Lomborg (C)	93	7	-*
Bloomberg (B)	93	7	-*
Bezos (B)	89	9	2
Wathuti (A)	100	-	-*
Thunberg (A)	83	5	12*
DiCaprio (A)	82*	12*	6
Baseline	93	5	2

Posts that contain either slur or rude words are both labelled as abusive. However, only slur words are personal attacks *per se*, whereas rude words can be used without any intention to offend anyone, as in the following example of a response to the activist Greta Thunberg:

(7) @FlowerPower88: *This deserves a fucking Oscar!!*
[Reaction to: Thunberg; Abusive Word: Rude; Sentiment: Positive]

Here, the rude word 'fucking'is used to compliment rather than to attack a person. The post, despite being labelled as abusive, does not represent a personal attack. Moreover, the sentiment expressed is positive, as the author praises the target's action as deserving of an Oscar. This contrasts with the use of the same word in responses to Donald Trump:

(8) @777superhero: Fucking deluded nutcase!
[Reaction to: Trump; Abusive Words: Rude, Slur; Sentiment: Negative]

The purpose of the word 'fucking' here is to emphasise the insult expressed by the accompanying slur word 'nutcase'. Both posts are labelled as abusive. Yet only the latter conveys a personal attack. This shows that although rude words typically are used to offend, as in Example (8), they can also represent a playful manner of speaking and express positive sentiment, as in Example (7). As a result, positive sentiment is possible even among abusive posts. This highlights that the purpose of offensive language is not always a personal attack, which has been already pointed out by the relevant literature (Dynel 2012; Jay and Janschewitz 2008; Locher and Watts 2005).

4.4 One Rotten Apple Spoils the Whole Barrel

The conventional wisdom about social media portrays them as a polarised environment permeated by hate speech and offensive language. This suggests that the average user occasionally engages in multimodal offensive language, across different forms of communication. According to Chatzakou et al. (2017) users who engage in language aggression less frequently are more popular, i.e., they have more followers than those who publish abusive reactions more often. On the other hand, ElSherief et al. (2018) demonstrate that hate instigators tend to be more popular than users who refrain from offensive language. This popularity may contribute to the wider dissemination of offensive communication styles. Mathew et al. (2020) provide empirical grounds for that supposition, as their study shows that the impact of hateful behaviour tends to disperse among the users of a social medium. Thus, relevant studies show that offensive users, due to their popularity, tend to spread their attitude among those who initially avoid hateful speech. Furthermore, among other factors, anonymity on social media has been proven to facilitate the use of hateful language and contribute to the high frequency of online offensiveness (Castano-Pulgarín et al. 2021; Mondal et al. 2017).Thus, extant literature provides

grounds to support the assumption that offensive language is common in online environments. The scope of our research is limited to the comments in the context of selected COPs, and we cannot make claims about user behaviour in general. Yet, our data does suffice to reach some conclusions on their activity related to the climate debate. This is also relevant, as these conclusions reveal key patterns of online communication that concern climate change, the topic widely discussed in the contemporary public debate. Thus, in Hypothesis 4, we assume that most users, while commenting on climate debate in the context of the selected COPs, publish abusive posts more frequently than non-abusive posts. In this section, we examine whether offensiveness spreads across social media, possibly as the result of the phenomenon known as "emotional contagion", as described by Kramer et al. (2014), or whether it stems instead from the actions of a minority of users, in line with the proverb "one rotten apple spoils the whole barrel".

Hypothesis 4: The majority of users, while commenting on climate debate in the context of the selected COPs, publish more abusive posts than non-abusive posts.

To investigate the extent to which users engage in offensive arguments, we divide them into three types, distinguished by the frequency of abusive posts: above 50%, i.e., users who are more often offensive than not; equal to 50%, i.e., users who post equally frequently offensive responses and not offensive ones; and lower than 50%, i.e., users who are offensive in less than 50% posts (see Table 10). In order to avoid bias, we included only users who posted at least three times. As offensive reactions turned out to be the most frequent in responses to politicians and the least frequent in reactions to activists, we also explore the distribution of the types of users among those who reacted to these groups.

Table 10: Distribution of users across the groups distinguished by the frequency of abusive posts: larger than 50%, equal to 50%, and lower than 50%. The numbers in the brackets mean deviations from the baseline: the ↑ symbol means an upward deviation and ↓ means a downward deviation.

	offensiveness of users %		
	>50	**50:50**	**<50**
Activists	**2.2** (3.6↓)	**0.4** (1.5↓)	**97.4** (5.1↑)
Politicians	**6.7** (0.9↑)	**2.3** (0.4↑)	**91** (1.3↓)
Baseline	**5.8**	**1.9**	**92.3**

Results show that our hypothesis is wrong: 92.3% of users publish abusive posts less frequently than non-abusive posts. The proportion of those for whom the offensive reactions are predominant amounts to just 5.8%. Apparently, users who are more likely to publish abusive than non-abusive posts are a minority. In other words, most offensive reactions are the result of

the activity of a minority group. Contrary to popular opinion, most users contribute to online offensiveness only to a lesser extent.

Furthermore, the proportion of users who publish predominantly abusive posts in response to climate activists amounts to 2.2% and is lower than the baseline by 3.6%. Although this percentage for politicians reaches 6.7%, it still represents a minority of people who comment on the activity of political leaders. Thus, users who frequently cross the boundaries of civility are in the minority even among the group with the highest frequency of offensive reactions.

Despite these unexpected results, it is still possible to maintain a negative assessment of social media by claiming that users who engage in offensive arguments publish a disproportionately high number of abusive posts. However, the inspection of our data does not support such a claim either. The abusive posts make up only 10% of the posts in the whole corpus. Thus, most users either entirely avoid offensive language or participate in uncivil disputes only to a minor extent, while much of the published content is within the confines of civility.

5 Conclusions

The large-scale quantitative analysis of the corpus of offensive reactions to public figures in social media and the qualitative inquiry into specific instances of the use of offensive language, as comprised in the present study, does not suffice to determine public figures' intentions, which are not the focus of our analysis anyway. Yet, the adopted method allowed us to identify factors correlated with the range of online offensiveness. We identified political discourse as a domain particularly affected by offensive language, with the right-wing populist rhetoric as a factor that possibly contributes to the online offensiveness in some contexts. Our data revealed no strict correlation between the popularity of a public figure and the frequency of offensive reactions to their posts. We based our analysis on the assumption that offensive language can be defined in lexical or semantic terms, i.e., as the use of specific words or the conveyance of insulting meanings.

Then, the sentiment analysis has been employed to establish to what extent the use of words commonly regarded as offensive also constitutes harmful language abuse. We found empirical evidence that the reactions that we classify as offensive on lexical criteriasometimes express positive sentiment. Thus, despite containing offensive words, they do not always carry insulting meanings. We also have shown that the majority of users publish abusive content less frequently than non-abusive.

Some of our observations are worthy of further exploration. An interesting question is the relationship between *offensive language* defined with the use of linguistic cues and *hate speech* understood as a harmful misuse of language. The significant presence of positive sentiment among abusive posts in responses to climate activists proves that offensive language and hate speech are phenomena that might overlap but are not identical. The second issue is the role of

populism as a factor that attracts offensive reactions and increases polarisation. Whether populist rhetoric has such a detrimental effect on public discourse, and if so, why, is a promising area for future study. Overcoming the detrimental impact of polarisation on public discourse is one of the most important challenges for contemporary deliberative democracies. Thus, the high frequency of offensive responses to populist leaders, as revealed in our data, points out the potentially beneficial direction of further inquiries.

Funding

We would like to acknowledge that the work reported in this paper has been supported in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621 and in part by VW foundation (VolkswagenStiftung) under grant 98542.

References

Alvanoudi, Angeliki. 2024. "Conventionalized impoliteness formulae in third-party assessments". *Journal of Language Aggression and Conflict* 12.

Bächtinger, Andre, and John Parkinson. 2019. *Mapping and Measuring Deliberation. Towards a New Deliberative Quality*. Oxford: Oxford University Press.

Brown, Penelope, and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Budzynska, Katarzyna, Marcin Koszowy, Ewelina Gajewska, Maciej Kulik, and Maciej Uberna. „A Computational Method for Quantitative Analysis of Ethos. In A. Hess and J. E. Kjeldsen (eds.). *Ethos and Technology*. New York, Abingdon, Oxon: Routledge, 2024.

Camacho-Collados, Jose, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, Gonzalo Medina, Thomas Buhrmann, Leonardo Neves, and Francesco Barbieri. 2022. "TweetNLP: Cutting-Edge Natural Language Processing for Social Media". In W. Che, and E. Shutova (eds.). Proceedings *of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–49. Abu Dhabi: Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-demos.5.

Carvalho, Paula, Danielle Caled, Claudia Silva, Fernando Batista, and Ricardo Ribeiro. 2023. "The Expression of Hate Speech Against Afro-Descendant, Roma, and LGBTQ+ Communities in YouTube Comments". *Journal of Language Aggression and Conflict* 12(2): 1–31.

Castano-Pulgarín, Sergio A., Natalia Suárez-Betancur, Luz M. Tilano Vega, and Harvey M. Herrera López. 2021. "Internet, Social Media and Online Hate Speech. Systematic Review". *Aggression and Violent Behavior* 58: 1–7.

Chatzakou, Despoina, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. “Mean Birds: Detecting Aggression and Bullying on Twitter”. In *WebSci ’17: Proceedings of the 2017 ACM on Web Science Conference*, 13–22.

Chojnicka, Joanna. 2013. “Nazis vs. Occupants: The Language of Ethnic Conflict in Latvian Parliamentary Debates”. *Journal of Language Aggression and Conflict* 1(2): 225–255.

Cohen, Jacob. 2018. “A Coefficient of Agreement for Nominal Scales”. *Educational and Psychological Measurement* 20(1): 37–46.

Culpeper, Jonathan. 2011. *Impoliteness: Using Language to Cause Offence*. Cambridge: Cambridge University Press.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language”. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017),* 512–515.

De Gibert, Ona, Naiara Perez, Aitor García-Pablos, and Montse Quadros. 2018. “Hate Speech Dataset from a White Supremacy Forum”. *arXiv preprint arXiv:1809.04444*.

Druckman, James N., Erik Peterson, and Rune Slothuus. 2013. “How Elite Partisan Polarization Affects Public Opinion Formation”. *American Political Science Review* 107(1): 57–79.

Dynel, Marta. 2012. “Swearing Methodologically: The (Im)Politeness of Expletives in Anonymous Commentaries on Youtube”. *Journal of English Studies* 10: 25–50.

ElSherief, Mai, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. “Peer to Peer Hate: Hate Speech Instigators and their Targets”. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, 52–61.

Eschmann, Rob, Jacob Groshek, Rachel Chanderdatt, Khea Chang, and Maysa Whyte. 2020. “Making a Microaggression: Using Big Data and Qualitative Analysis to Map The Reproduction and Disruption of Microaggressions Through Social Media”. *Social Media + Society* 6(4): 1–13.

Falkenberg, Max, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, and Andrea Baronchelli. 2022. “Growing Polarization Around Climate Change on Social Media”. *Nature Climate Change* 12: 1114–1121.

Frimer, Jeremy A., and Linda J. Skitka. 2018. “The Montagu Principle: Incivility Decreases Politicians’ Public Approval, Even with Their Political Base”. *Journal of Personality and Social Psychology* 115(5): 845–846.

Gaupp, Franziska, and Sibel Eker. 2024. “Climate Activism, Social Media and Behavioural Change: A Literature Review” (IIASA Working Paper).

Gervais, and Bryan T. 2020. “More than Mimicry? The Role of Anger in Uncivil Reactions to Elite

Political Incivility". *International Journal of Public Opinion Research* 29(3): 384–405.

Gorrell, Genevieve, Mehmet E. Bakir, Ian Roberts, Mark A Greenwood, and Kalina Bontcheva. 2020. "Online Abuse Toward Candidates During the UK General Election 2019". *arXiv preprint arXiv:2001.08686*.

Hardaker, Claire, and Sage L. Graham. 2017. "(Im)politeness in Digital Communication". In Jonathan Culpeper, Michael Haugh, and Dániel Z. Kádár (eds.). *The Palgrave Handbook of Linguistic (Im)politeness*. London: Palgrave Macmillan: 785–814.

Hong, Traci, Zilu Tang, Manyuan Lu, Yunwen Wang, Jiaxi Wu, and Derry Wijaya. 2023. "Effects of Coronavirus Content Moderation on Misinformation and Anti-Asian Hate on Instagram". *New Media & Society* 1–24.James, David V., Frank Farnham, Seema Sukhwal, Katherine Jones, Josephine Carlisle, and Sara Henley. 2016. "Aggressive/Intrusive Behaviours, Harassment and Stalking of Members of the United Kingdom Parliament: A Prevalence Study and Cross-National Comparison". *The Journal of Forensic Psychiatry Psychology* 27(2): 118–134.

Jaszczyk-Grzyb, Magdalena, Anna Szczepaniak-Kozak, and Sylwia Adamczak-Krzysztofowicz. 2023. "A Corpus-Assisted Critical Discourse Analysis of Hate Speech in German and Polish Social Media Posts". *Moderna Språk* 117(1): 44-71.

Jay, Timothy, and Kristin Janschewitz. 2008. "The Pragmatics of Swearing". *Journal of Politeness Research: Language, Behaviour, Culture* 4(2): 267–288.

Karapetjana, Indra, Gunta Roziņa, and Margarita Spirida. 2023. "Critical Discourse Analysis of Hate Speech from a Linguistic Perspective". *Meaning and Form* 14: 74–90.

Kelleher, John D., and Brendan Tierney. 2018. *Data Science*. The MIT Press.

Kramer, Adam D. I., Jamie E. Guillory, J. E., and Jeffrey T. Hancock, 2014. "Experimental evidence of massive-scale emotional contagion through social networks". *Proceedings of the National Academy of Sciences of the United States of America* 111(24), 8788–8790.

Landowska, Alina, Katarzyna Budzynska, and He Zhang. 2024. „Quantitative and qualitative analysis of moral foundations in argumentation". *Argumentation* 38: 405–434.

Lawrence, John, Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. "Argument Analytics". In *Computational Models of Argument. Proceedings from the Sixth International Conference on Computational Models of Argument (COMMA)*, 371–378. IOS Press.

Lawrence, John, Mark Snaith, Barbara Konat, Katarzyna Budzynska, and Chris Reed. 2017. "Debating Technology for Dialogical Argument: Sensemaking, Engagement, and Analytics". *ACM Transactions on Internet Technology (TOIT)* 17(3): 1–23.

Leech, Geoffrey N. 1983. *Principles of Pragmatics*. London, New York: Longman.

Leezenberg, Michiel. 2015. "Discursive Violence and Responsibility: Notes on the Pragmatics of Dutch Populism". *Journal of Language Aggression and Conflict* 3(1): 200–228.

Leiserowitz, Anthony, Edward Maibach, Seth Rosenthal, John Kotcher, Parrish Bergquist, Abel Gustafson, Matthew Ballew, and Matthew Goldberg. 2019. *Climate Activism: Beliefs, Attitudes, and Behaviors, November 2019*. New Haven, CT: Yale University and George Mason University.

Lepoutre, Maxime, Sara Vilar-Lluch, Emma Borg, and Nat Hansen. 2024. "What is Hate Speech? The Case for a Corpus Approach". *Criminal Law and Philosophy* 18: 397–430.

Locher, Miriam A., and Richard J. Watts. 2005. "Politeness Theory and Relational Work". *Journal of Politeness Research: Language, Behaviour, Culture* 1(1): 9–33.

Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. "TimeLMs: Diachronic Language Models from Twitter". In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260, Dublin: Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-demo.25.

Määttä, Simo. 2023. "Linguistic and Discursive Properties of Hate Speech and Speech Facilitating the Expression of Hatred: Evidence from Finnish and French Online Discussion Boards". *Internet Pragmatics* 6(2): 156–172.

Maisto, Alessandro, Serena Pelosi, Simonetta Vietri, and Pierluigi Vitale. 2017. "Mining Offensive Language on Social Media". In *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it,* 252–256.

Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. "Spread of Hate Speech in Online Social Media". In *WebSci '19: Proceedings of the 10th ACM Conference on Web Science*, 173–182.

Mathew, Binny, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. "Hate Begets Hate: A Temporal Study of Hate Speech". In *Proceedings of the ACM on Human-Computer Interaction*, volume 4, 1–24.

Meutia, Rita, and Bambang Gulyanto. 2022. "Verbal Aggressiveness Against Public Figures Language: An Analysis of Tweeps' Comments on Twitter". *Asian Journal of Behavioural Sciences* 4(2): 118–134.

Mondal, Mainack, Leandro Araújo Silva, and Fabricio Benevenuto. 2017. "A Measurement Study of Hate Speech in Social Media". In *HT '17: Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94.

Parkinson, John, and Jane Mansbridge. 2012. *Deliberative Systems. Deliberative Democracy at the Large Scale*. Cambridge: Cambridge University Press.

Park, Chang S., Qian Liu, and Barbara K. Kaye. 2021. "Analysis of Ageism, Sexism, and Ableism in User Comments on YouTube Videos about Climate Activist Greta Thunberg". *Social Media +*

Society 7(3): 1–14.

Parvaresh, Vahid. 2023. "Covertly Communicated Hate Speech: A Corpus-Assisted Pragmatic Study". *Journal of Pragmatics* 205: 63–77.

Pengelly, Martin. 2020. "Mike Bloomberg Rocked by Re-Emergence of Sexist Remarks". In *The Guardian*. URL https:// www.theguardian.com/us-news/2020/feb/15/michael-bloomberg-booklet-sexist-remarks-abortion (visited: 2024-04-23).

Pereira-Fariña, Martin, Marcin Koszowy, and Katarzyna Budzynska. 2022. „It Was Never Just About The Statue: Ethos of Historical Figures in Public Debates on Contested Cultural Objects". Discourse and Society, 33(2): 193–214.

Plaza del Arco, Flor M., Sercan Halat, Sebastian Padó, and Roman Klinger. 2022. "Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language". *arXiv:2109.10255v4*.

Rasulo, Margaret. 2021. "Are Gold Hoop Earrings and a Dab of Red Lipstick Enough to Get Even Democrats on the Offensive? The Case of Alexandria Ocasio-Cortez". *Journal of Language Aggression and Conflict* 9(1): 155–183.

Rega, Rossella, Rita Marchetti, and Anna Stanziano. 2023. "Incivility in Online Discussion: An Examination of Impolite and Intolerant Comments". *Social Media + Society* 9(2): 1–12.

Retta, Mattia. 2023. "A Pragmatic and Discourse Analysis of Hate Words on Social Media". *Internet Pragmatics* 6(2): 197–218.

Riedl, Martin J., Katie Joseff, Stu Soorholtz, and Samuel Woolley. 2022. "Platformed Antisemitism on Twitter: Anti-Jewish Rhetoric in Political Discourse Surrounding the 2018 US Midterm Election". *New Media & Society*, 1–21.

Rodríguez, Axel, Carlos Argueta, and Yi-Ling Chen. 2019. "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis". In *Conference: 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 169–174.

Rösner, Leonie and Nicole C. Krämer. 2016. "Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments". *Social Media + Society* 2(3): 1–13.

Schäfer, Mike S. and Peter North. 2019. "Are social media making constructive climate policymaking harder?" In *Contemporary Climate Change Debates.* London: Routledge.

Schroeter, Melani. 2018. "How Words Behave in Other Languages: The Use of German Nazi Vocabulary in English". *Pragmatics and Society* 9(1): 91–116.

Simchon, Almog, William J. Brady, and Jay J. Van Bavel. 2022. "Troll and Divide: The Language of Online Polarization". *PNAS Nexus* 1(1): 1–12.

Stahel, Lea and Dirk Baier. 2023. “Digital Hate Speech Experiences Across Age Groups and their Impact on Well-Being: A Nationally Representative Survey in Switzerland”. *Cyberpsychology, Behavior, and Social Networking*, 26(7): 519–526.

Teneketzi, Korallia. 2022. “Impoliteness across social media platforms: A comparative study of conflict on YouTube and Reddit”. *Journal of Language Aggression and Conflict*, 10(1): 38–63.

Terkourafi, Marina, Lydia Catedral, Iftikhar Haider, Farzad Karimzad, Jeriel Melgares, Cristina Mostacero-Pinilla, Julie Nelson, and Benjamin Weissman. 2018. “Uncivil Twitter: A Sociopragmatic Analysis”. *Journal of Language Aggression and Conflict* 6(1): 26–57.

Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, and Sebastian A. Popa. 2020. “The Dynamics of Political Incivility on Twitter”. *Sage Open*, 1–15.

Thomas-Walters, Laura, Eric G. Scheuch, Abby Ong, and Matthew H. Goldberg. 2025. “The Impacts of Climate Activism”. *Current Opinion in Behavioral Sciences* 63: 101498.

Vasist, Pramukh N., Debashis Chatterjee, and Satish Krishnan. 2023. “The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-Country Configural Narrative”. *Information Systems Frontiers* 26(2): 1–26.

Vergani, Matteo, Alfonso Martinez Arranz, Ryan Scrivens, and Liliana Orellana. 2022. “Hate Speech in a Telegram Conspiracy Channel During the First Year of the COVID-19 Pandemic”. *Social Media + Society* 8(4): 1–14.

Walker, Russell. 2015. *From Big Data to Big Profits: Success with Data and Analytics*. New York: Oxford University Press.

Wilson, Anne E., Victoria A. Parker, and Matthew Feinberg. 2020. “Polarization in the Contemporary Political and Media Landscape”. *Current Opinion in Behavioral Sciences* 34(8): 223–228.

Wodak, Ruth, Jonathan Culpeper, and Elena Semino. 2020. “Shameless Normalisation of Impoliteness: Berlusconi’s and Trump’s Press Conferences”. *Discourse & Society* 32(3): 369–393.

Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian Davison, April Kontostathis, and Lynne Edwards. 2009. “Detection of Harassment on Web 2.0”. In *Proceedings of the Content Analysis in the WEB*, volume 2, 1–7.